

# GISRUK 2015 Proceedings

Nick Malleeson, Nicholas Addis, Helen Durham, Alison Heppenstall, Robin Lovelace, Paul Norman and Rachel Oldroyd

15 – 17 April 2015

## Preface

This volume contains the papers presented at GIS Research UK 2015 (GISRUK2015) held at the School of Geography, University of Leeds, on 15-17 April 2015.

In recognition of the recently funded £6M+ ESRC Consumer Data Research Centre at Leeds, the theme for the conference this year was: *Big Data and the Future of GIS*

## History

Established in 1993, the GISRUK conference series is the UK's national GIS research conference. Each annual conference attracts delegates from all parts of the world from disciplines including geography, computer science, planning, archaeology, geology, geomatics and engineering. The GISRUK conferences aim to act as a focus for GIS research in the UK and provide a mechanism for the announcement and publication of GIS research. They also provide a forum for the discussion of research ideas, promote collaboration amongst researchers from diverse disciplines and offer postgraduate students the opportunity to showcase their work in an international context.

## Program Committee

Nick Malleeson (Chair)	University of Leeds
Nicholas Addis	University of Leeds
Helen Durham	University of Leeds
Alison Heppenstall	University of Leeds
Robin Lovelace	University of Leeds
Paul Norman	University of Leeds
Rachel Oldroyd	University of Leeds

## GISRUK2015 Proceedings

All submissions to the conference were reviewed by at least one reviewer. 106 papers were accepted to the conference and subsequently invited to appear in the proceedings.

For more information about the conference, see the conference website:

<http://leeds.gisruk.org/>

I would like to thank the GISRUK2015 programme committee for their ongoing work on organising the conference and members of the Centre for Spatial Analysis and Policy (CSAP) for reviewing papers.

I would also like to thank our four keynote speakers:

- Ed Parsons, *Geospatial Technologist, Google*
- Steven Ramage, *Director of Strategy, what3words*
- Dominic Stubbins, *Chief Architect, ESRI UK*
- Sarah Williams, *Director of the Civic Data Design Lab, MIT School of Architecture and Planning*

Finally, I would like to thank the conference sponsors:

- Ordnance Survey
- ESRI UK
- EDiNA
- UK Data Service (UKDS) Census Support
- Google
- Royal Geographical Society (RGS)
- Association for Geographic Information (AGI)

*Nick Malleon, GISRUK2015 chair*



## Contents

Preface	i
History	i
Program Committee	i
GISRUK2015 Proceedings	ii
<b>51. A new metric of crime hotspots for operational Policing</b> <i>Monsuru Adepeju, Tao Cheng, John Shawe-Taylor and Kate Bowers</i>	<b>1</b>
<b>41. Exploring the geo-temporal patterns of the Twitter messages</b> <i>Muhammad Adnan, Guy Lansley and Paul Longley</i>	<b>7</b>
<b>64. Participatory mapping for transformation: multiple visual representation of foodscapes and environment in informal settlements in Nairobi</b> <i>Sohel Ahmed, Muki Haklay, Adriana Allen, Cecilia Tacoli and Dvila Julio</i>	<b>14</b>
<b>23. Can Administrative Data Be Used To Create A Geodemographic Classification?</b> <i>Mildred Ajebon and Paul Norman</i>	<b>20</b>
<b>25. A geospatial relational database schema for interdependent network analysis and modelling</b> <i>David Alderson, Stuart Barr, Tomas Holderness, Craig Robson and Alistair Ford</i>	<b>33</b>
<b>68. The Role of Geographical Context in Building Geodemographic Classifications</b> <i>Alexandros Alexiou and Alexander Singleton</i>	<b>40</b>
<b>132. Utilising GIS capabilities to study and analyse the spatial distribution of crimes in Kuwait</b> <i>Nawaf Alfadhli, Graham Clarke and Mark Birkin</i>	<b>46</b>
<b>86. Characterisation and Classification of Hydrological Catchments in Alberta Canada Using Growing Self-Organising Maps</b> <i>Michael Allchin</i>	<b>55</b>
<b>73. New approaches to measure the spatial structure(s) of cities</b> <i>Daniel Arribas-Bel and Emmanouil Tranos</i>	<b>64</b>
<b>92. Big Data Analysis of Population Flow between TfL Oyster and Bicycle Hire Networks in London</b> <i>Nilufer Sari Aslam, James Cheshire and Tao Cheng</i>	<b>69</b>

109. <b>Researching long-run trends in South East England 1841-2011 for the European Union and Greater London Authority</b> <i>Paula Aucott and Humphrey Southall</i>	76
95. <b>Understanding Spatio Temporal Patterns of Crime Using Hotspot AND Coldspot Analysis</b> <i>Ellie Bates and William Mackaness</i>	80
24. <b>TravelOAC: development of travel geodemographic classifications for England and Wales based on open data</b> <i>Nick Bearman and Alex Singleton</i>	87
105. <b>Geodemographics and Big Data: A New Research Agenda</b> <i>Mark Birkin</i>	93
106. <b>Integrating BIM and GIS : Exploring the use of IFC space objects and boundaries</b> <i>Gareth Boyes, Charles Thomson and Claire Ellul</i>	98
83. <b>Spoilt for Choice? An Investigation Into Creating Gastner and Newman-style Cartograms</b> <i>Chris Brunsdon and Martin Charlton</i>	107
75. <b>Evolutionary Computing for Multi-Objective Spatial Optimisation</b> <i>Daniel Caparros-Midwood, Stuart Barr and Richard Dawson</i>	116
15. <b>sDNA: how and why we reinvented Spatial Network Analysis for health economics and active modes of transport</b> <i>Crispin Cooper and Alain Chiaradia</i>	122
31. <b>HAG-GIS: A spatial framework for geocoding historical addresses</b> <i>Konstantinos Daras, Zhiqiang Feng and Chris Dibben</i>	128
82. <b>Quantifying the deterrent effect of police patrol via GPS analysis</b> <i>Toby Davies and Kate Bowers</i>	135
2. <b>Ephemeral Londoners: Modelling Lower Class Migration to Eighteenth Century London</b> <i>Adam Dennett, Adam Crymble, Tim Hitchcock and Louise Falcini</i>	139
87. <b>Identifying perpetuation in processes driving fish movement</b> <i>Matt Duckham, Antony Galton and Alan Both</i>	142

111. <b>Inequality in access to education and inequality in access to information about allocation of school places</b> <i>Oliver Duke-Williams, Elizabeth Shepherd and Alexandra Eveleigh</i>	148
101. <b>Assessing spatial distribution and variability of destinations in inner-city Sydney from travel diary and smartphone location data</b> <i>Richard Ellison, Adrian Ellison and Stephen Greaves</i>	152
98. <b>Spatiotemporal Identification of Trip Stops from Smartphone Data</b> <i>Adrian Ellison, Richard Ellison, Asif Ahmed and Stephen Greaves</i>	159
42. <b>Exploring new ways of digital engagement: a study on how mobile mapping and applications can contribute to disaster preparedness</b> <i>Gretchen Fagg, Enrica Verrucci and Patrick Rickles</i>	164
37. <b>Comparing Methods: Using Multilevel Modelling and Artificial Neural Networks in the Prediction of House Prices based on property location and neighbourhood characteristics</b> <i>Yingyu Feng and Kelvyn Jones</i>	168
49. <b>Assessing the need for infrastructure adaptation by simulating impacts of extreme weather events on urban transport infrastructure</b> <i>Alistair Ford, Maria Pregnolato, Richard Dawson, Stuart Barr and Katie Jenkins</i>	175
124. <b>Calculating the Overbuilding Potential of Municipal Buildings in London</b> <i>Joanna Foster, Claire Ellul and Philippa Wood</i>	183
119. <b>Assessing the quality of OpenStreetMap building data and searching for a proxy variable to estimate OSM building data completeness</b> <i>Claire Fram, Katerina Chistopoulou and Claire Ellul</i>	195
38. <b>Assessing geographic data usability in analytical contexts: Undertaking sensitivity analysis of geospatial processes</b> <i>Robin Frew, Gary Higgs, Mitchel Langford and Jenny Harding</i>	206
123. <b>Profiling Burglary in London using Geodemographics</b> <i>Chris Gale, Alex Singleton and Paul Longley</i>	214
70. <b>Assessment of social vulnerability under three flood scenarios using an open source vulnerability index</b> <i>Kurtis Garbutt, Claire Ellul and Taku Fujiyama</i>	220
3. <b>Mapping of Spatial Distribution of Tuberculosis Cases in Kebbi state Nigeria 2008-2011</b> <i>Usman Lawal Gulma</i>	227

19. <b>The Influence of Familiarity on Route Choice: Edinburgh as a Case Study</b> <i>Maud van Haeren and William Mackaness</i>	236
45. <b>Real time coupled network failure modelling and visualisation</b> <i>Neil Harris, Craig Robson, Stuart Barr and Phil James</i>	246
52. <b>Football fan locality- An analysis of football fans tweet locations</b> <i>Neil Harris and Phil James</i>	252
14. <b>Do Geospatial &amp; Heritage standards work and do they work together?</b> <i>Glen Hart</i>	262
47. <b>Objectively scrutinising the impact of the obesogenic environment on obesity in Yorkshire England: a multi-level cross-sectional study</b> <i>Matthew Hobbs, Jim McKenna, Mark Green, Hannah Jordan and Claire Griffiths</i>	266
13. <b>Evaluating the Spraycan: understanding participant interaction with a PPGIS</b> <i>Jonathan Huck, Duncan Whyatt and Paul Coulton</i>	270
77. <b>Abstract Feature Representation as a Cartographic Device for Mixed-Reality Location-Based Games</b> <i>Jonny Huck, Paul Coulton, Adrian Gradinar and Duncan Whyatt</i>	277
32. <b>Development of public transport accessibility in the Czech Republic</b> <i>Igor Ivan</i>	286
81. <b>SAFEVolcano: Spatial Information Framework for Volcanic Eruption Evacuation Site Selection-allocation</b> <i>Jumadi Jumadi, Steve Carver and Duncan Quincey</i>	291
126. <b>Geodemographics and spatial microsimulation: using survey data to infer health milieu geographies</b> <i>Jens Kandt</i>	302
78. <b>Designing a location model for face to face and on-line retailing for the UK grocery market</b> <i>Elena Kirby-Hawkins, Graham Clarke and Mark Birkin</i>	312
40. <b>Data-driven modelling of police route choice</b> <i>Kira Kowalska, John Shawe-Taylor and Paul Longley</i>	317
76. <b>Optimising sentiment analysis in commercial context</b> <i>Radoslaw Kowalski</i>	324

26. <b>Spatio-Temporal Patterns of Passengers' Interests at London Tube Stations</b> <i>Juntao Lai, Tao Cheng and Guy Lansley</i>	329
79. <b>Mapping to Disrupt unjust urban trajectories</b> <i>Rita Lambert and Adriana Allen</i>	334
102. <b>Creating an Output Area Classification of Cultural and Ethnic Heritage to Assist the Planning of Ethnic Origin Foods in Supermarkets in England and Wales</b> <i>Guy Lansley, Yiran Wei and Tim Rains</i>	339
27. <b>Towards a Seamless World Names Database</b> <i>Alistair Leak, Paul Longley and Muhammad Adnan</i>	347
90. <b>Land-use Simulation at Large-scale using Big Data</b> <i>Dan Li</i>	355
110. <b>UK internal migration by ethnicity</b> <i>Nik Lomax and Phil Rees</i>	360
16. <b>Mapping Interactive Behaviour in Wildlife from GPS Tracking Data</b> <i>Jed Long</i>	366
71. <b>Crowd sourced vs centralised data for transport planning: a case study of bicycle path data in the UK</b> <i>Robin Lovelace</i>	373
55. <b>Strategies in the Use of Referring Expressions to Describe Things Urban</b> <i>William Mackaness, Phil Bartie and Philipp Petrenz</i>	380
115. <b>Using Mobile Phone Traces to Understand Activity and Mobility in Dakar Senegal</b> <i>Ed Manley, Adam Dennett and Michael Batty</i>	387
129. <b>A Spatiotemporal Population Subgroup Model of Radiation Exposure</b> <i>Becky Martin, David Martin and Samantha Cockings</i>	396
54. <b>Understanding the urban experience of people with visual impairments</b> <i>Panos Mavros, Katerina Skroumpelou and Andrew Hudson Smith</i>	401
20. <b>Traffic Prediction and Analysis using a Big Data and Visualisation Approach</b> <i>Declan McHugh</i>	408
85. <b>Beyond Visualisation in 3D GIS</b> <i>James Milner, Kelvin Wong and Claire Ellul</i>	421

<b>5. Visualize and interactively design weight matrices</b> <i>Angelos Mimis</i>	<b>430</b>
<b>114. Census statistics for Civil Parishes: When best-fitting just isn't good enough</b> <i>Bruce Mitchell</i>	<b>436</b>
<b>58. Comparing the Quality of Local Authority Spatial Data</b> <i>Amy Mizen, Sarah Rodgers and Richard Fry</i>	<b>443</b>
<b>65. Is VGI Big Data?</b> <i>Peter Mooney and Adam Winstanley</i>	<b>448</b>
<b>97. Exploring the role of consumer data for food in national survey reporting</b> <i>Michelle Morris, Graham Clarke and Mark Birkin</i>	<b>454</b>
<b>91. Retail Modelling in Tourist Resorts: A case study of Looe Cornwall</b> <i>Andy Newing, Graham Clarke and Martin Clarke</i>	<b>460</b>
<b>21. The changing geography of deprivation in Britain: exploiting small area census data 1971 to 2011</b> <i>Paul Norman</i>	<b>465</b>
<b>125. Data Exploration with GIS Viewsheds and Social Network Analysis</b> <i>Giles Oatley, Tom Crick and Ray Howell</i>	<b>475</b>
<b>113. A Framework for Big Data in studies of Urban Mobility and Movement</b> <i>Eusebio Odiari, Mark Birkin, Susan Grant-Muller and Nick Malleeson</i>	<b>480</b>
<b>34. A national-scale application of the Huff gravity model for the estimation of town centre retail catchment area</b> <i>Michail Pavlis, Les Dolega and Alexander Singleton</i>	<b>481</b>
<b>36. Development and application of a two stage hybrid spatial microsimulation technique to provide inputs to a model of capacity to walk and cycle</b> <i>Ian Philips</i>	<b>486</b>
<b>11. Combining Statistics and Texts Using GIS: Nineteenth Century Health Reports</b> <i>Catherine Porter, Paul Atkinson and Ian Gregory</i>	<b>492</b>
<b>127. Alternative Approaches to Forecasting Migration: Framework and UK Illustrations</b> <i>Philip Rees, Nikolaos Lomax and Peter Boden</i>	<b>500</b>

128. <b>Learning Lessons from Population Projections: How Well Did We Forecast the Ethnic Transition?</b> <i>Philip Rees and Pia Wohland</i>	522
30. <b>Spatially modelling dependent infrastructure networks</b> <i>Craig Robson, Stuart Barr, Philip James and Alistair Ford</i>	536
35. <b>The Complexity of Exclusion</b> <i>Jacobus Van Rooyen and Joana Barros</i>	542
33. <b>A self-exciting point process model for predictive policing: implementation and evaluation</b> <i>Gabriel Rosser and Tao Cheng</i>	547
9. <b>Constrained clustering of the precipitation regime in Greece</b> <i>Eftychia Rousi, Christina Anagnostopoulou, Angelos Mimis and Marianthi Stamou</i>	553
50. <b>Accessibility-based simulation of urban expansion in Brazil</b> <i>Marcus Saraiva, Joana Barros and Mauricio Polidori</i>	559
22. <b>Group Behaviour Analysis of London Foot Patrol Police</b> <i>Jianan Shen and Tao Cheng</i>	565
53. <b>Are we there yet? Exploring distance perception in urban environments with mobile Electroencephalography</b> <i>Katerina Skroumpelou, Panagiotis Mavros and Andrew Hudson Smith</i>	570
43. <b>Semantic and geometric enrichment of 3D geo-spatial building models with photo captions and illustration labels</b> <i>Jon Slade, Christopher Jones and Paul Rosin</i>	583
17. <b>Assessing the impact of seasonal population fluctuation on regional flood risk management</b> <i>Alan Smith, Andy Newing, Niall Quinn, David Martin and Samantha Cockings</i>	590
122. <b>Creating a spatio-temporal Data Feed API for a large and diverse library of historical statistics for areas within Britain</b> <i>Humphrey Southall and Michael Stoner</i>	599
4. <b>GIS Big Data and Lessons from John Snow</b> <i>Doug Specht</i>	603
39. <b>The Impact of Task Workflow Design on VGI Citizen Science Platforms</b> <i>James Sprinks, Jeremy Morley, Robert Houghton and Steven Bamford</i>	607

18. <b>Assessing the risk landslides pose to road and rail networks</b> <i>Khaled Taalab and Tao Cheng</i>	612
56. <b>Comparing different spatial microsimulation frameworks</b> <i>Melanie Tomintz and Bernhard Kosar</i>	620
100. <b>Using GeoTools to explore Advice Leeds client data</b> <i>Andy Turner and Stuart Hodgkinson</i>	624
121. <b>Temporal profile of daily sales in retail stores in London</b> <i>Syed Rakib Uddin and Professor Paul Longley</i>	637
118. <b>Understanding the spatial pattern of urban crime: a developing country's perspective</b> <i>Faisal Umar, James Cheshire and Shane Johnson</i>	641
46. <b>Reconstructing the Agricultural Landscape of the South Downs England: an Examination of the 1940 and 1941 World War II Plough-up Campaigns</b> <i>Nigel Walford</i>	650
69. <b>Visualisation of Spread of Chalara Ash Dieback for Raising Public Awareness and Responsible Woodland Access</b> <i>Chen Wang, David Miller, Paula Horne, Yang Jiang, Gillian Donaldson-Selby and Jane Morrice</i>	653
28. <b>Is the use of 'mobile computer technology' appropriate for locating people with dementia?</b> <i>Steve Williams and Mark Ware</i>	659
60. <b>Exploratory spatiotemporal data analysis of public confidence in the police in London</b> <i>Dawn Williams, James Haworth and Tao Cheng</i>	665
89. <b>Can the sentiment expressed in trail users' tweets help to assess the effectiveness of Environmental Stewardship Agreements? An exploratory analysis of the Pennine Way National Trail England.</b> <i>Tom Wilson and Robin Lovelace</i>	673
48. <b>A Model Officer: An Agent-based Model of Policing</b> <i>Sarah Wise and Tao Cheng</i>	680
112. <b>Estimates of ethnic mortality in the UK revisited</b> <i>Pia Wohland and Phil Rees</i>	686



103. <b>Designing 3D Geographic Information for Navigation Using Google Glass</b> <i>Kelvin Wong and Claire Ellul</i>	<b>692</b>
84. <b>The role of the age() function in a GIS for ecological conservation</b> <i>Greg Wood, Duncan Whyatt and Carly Stevens</i>	<b>699</b>
66. <b>Understanding car ownership elasticities in England and Wales: Advancing the evidence base with new data sources</b> <i>Godwin Yeboah, Jillian Anable, Tim Chatterton, Jo Barnes, Eddie Wilson, Oliver Turnbull and Sally Cairns</i>	<b>704</b>
96. <b>Modelling the long-term economic and demographic impacts of major infrastructure provision: a simultaneous model approach</b> <i>Chengchao Zuo and Mark Birkin</i>	<b>714</b>

# A new metric of crime hotspots for Operational Policing

Monsuru Adepeju<sup>\*1</sup>, Tao Cheng<sup>†1</sup>, John Shawe-Taylor<sup>‡2</sup>, Kate Bowers<sup>‡3</sup>

<sup>1</sup>SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental and Geomatic Engineering, University College London

<sup>2</sup>Department of Security and Crime Science, University College London

<sup>3</sup>Department of Computer Science, University College London

November 07, 2014

## Summary

This study examines the existing metrics used in evaluating the effectiveness of area-based crime hotspots for operational policing. We identified some of the limitations of the metric (i.e. Area-to-Perimeter (AP) ratio) used for measuring compactness of hotspots and then proposed a new improved metric called “Clumpiness Index (CI)”. The case study of London Metropolitan police crime dataset features the prediction of 3 different crime types using two different crime predictive methods. The effectiveness of the hotspots was then measured using both AP ratio and CI. The comparison of the results clearly shows that CI is a better metric for measuring the effectiveness of crime hotspots for operational policing.

**KEYWORDS:** Effective hotspots, Area-to-perimeter (AP) ratio, Hit rate, Clumpiness Index (CI), Operational Policing.

## 1. Introduction

As police resources are becoming increasingly limited due partly to budget constraints, there is a growing interest in strategies that can enhance optimisation of their resources towards achieving the desired crime prevention goals. An effective policing strategy is one that offers the police high crime prevention potential with a small amount of police deployment (Weisburd, 2008). Over the last decade, attempts to increase police effectiveness have resulted in operational policing being informed by predictive analysis of crime. The predictive methods are used to identify locations of high future crime risk. These locations are referred to as crime hotspots. The types of hotspots include point-based, network-based and area-based hotspots. Several studies have suggested that police can be more effective in intervening in crime by focussing on small geographical units with high crime rates (hotspots) rather than actual people (offenders) committing the crimes (Telep & Weisburd, 2014). As a result, most predictive methods of crime have been aimed at identifying area-based hotspots.

To estimate how effective the detected hotspots are for operational policing, two metrics have been used, proposed by Bowers et al. (2004). These metrics are Hit rate (HR) and Area-to-Perimeter (AP) ratio. The HR measures the proportion of future crime accurately captured by the purported hotspots while AP ratio measures compactness (easiness of covering) the hotspots. Bowers et al. (2004) recommended that the two measures should be used together for meaningful evaluation. This is because hotspots with high HR may not necessarily be easily coverable based on their geometric shape. Thus, hotspots with moderate HR and a high AP ratio are preferred to ensure effective policing. However, certain limitations can be identified with the AP ratio which renders it less appealing for evaluating hotspots for effective policing.

The AP ratio is used to measure the geometric complexities (compactness) of hotspots. The

---

<sup>\*</sup> monsuru.adepeju.11@ucl.ac.uk

<sup>†</sup> tao.cheng@ucl.ac.uk

<sup>‡</sup> j.shawe-taylor@ucl.ac.uk

<sup>‡</sup> kate.bowers@ucl.ac.uk

assumption is that regular-shaped hotspots (e.g. squares) can be covered quicker and more easily than irregularly-shaped hotspots, if we ignore the underlying network structure. The AP ratio has limitations. They are:

- a) Holding the shape of a hotspot constant, the AP ratio varies with the spatial scale, (ranging from zero to infinity). This makes it difficult to compare similar hotspots across different study areas.
- b) The level of disaggregation or dispersion of the hotspots (grid units) cannot be inferred from the value of the AP ratio. Therefore, the AP ratio cannot give us an idea of randomly distributed hotspots as a baseline for comparison.
- c) The AP ratio is relatively insensitive to differences in the structure of hotspots. Thus, although hotspots may possess very different shapes, they may have identical area and perimeters.

Therefore, the goal of this paper is to propose a new metric for measuring effectiveness of crime hotspots for operational policing. Specifically, we are proposing a new metric called Clumpiness Index as an alternative to AP ratio given the limitations of AP ratio listed above.

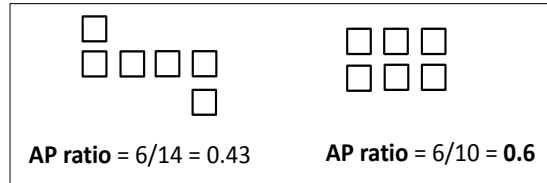
## 2. Existing Metrics – Hit Rate and AP Ratio

**2.1 Hit Rate:** the proportion of new crimes captured by the defined hotspot. Evaluated at a certain area coverage (e.g. 20% area coverage)

$$Hit\ Rate = \left( \frac{\sum_{k=1}^m (number\ of\ crimes)}{\sum_{i=1}^n (number\ of\ crimes)} \right) \times 100 \quad (1)$$

Where  $i$  = number of ranked grids;  $k$  = number of percentile of ranked grids e.g. 20th;  $n$  = total number of grids.

**2.2 Area-to-perimeter (AP) ratio:** a measure of how compact an identified cluster (hotspot) is. The more compact a hotspot is, the easier and quicker it will be to cover operationally. Higher AP ratio corresponds to better compactness (Figure 1).



**Figure 1** Area-to-Perimeter (AP) Ratio. The hotspot on the right pane is more compact and therefore has higher AP ratio and may be seen as more efficient in operational policing terms.

## 3. A new metric - Clumpiness Index (CI)

Provided a modest hit rate, the actual effectiveness of a predictive solution is measured in terms the geometric complexity (compactness) and distribution of the hotspots across a geographical area. We propose a new metric called “Clumpiness Index (CI)” which measures the compactness and distribution of hotspots and is robust to scaling issues of AP ratio.

The Clumpiness Index (CI) was originally proposed by Turner (1989) as Contagion Index for measuring the overall clumpiness of categorical patches on a landscape. CI is able to measure effectively both patch type interspersion (i.e. the intermixing of units of different patch types) as well as patch dispersion (i.e. the spatial distribution of a patch type) at the landscape level. CI is computed by first summarising the adjacency of all cells in an adjacency matrix, which shows the frequency with which different pairs of patch types (including adjacencies between the same patch type) appear side-by-side on the map. CI is defined as follows:

$$G_i = \left( \frac{g_{ii}}{\sum_{k=1}^m g_{ik}} \right); CI = \begin{cases} \frac{G_i - P_i}{1 - P_i} & \text{for } G_i \geq P_i \\ \frac{G_i - P_i}{1 - P_i} & \text{for } G_i < P_i; P_i \geq 0.5 \\ \frac{P_i - G_i}{-P_i} & \text{for } G_i < P_i; P_i < 0.5 \end{cases} \quad (2)$$

$g_{ii}$  = number of like adjacencies (joins) between pixels of patch type (class)  $i$

$g_{ik}$  = number of adjacencies (joins) between pixels of patch types (classes)  $i$  and  $k$

$P_i$  = proportion of the landscape occupied by patch type (class)  $i$ .

The goal of CI is to determine the maximum value of  $g_{ii}$  for any  $P_i$

The CI takes values between -1 (when the class is maximally disaggregated) to 1 (when the class is maximally aggregated corresponding to a checkerboard arrangement).

## 4. Case Study

### 4.1 Camden Borough of London

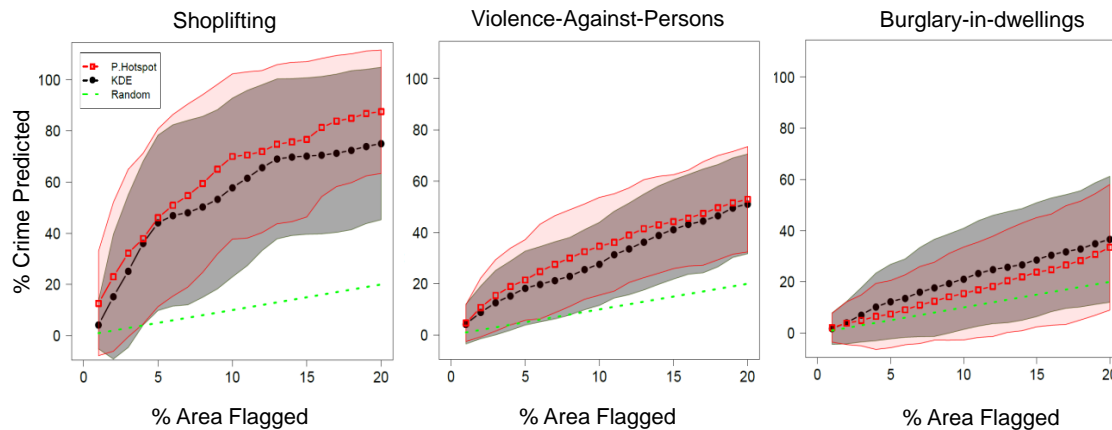
Camden Borough is one of the 12 inner boroughs of London City with an estimated 224,962 inhabitants as of 2011. The population density is estimated as approximately 10,000 people per square kilometre (2011 Census, Office of National Statistics). The Borough contains a mixture of commercial and residential areas with the busiest parts being the *Camden Market* and *Covent Garden and Holborn*. The borough recorded a crime rate of 145 crimes per 1,000 people in 2010/11, the national average being 75 crimes per 1,000 people (Source: Metropolitan Police Service website, 2014).

### 4.2 Effectiveness of hotspot for operational policing

Three different crime types are used in this analysis. They are shoplifting, violence-against-persons and burglary (in-dwellings) crimes. The data points are aggregated to a grid system of 250m by 250m and have temporal resolution of 1 day. The time period predicted is between 28/09/2011 and 6/01/2012, predicting 2 days ahead using the crime risk surface produced on each day.

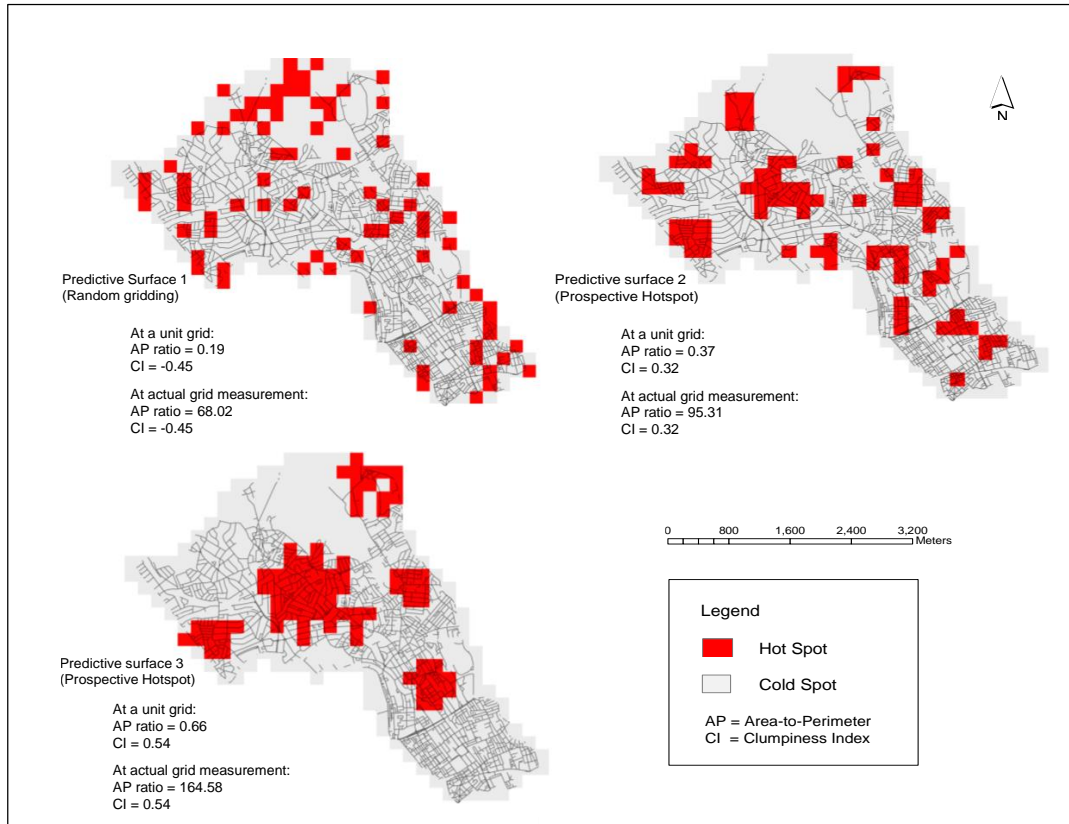
To check the predictions against the real dataset, we overlay future crime on the predictive surface generated. For example, validating the prediction on day  $t_n$  means overlaying crime data from day  $t_{n+1}$  to day  $t_{n+2}$  on the predictive surface generated on day  $t_n$ . By so doing, the proportion of crimes that are captured by the ranked top 20% of the grids squares (hotspots) is evaluated (assuming that police only has resources to cover just 20% of the Camden).

Two hotspots predictive methods are used, namely (i) Prospective Hotspot (Bowers et al. 2004) and (ii) Kernel Density Estimation (KDE) method. Figure 2 represents the average of percentage hit rate over the prediction period. Also included in Figure 2 is baseline prediction which is generated by way of picking grid squares with equal probability (random) until 20% coverage is attained. This is represented with the green line. The general performance of these predictive methods in terms of hit rates follows the spatial concentration of different crime types with highly spatially concentrated crime type showing highest hit rates. For example, shoplifting crime is highly concentrated in a few regions near commercial areas and therefore shows the highest hit rates whereas prediction of burglary crimes is lowest as residential properties are dispersed across the entire borough.



**Figure 2** Average % hit rates over the prediction period

In measuring hotspot compactness, we adapted CI to crime hotspots by classifying grid squares constituting the predictive surface into two types, namely (1) Hot Spot – the top 20% ranked grids and (2) Cold Spot – the remaining 80%. Figure 3 shows examples of predictive surfaces generated by Prospective Hotspot method and random grid selection to illustrate how AP ratio and CI vary with different hotspot configurations. In each example, we consider two spatial scalings: the original data (unchanged), and data scaled such that each grid square has unit length. The AP ratio is observed to change at different scales of measurement of the same surface, making it difficult to compare. However, CI remains the same at any given scale. This is because CI is based on the adjacencies as well as the proportion of the hotspot grids across total surface. Therefore, CI is able to provide a sense of dispersion from a complete disaggregation ( $CI = -1$ ) of the hotspot units.



**Figure 3** Evaluating hotspot compactness with AP ratio and CI

## 5. Discussion and Conclusion

This study examines the use of existing metrics for measuring the effectiveness of predictive hotspots for operational policing. We highlighted some of the limitations of AP ratio, a metric that is specifically designed to evaluate effectiveness of predictive hotspots. We then proposed a new metric called *Clumpiness Index* CI which is able to eliminate the limitations of AP ratio. This study first established that the two predictive methods used (i.e. Prospective Hotspot and KDE) are able to predict crimes well above the baseline predictions (random), and their performances are observed to vary according to the spatial concentration of different crime types. The CI was then used to provide more interpretable assessment of hotspot compactness which is found to be very robust to change in scales of spatial units of analysis. In addition, the CI calculation provided a baseline of comparison i.e. either of maximally aggregated (CI = -1) or maximally disaggregated (CI = 1) hotspot configuration, giving a sense how easy the hotspots can be covered by the police. As with AP ratio however, the CI being a single-valued metric requires visualisation of the purported hotspot to make perfect meaning out of it.

## 6. Acknowledgements

This research is part of the CPC (Crime, Policing and Citizenship) project supported by UK EPSRC (EP/J004197/1), in collaboration with the London Metropolitan Police.

## 7. Biography

Monsuru Adepeju is currently 2<sup>nd</sup> year PhD student at the SpaceTimeLab for Big Data Analytics, at University College London. His PhD focuses on predictive modelling of space-time hotspot of crime and his research interests include validation of crime predictive models for predictive policing and development of usable predictive tools for operational environment.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimeLab>), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

John S Shawe-Taylor is a professor at University College London (UK) where he is co-Director of the Centre for Computational Statistics and Machine Learning (CSML). His main research area is Statistical Learning Theory, but his contributions range from Neural Networks, to Machine Learning, to Graph Theory. He has coordinated a number of European wide projects investigating the theory and practice of Machine Learning, including the NeuroCOLT projects.

Kate Bowers is a Professor in Crime Science at the UCL Department of Security and Crime Science. Kate has worked in the field of crime science for almost 20 years, with research interests focusing on the use of quantitative methods in crime analysis and crime prevention.

## References

- Bowers K J, Johnson S D and Pease K (2004). Prospective hot-spotting the future of crime mapping? *British Journal of Criminology*, 44, 641-658.
- Telep C W and Weisburd D (2014). Hot Spots and Place-Based Policing. In *Encyclopedia of Criminology and Criminal Justice*, Springer New York, pp. 2352-2363.
- Turner M G (1989). Landscape ecology: the effect of pattern on process. *Annual Review of Ecology and Systems*, 20: 171–197
- Weisburd D (2008). Place-based policing, Ideas in American policing. *Police Foundation*, Washington, DC. <http://www.policefoundation.org/content/place-based-policing>. Accessed on 7/11/2014

# Exploring the geo-temporal patterns of the Twitter messages

Muhammad Adnan<sup>\*</sup>, Guy Lansley<sup>†</sup>, Paul A. Longley<sup>‡</sup>

Department of Geography, University College London, Gower Street, London, WC1E 6BT.

Email: [m.adnan@ucl.ac.uk](mailto:m.adnan@ucl.ac.uk), [g.lansley@ucl.ac.uk](mailto:g.lansley@ucl.ac.uk), [p.longley@ucl.ac.uk](mailto:p.longley@ucl.ac.uk)

November 07, 2015

## Summary

This paper explores the data recorded through the Twitter social media service. In particular we are interested in the analysis of the content of Tweet messages. A large corpus of Twitter messages was analyzed and Index of Dissimilarity measure was used to identify interesting words having spatial concentrations. The paper presents an initial exploration of the spatial and temporal pattern of the identified interesting words. At the finest geographical level, this type of analysis can gage very useful information to local planners in general and retail planners in particular.

**KEYWORDS:** Social Media, Geo-Temporal Analysis, Twitter, Content Analysis

## 1. Introduction

Recent years have seen an increased use of social media data as a cheaper alternative to more traditional methods of market research. Social media services generate a large quantity of data every day and some of the data is available through their Application Programming Interfaces (APIs). Social media services such as Twitter allow users to share information via short messages. These services are used not only for communicating with friends, family, and colleagues, but also for real-time news feeds and content sharing about venues (Pennacchiotti and Popescu, 2011). According to recent figures, the Twitter service has more than 200 million active users around the world (Twitter, 2012a). Its major user base is in European countries: in the context of the present paper, usage in the city of London, New York and Paris is the 3rd, 5th, and 7th highest in the world (Bennett, 2012). Twitter users generate a huge quantity of data every day, and our motivation here is to explore the geo-temporal patterns which exist in the text messages themselves. This paper presents an analysis of a large dataset of Twitter messages by the identification of a range of interesting words. Words were assigned to different categories and an initial exploration of the spatial and temporal pattern of the categories is presented. At the finest geographical level, this type of analysis can provide very useful information to local and retail planners.

Analysis of the social media content is a promising research area. Whilst past research on the Tweets' content has emphasized on exploring the sentiments users express in their messages, there has been limited attempts to link the geography of user generated topics across space to land use and activity. Some related work includes: the use of social media messages to classify areas into homogeneous groups (Birkin et al, 2013), the analysis of the personal information included in the tweet messages (Humphreys et al, 2013), historicizing Twitter within a longer historical framework of diaries

---

<sup>\*</sup> [m.adnan@ucl.ac.uk](mailto:m.adnan@ucl.ac.uk)

<sup>†</sup> [g.lansley@ucl.ac.uk](mailto:g.lansley@ucl.ac.uk)

<sup>‡</sup> [p.longley@ucl.ac.uk](mailto:p.longley@ucl.ac.uk)



(Humphreys et al, 2014), the content analysis of Tobacco-related Twitter posts (Myslín et al, 2012), and a forecasting model to predict the spread of a news (Naveed et al, 2011).

This paper is comprised of 5 sections. Section 2 of this paper describes the data used in the analysis. Data processing is described in the section 3, while section 4 and 5 present the results and conclusion.

## 2. Data

The Twitter Streaming API (Twitter, 2012b) can be used to download a 1% sample of the geotagged tweets. For this paper, the Twitter Streaming API was used to download geo-tagged Tweets for the Greater London during July to December, 2013. The fields downloaded from the API included the user name, latitude and longitude from which the Tweet was sent, time and tweet message content. A total of 4.6 million (4,633,139) geo-tagged Tweets were downloaded. These tweets were sent by a total of 272,248 unique users. Following map (Figure 1) shows a map of the 4.6 million tweets. This map shows that more Tweets were sent by users located in the central part of the city than the surrounding areas of Outer London.

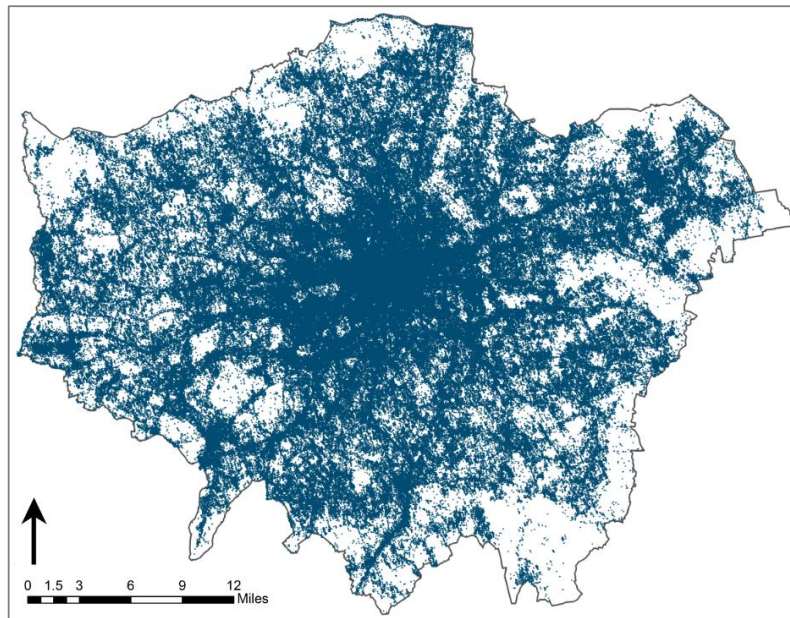


Figure 1: The Greater London geography of the 4.6 million tweets

Few users sent more tweets than others. 2,000 or more tweets were sent by the top 45 users. Following figure (2) shows the number of tweets by individual users.

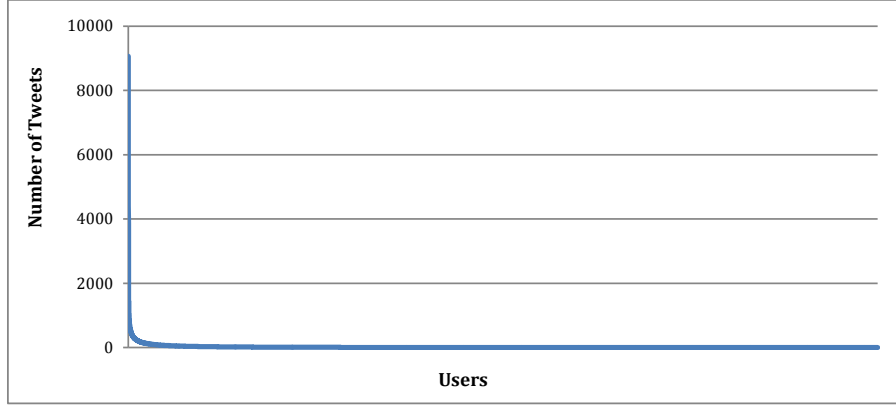


Figure 2: Number of tweets by individual users

### 3. Data processing

In the first step, 4.6 million tweets messages were divided into a series of ‘words’ i.e. a group of characters separated by a full-stop, comma, semi-colon, colon, apostrophe, or double quotes. This resulted in a dataset of 35,028,273 words. For the investigation of the spatial patterns of individual words, all the words were aggregated to 633 wards in the Greater London. For each word ( $y$ ), an Index of Dissimilarity (Birkin et al, 2013) was calculated across the 633 wards. The Index of Dissimilarity is defined in the following equation.

$$\theta(x, z) = 0.5 \times \sum_x \left| \frac{X_x^y}{X_*^y} - \frac{X_x^*}{X_*^*} \right|$$

Where  $x = (1, \dots, 633)$  wards in the Greater London and an asterisk (\*) denotes summation across a missing index. The resulting Index of Dissimilarity value for each word ( $y$ ) is a standardized value between 0 and 1. Where 0 indicates a uniform distribution and a 1 indicates a spatial concentration.

Index of Dissimilarity was calculated for each of the 35,028,273 words in the dataset. In the second step, in order to select the words which are spatially concentrated, the words having Index of Dissimilarity less than 0.5 were deleted from the database. This resulted in 122 remaining words which are listed in the following table (1). The table also assigns each word to one of the 8 distinct categories.

Table 1: 122 spatial concentrated words

Categories	Words
Travel	LHR, PANCRA, PADDINGTON, HEATHROW, RAILWAY, UNDERGROUND, FLIGHT, STATION, @HEATHROWAIRPORT, AIRPORT, TERMINAL, TUBE
Sports	#THFC, FULHAN, #ARSENAL, #LFC, #AFC, #ASHES, #CFC, @ ARSENAL, CHELSEA, SPURS, FOOTBALL, #MUFC
Places in London	HOUSNLOW, MARYLEBONE, MIDDLESEX, BROMLEY, GREENWICH, ISLINGTON, SHOREDITCH, OXFORD, PICCADILLY, WHARF, KINGSTON, SHARD, HACKNEY, BRIXTON, BRICK, MARKET, KENSINGTON, LEICESTER, KNIGHTSBRIDGE, CROYDON, HAMMERSMITH, CIRCUS, TOTTENHAM, WATERLOO, NOTTING, COVENT, REGENT, ARENA, WESTFIELD, ROMFORD, CAMDEN, RICHMOND, CLAPHAM, STRATFORD
Tourism	MUSEUM, TOWER, GALLERY, BRIDGE, PALACE, ROYAL, HOTEL, COURT, TRAFALGAR, HYDE, WESTMINSTER, ALBERT, BUCKINGHAM
Food & Drink	@STARBUCKSUK, STARBUCKS, COCKTAILS, BAR, COSTA, PUB, DRINK, COFFEE, JUICE, CAFE, MCDONALDS, COOKING, RESTAURANT

<b>Leisure</b>	LOUNGE, STUDIOS, THEATRE, PARK, EVENT, CINEMA, XFACTOR, KITCHEN, HOLIDAY, XBOX, HANGING, GARDEN, SHOPPING, MUSIC
<b>Emotions</b>	ENJOYED, #EXCITED, OMG, MISSING, SURPRISED, DISGUSTING, EMBARRASING, ANNOYING, GAY, MADNESS, WTF, FANTASTIC, SHOCKING, RIDICULOUS, BORED, AWFUL, HAPPINESS, PLZ
<b>Other</b>	GOODNIGHT, DUDE, DAD, DADDY, BOYS, FAMILY, FRIEND

#### 4. Results and Discussion

Following figure (3) shows an example of the spatial concentration of the words. This figure shows two maps of the individual tweets where ‘TRAFALGAR’ (map on the left) and ‘LHR’ (map on the right) were mentioned in the tweet messages. The Index of Dissimilarity value for both the words was 0.833 and 0.96 respectively, indicating a spatial concentration of the tweets. This could also be seen in the maps.

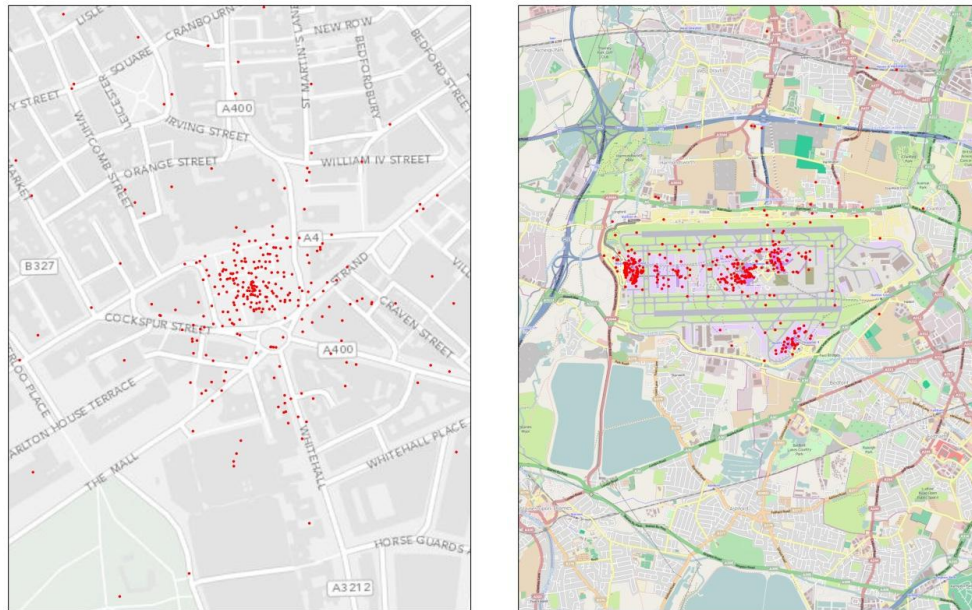
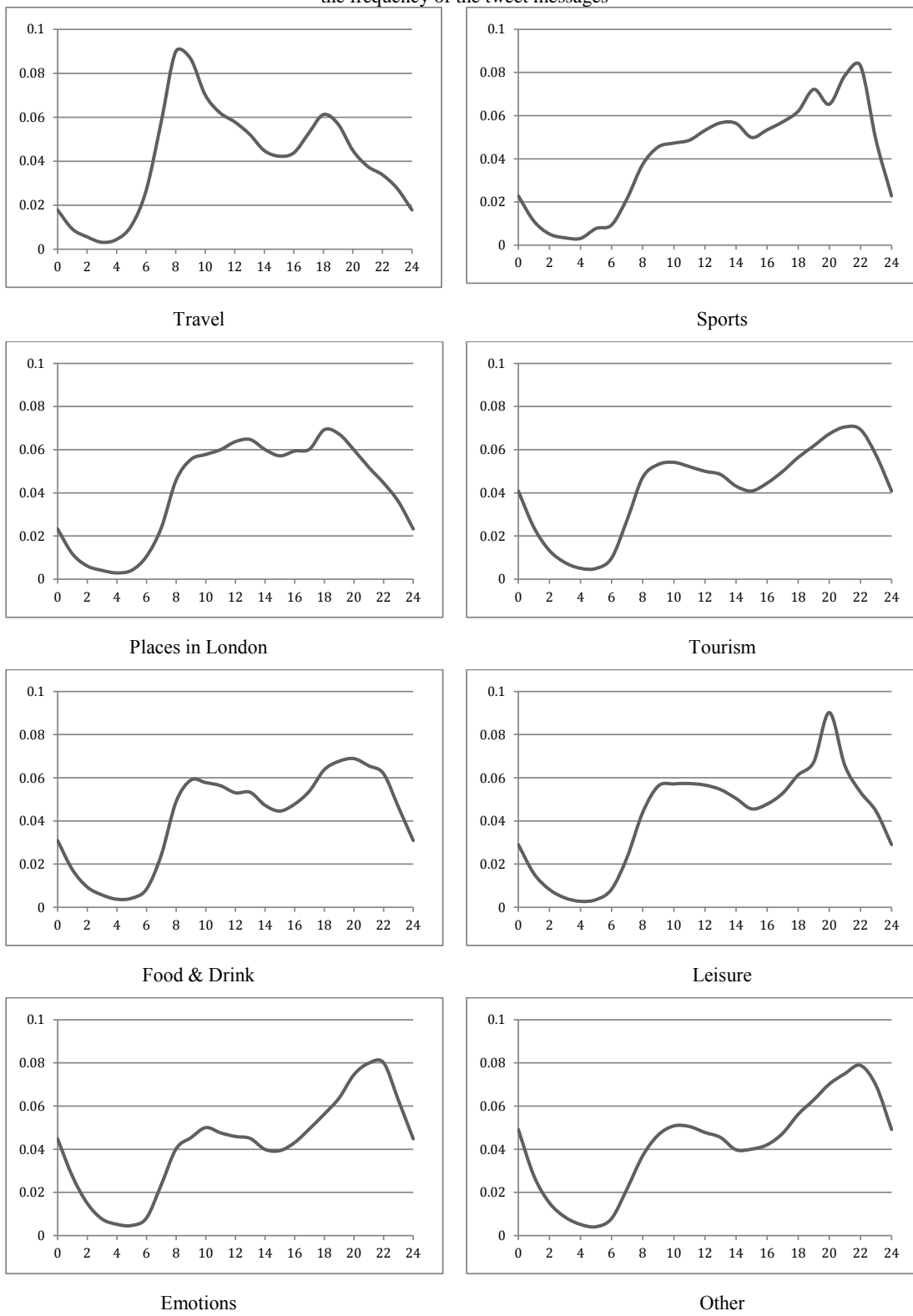


Figure 3: Tweets around the area of Trafalgar Square (left) and London Heathrow Airport (right)

The following table (2) shows the temporal graphs of the 8 word categories listed in section 3. The temporal graphs show the distinct temporal patterns of these categories. Words of the ‘Travel’, ‘Sports’, and ‘Leisure’ categories have the most distinct patterns. There is high number of tweets mentioning ‘Travel’ category words during the morning and evening rush hours. More tweets of the ‘Sports’ and ‘Leisure’ category words are sent during the night time. There are also more tweet mentions of the tourist places after 3pm during the day.

Table 2: Temporal graphs of the word categories. X-axis represents the hours of the day and Y-axis represents the frequency of the tweet messages



These temporal graphs show the footprints of the Twitter activity throughout the city. These also show an overall pattern of the behavior of the users in the Greater London.

## 5. Conclusion and future work

This paper has presented a preliminary analysis of the Twitter messages to explore the inherent spatial and temporal patterns of activity. A large dataset of Twitter messages was analyzed and decomposed into 35,028,273 words. For each word, the Index of Dissimilarity was calculated to identify interesting words having spatial concentrations. This resulted in a total of 122 words which were assigned to 8 distinct categories. The paper has also presented an initial exploration of the spatial and temporal pattern of the word categories.

This is a very promising research area, and we plan to enhance this work in the future. We plan to perform a fine scale temporal activity pattern analysis on the dataset to identify the areas of distinct attributes and behaviors e.g. the areas of leisure activities vs. work place areas. We also plan to use various topic unsupervised modelling techniques such as Latent Dirichlet Allocation (LDA) to generate topics in small geographical areas, and analysing the temporal variations in topic formulation and popularity, both daily temporally and seasonally.

## Acknowledgements

This work was completed as part of the ESRC research Grant "Retail Business Datasafe" (ES/L011840/1).

## References

- Bennet, S. 2012. Revealed: The Top 20 Countries and Cities of Twitter [STATS]. Retrieved 31st December, 2012, from [http://www.mediabistro.com/alltwitter/twitter-top-countries\\_b26726](http://www.mediabistro.com/alltwitter/twitter-top-countries_b26726).
- Birkin, M., Harland, K., Malleson, N. (2013). The classification of space-time behavior patterns in a British city from crowd-sourced data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7974, pp.179-192.
- Humphreys, L., Gill, Phillipa., Krishnamurthy, B. 2013. Historicizing New Media: A Content Analysis of Twitter. *Journal of Communication*, 63, 413-431.
- Humphreys, L., Gill, Phillipa., Krishnamurthy, B. 2014. Twitter: a content analysis of personal information. *Information, Communication & Society*. 17 (7).
- Myslín, M., Zhu, Shu-Hong., Conway, Michael. 2012. Content Analysis of Tobacco-related Twitter Posts. In the proceedings of the 2012 International Society for Disease Surveillance Conference.
- Pennacchiotti, M. and Popescu, A. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the Fifth International AAAI conference on Weblogs and Social Media*.
- Naveed, N., Gottron, T., Kunegis, Jérôme., Alhadi, Arifah Che. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In the proceedings of the WebSci'11. Koblenz, Germany. June 14-17, 2011.

Twitter. 2012a. What is Twitter ?. Retrieved 31st December, 2012, from <https://business.twitter.com/basics/what-is-twitter/>.

Twitter. 2012b. The Streaming APIs ?. Retrieved 22nd January, 2012, from <https://dev.twitter.com/docs/streaming-apis>.

## **Biographies**

Muhammad Adnan is a Senior Research Associate at Consumer Data Research Centre, University College London. His research interests are in data mining, social media analysis, and visualisation of large spatio-temporal databases.

Guy Lansley is a Research Associate at the Consumer Data Research Centre, UCL, an ESRC Data Investment. His previous research has included exploring the temporal geo-demographics derived from social media data, and identifying socio-spatial patterns in car model ownership in conjunction with the Department for Transport. Whilst, his current work entails exploring population data derived from large consumer datasets.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include 14 books and more than 125 refereed journal articles and book chapters. He is a former co-editor of the journal Environment and Planning B and a member of four other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.

# Participatory mapping for transformation: multiple visual representation of foodscapes and environment in informal settlements in Nairobi

Sohel Ahmed<sup>1\*</sup>, Muki Haklay<sup>2</sup>, Adriana Allen<sup>1</sup>, Cecilia Tacoli<sup>3</sup>, Edwin Simiyu<sup>4</sup>  
and Julio Davila<sup>1</sup>

<sup>1</sup>The Bartlett Development Planning Unit (DPU), University College London (UCL), London, UK

<sup>2</sup>Department of Civil, Environment & Geomatic Engineering (CEGE), UCL, London, UK

<sup>3</sup>International Institute for Environment and Development (IIED), London, UK

<sup>4</sup>Slum/Shack Dwellers International (SDI)- Kenya, Nairobi

## Summary

Although branded as ‘obstructionists’ and major agents of ‘disease and filth’ by city authorities, food vendors remain the pivotal node in the local food system in most informal settlements; therefore, their interaction with the environment and infrastructure services, and challenges they face to keep the food safe to eat, requires further grounded exploration. Food vendors from informal settlements in Nairobi, Kenya, who are acting as mappers and change agents, are building multi-layered views of places through the deliberative process of knowledge coproduction by participatory sensing, which lead to opportunities and challenges to improve those places.

**KEYWORDS:** Volunteered Geographic Information, food vending, Nairobi, participatory mapping, participatory sensing

## 1. Background

The households of the urban poor often rely on the food resources that are generated within the informal sector typical of many African urban centres, including Nairobi, Kenya (Tacoli, 2013). Yet these seemingly small-scale but significant numbers of vendors are not considered the ideal fit to modernist and elitist centred nature of planning and management of many cities in the Global South. They are, thus, often considered by the local authorities as- the ‘obstructionists’ as their stalls increase congestion in the very limited public spaces of the settlements; - and major agents of ‘disease and filth’ for demonstrating inadequate food safety measures, including poor storage facilities, often contaminated from road dirt, nearby waste dumps and open sewers. These vendors often suffer removal or forced closure by city authorities during disease outbreaks which not only put their livelihoods at risks but also affect access to food for the poorest residents of low-income settlements, who tend to be most dependent on street vendors (Keck & Etzold, 2013; Tacoli et al, 2013). Despite all these adversities, food vendors continue to be the pivotal node in the local food in most informal settlements (Tacoli et al, 2013); therefore, their interaction with the environment and infrastructure services, and challenges they face to keep the food safe to eat, requires further grounded exploration. Hence, we engage with food vendors in a few informal settlements in Nairobi, recognising their role as a major entry point for increasing urban food security and safety. Since vendors can both affect, and be affected by urban spatial structure/form, land-use and how infrastructure and services are provided, a crucial first step is to understand the physical constraints in the space within which street vendors operate. Thus the local communities have started conversation with us to explore how food-scapes (i.e. all types of food they eat) is connected with places where they live, work and walk within the settlements. To put it in another way, *how participatory mapping as a process and product, involving local participants, can contribute to and- from situated knowledge co-production when positioned to explore the environment-human-*

---

\* Sohel.ahmed@ucl.ac.uk

*food nexus?* This had opened up the need for multiple layers of data that are required for positioning ‘multiple ways of knowing’ the community food-scapes and their relation with the environment; on one hand, to capture the differential conditions within the settlements where people live, buy their food, and eat the food while walking (‘snack foods’) is very much part of their main meal as put by one of the participants-

“The way we eat in informal settlements has changed over time; this is because we lack adequate cooking spaces in our shanties and more so we are prone to fire outbreaks. This is why we prefer ready cooked food.”

-and on the other, to capture social construct and narratives around food-scapes that are unique to these informal settlements. The community realised that they need access to more innovative tools to be included to the existing repertoire of mapping and knowledge producing tools.

## **2. Bridging citizen science and Participatory GIS: multi-layered visual representation practices**

Use of GIS for data creation, analysis and dissemination of information has becoming ubiquitous across various disciplines because of its ability to bring more life to data by embedding it to a location/place/space, which also allow wider flexibility of visualisation in the policy making areas for multiple disciplines (Sieber, 2006). But since its journey in the 1980s, it primarily stayed as a tool for surveillance, control and authority – an expert power in the hands of a few advantaged- a very elitist and positivist tool and technique critiqued by many in the 1990s (Sieber, 2006; Cope and Elwood, 2009). To make the spatial platform more open to public, Public Participation GIS (PPGIS) evolved in the North primarily, which in the late-1990s transformed to give voice to marginalised communities in the Global South and evolved as Participatory GIS (PGIS) – a merger between participatory Learning and Action (PLA) and GIS, alternatively put ‘community GIS/mapping’ or ‘GIS-in-practice’ (Corbett et al, 2006).

Participatory GIS (PGIS) involves a collaborative process of using geospatial technologies in collecting and storing spatial data to have diverse perceptions and realities of space and place which includes collaborative collection of ‘field data’ that includes spatial data, and non-spatial qualitative data – e.g. community narratives representing local knowledge that is initiated and directed by the participants in the participatory development process of the Global South (Rambaldi, 2005, 2004, 2006). PGIS gave rise to a community of ‘Grassroot GIS users’<sup>2</sup>, promoting GIS and mapping practices that can situate and navigate the local initiative and knowledge for diverse reasons but for priorities important to the local community. Along with many other examples to PGIS and community mapping (e.g. Livengood & Kunte, 2012; Corbett et al., 2006; Makau et al., 2012), the Federation of the Kenyan Urban Poor (Muungano wa Wanavijiji) has already engaged in PGIS tools for enumerating and mapping urban informal settlements in Kenya (Kerenja, 2010).

We are using PGIS for mapping the issues that the community identified are supported by local and external experts and analysts in a collaborative way for collecting and analysing the data. We also adopt innovative way of multiple representation of places (as illustrated in Fig 1 and described below) by bridging citizen science<sup>3</sup> tools (mobile apps and balloon mapping) with PGIS tools to create appropriate community data and knowledge platform that ensemble all partial and situated knowledge through ‘multi-layered way of knowing’ in inclusive and empowering manner. Such –‘representational flexibility inherent in existing forms of the technology, creatively mixing and shifting representations, epistemologies, and signification strategies’ (Elwood, 2009, p60).

Cognitive ability of the community participants are used whenever possible to harness and situate local knowledge. For instance, community view on issues on food-scapes and environment are captured

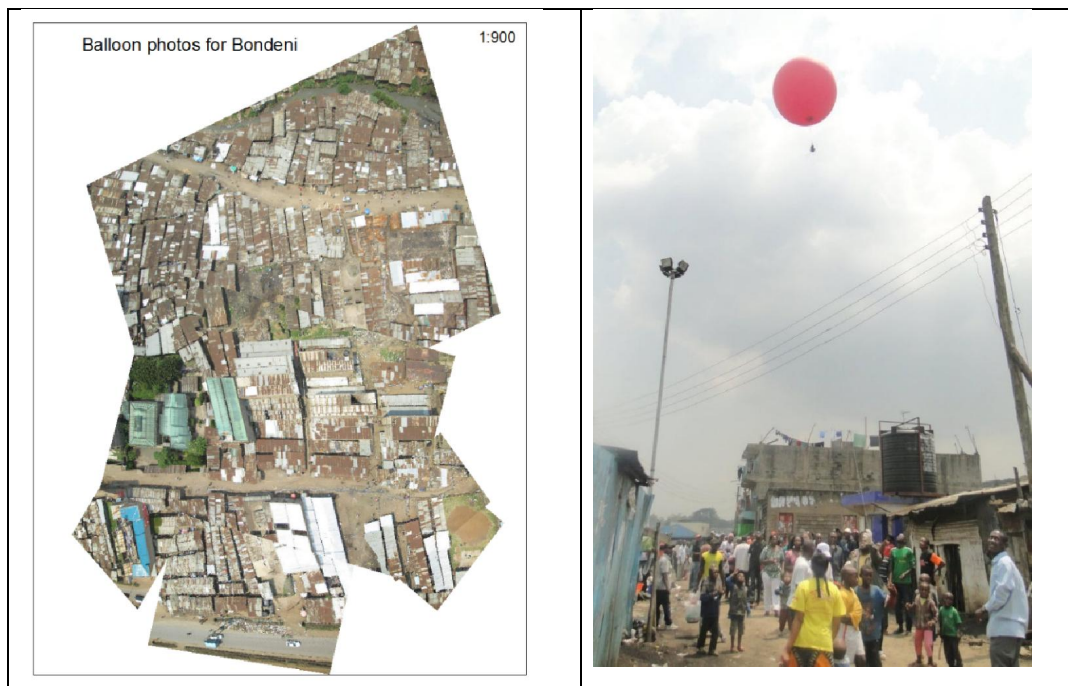
---

<sup>2</sup> Refers to the kinds of individuals and organisations that tend to be involved in PGIS initiatives: smaller NGOs, activist groups, community organisations’ (Elwood, 2009, p.59).

<sup>3</sup> ---‘ scientific activities in which non-professional scientists voluntarily participate in data collection, analysis and dissemination of a scientific project’ (Cohn 2008; Silvertown 2009 in Haklay, 2013)



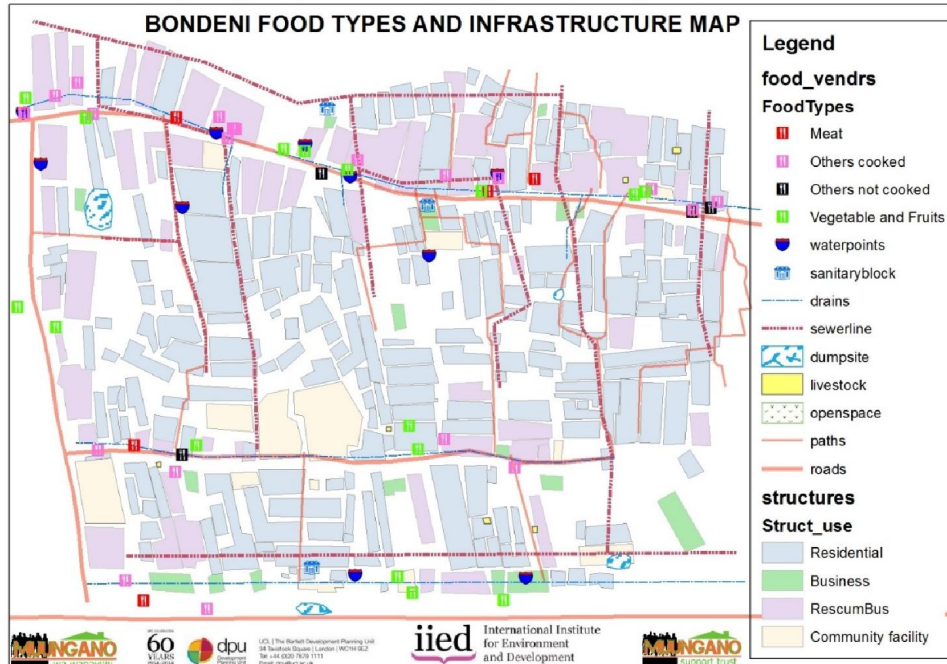
through narratives but cognitive mapping tools using satellite images and paper maps are gleaning local spatial knowledge from the community and helping them to identify locations that require mapping (on food and environment) on the ground. At this stage, the community is also using ‘participatory sensing’ techniques (Haklay 2013) that are now readily available in smartphones; The GPS and camera abilities are used of the mobile location-aware devices for mapping food vending types, in addition to general demographics of the vendors (age, sex), locations and food safety measures observed to have a rapid scan of food consumption sites in public spaces within the settlements as this is the predominant mode of consumption within such settlements in Nairobi, previously identified by Tacoli et al. 2013 and later confirmed by this study in community discussions. They are also using external sensors like balloon mapping tools (for more information on the technique, see publiclab.org) for generating cheaper DIY high resolution community aerial photos that provide different and unique bird-eye visual representations of the community, particularly having scalar view of environmental hazards (open sewerage lines, dumping sites, and so one), and also helps to update local base mapping and enumeration activities. Cognitive ability of the community participants are also being used to stitch the images to have settlement-wide images (Figure 1) and also to have a purview of scale of environmental problem they are facing.



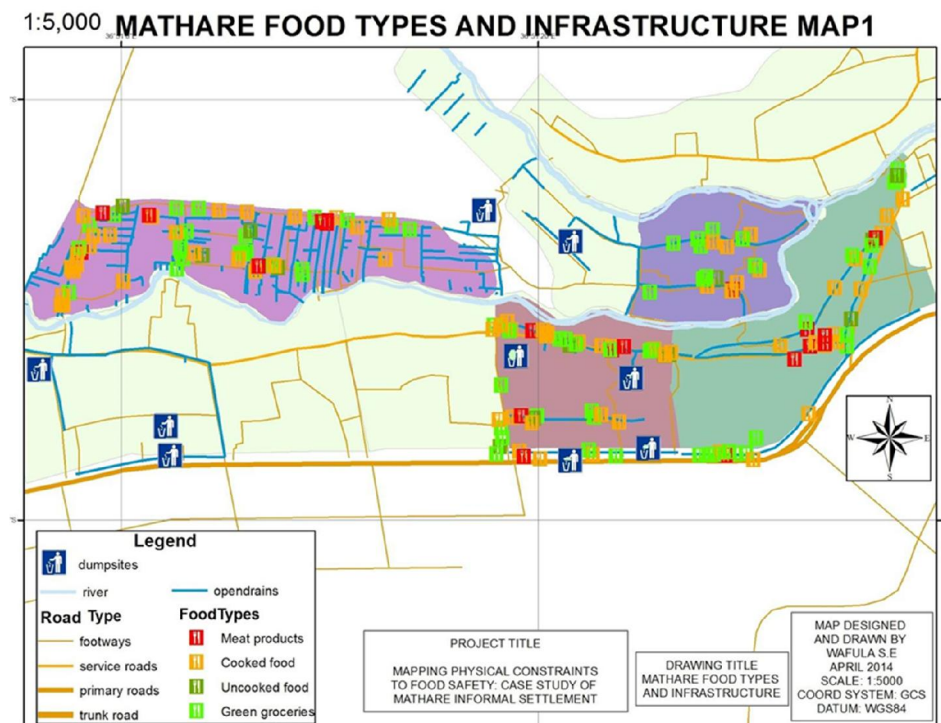
**Figure 1** visual representation of the same village captured through balloon mapping

We argue that the *raison d'être* behind such multi-representations are helping the community to frame and reframe the community narratives to situate to ‘representational practices’ (as coined by Elwood, 2009, p61) that are navigating and shifting to new priorities while also challenging existing meaning and identities embedded in those spaces. For example, the Federation has developed local base maps (as Mathare Zonal Plan<sup>4</sup>) highlighting inadequate infrastructure provision and thus showcasing and advocating the need for resource allocation with explicit pointers to areas that require more attention within this settlements; As food came to the fore of their agenda, the representation spaces and practices started to take shape around it which actually culminated in the study that this paper is referring to – food-environment-human interface (fig 2).

<sup>4</sup> <http://www.mustkenya.or.ke/index.php/settlement-zonal-plans/mathare-zonal-plan>



**Figure 2** Food-environment-human interface for the village, Bondeni in Mathere



**Figure 3:** Illustration of how food and infrastructure are interacting

We also argue that such fixed production of places through maps are ‘travelling’ as their causes are

changing. Using ‘expert power’<sup>5</sup> of GIS with other powerful visual representations like those from the ground and from the air linked to maps can only make the claim firmer.

### 3. Conclusion: knowledge that leads to action

Bridging Participatory GIS with citizen science tools such as food mapping with mobile apps and capturing high resolution community top-view with Balloon mapping with conventional GIS functionalities is allowing the community to have a deeper contextualisation than simple digital cartography cannot afford, and is also acting as a knowledge building tool, platform that empowers community. Using the narratives coming from community discussions are being translated into planning for immediate and future location-specific and settlement-wide interventions- e.g. settlement-wide awareness building by showcasing these multiple forms of visual representations ; with such multi-layers synoptic geographic overviews of settlements, communities/neighbourhoods are identifying hazardous areas in relation to food spaces and infrastructure provisions (road networks, water and sanitation provisions etc.) e.g. inadequate solid waste collection fosters food contamination(fig. 3), which is allowing the community to prioritise areas for clean-up and putting their priorities forward to local authorities. There are signs that this study is gathering quite a momentum as the community managed to get in touch with the Nairobi County Government with the preliminary findings which helped them to initiate Public-Private (PP) based solid waste collection effort as well as opportunity to consult the findings in a Parliamentary Committee. In other settlements in Nairobi, the communities are also forming food vendors association (FVA) like the one in Mathere which can make a big difference and can help in making the impact of the project sustained for longer.

### 4. References

- Corbett, J., Rambaldi, G., Kyem, P., Weiner, D., Olson, R., Muchemi, J., & Chambers, R. (2006). Overview: mapping for change—the emergence of a new practice. *Participatory learning and action*, 54(1), 13-19.
- Livengood, A., & Kunte, K. (2012). Enabling participatory planning with GIS: a case study of settlement mapping in Cuttack, India. *Environment and Urbanization*, 24(1), 77-97.
- Makau, J., Dobson, S., & Samia, E. (2012). The five-city enumeration: the role of participatory enumerations in developing community capacity and partnerships with government in Uganda. *Environment and Urbanization*, 24(1), 31-46.
- Karanja, I. (2010). An enumeration and mapping of informal settlements in Kisumu, Kenya, implemented by their inhabitants. *Environment and Urbanization*, 22(1), 217-239.
- .Sieber, R. (2006). Public Participation Geographic Information Systems: A Literature Review and Framework, *Annals of the Association of American Geographers*, 96:3, 491-507.
- Elwood, S. (2009). Multiple representations, significations, and epistemologies in community-based GIS. *Qualitative GIS: a mixed methods approach*. London: Sage Publications, 57-74.
- Sarah Elwood (2006). Negotiating Knowledge Production: The Everyday Inclusions, Exclusions, and Contradictions of Participatory GIS Research , *The Professional Geographer*, 58:2, 197-208.
- Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing Geographic Knowledge* (pp. 105-122). Springer Netherlands.
- Tacoli, C, Thanh, H., Owusu, M., Kigen, L. and Padgham, J. (2013). The role of local government in

---

<sup>5</sup> As other actors, particularly public and local government actors, usually treat ‘GIS-based data and maps as accurate representation of local conditions’, and take it more seriously as ‘expert or legitimate portrayal on the ground’(Elwood, 2009, p.70).

urban food security, Policy Briefing series, International Institute of Environment and Development (IIED), London.

Keck, M. & Etzold, B. (2013). Resilience refused: wasted potential of food security in Dhaka, *Erdkunde*, vol. 67 (1), pp. 75–91.

## 5. Acknowledgements

We acknowledge funding from the Department for International Development (UKAID). We also acknowledge the Medical Research Council, Natural Environment Research Council, Economic and Social Research Council, Biotechnology and Biosciences Research Council for the funding received for this project through the Environmental and Social Ecology of Human Infectious Diseases Initiative (ESEI), Grant Reference: G1100783/1.

## 6. Biography

**Sohel Ahmed** is a postdoctoral researcher at the Development Planning Unit (DPU) in the University College London (UCL), London, UK. His primary research interests revolve around urbanization, land-use change, social and spatial equity at the intersections of Planning, Environment, Development and Health, along with GIS and mapping applications for cross-disciplinary research.

**Mordechai (Muki) Haklay** is a Professor of Geographic Information Science at UCL, where he is also the co-director of the Extreme Citizen Science research group. His research focuses on usability engineering aspects of geospatial technologies, participatory mapping, citizen science and public access to environmental information.

**Adriana Allen** is Professor of Development Planning and Urban Sustainability at The Bartlett Development Planning Unit, UCL. She has over 25 years international experience in research and consultancy undertakings in Latin America, Africa and Asia. Both as an academic and practitioner, her work focuses on the interface between environmental justice and resilience in the urban global south.

**Cecilia Tacoli** is Principal Researcher and co-Head of the Human Settlements Group, International Institute for Environment and Development, London.

**Edwin Simiyu** is a program officer, community planning at SDI- Kenya/ Muungano Support Trust (MuST) and a surveyor by profession.

**Julio D. Dávila** is Professor of Urban Policy and International Development and Director of the Development Planning Unit, University College London. Much of his recent research work focuses on the role of local government in progressive social and political transformation in developing countries; the governance dimensions of urban and peri-urban infrastructure, especially public transport, and water & sanitation; the intersection between planning and urban informality; and the linkages between rapid urbanisation and health.

# CAN ADMINISTRATIVE DATA BE USED TO CREATE A GEODEMOGRAPHIC CLASSIFICATION?

Mildred Oiza Ajebon <sup>1</sup> and Paul Norman <sup>2</sup>

<sup>1</sup> Department of Geography, University of Durham, Durham, DH1 3LE, UK

Email: [m.o.ajebon@durham.ac.uk](mailto:m.o.ajebon@durham.ac.uk)

Tel: +447405890795

<sup>2</sup> Centre for Spatial Analysis & Policy, School of Geography, University of Leeds, Leeds, LS2 9JT, UK

Email: [p.d.norman@leeds.ac.uk](mailto:p.d.norman@leeds.ac.uk)

Tel:+44 (0)113 34 38199 Fax:+44 (0)113 34 33308

## Acknowledgements

This work used Census data obtained via MIMAS' CASWEB and GIS boundary data obtained via EDINA's UKBORDERS; services supported by ESRC and JISC. These data are Crown copyright and are reproduced with permission of OPSI. Adapted data from the Office for National Statistics and obtained via the Neighbourhood Statistics and NOMIS websites licensed under the Open Government Licence v.2.0 are also used. All data are subject to Crown Copyright.

## Abstract

This paper aims to contribute to the wider research scheme of the ONS 'Beyond 2011' project by assessing the feasibility of creating geodemographic classifications from administrative statistics as a way of eliminating the need for a full population survey. The classification is created using K-Means clustering algorithm which is then compared with OAC super-groups as a benchmark in maps and cross-tabulations. Results show similar classification of area types and health variations in England suggesting that the range of administrative datasets examined in this study could be explored as viable alternatives to the traditional census approach.

**Keywords:** Geodemographic Classification, Health Variation, Administrative Data

## 1.0 Introduction

The decennial census, though a complete source of socioeconomic data on UK population, has been criticized as being too costly and becoming increasingly difficult to carry out due to contemporary changes in the way society is organised. The challenges of high population mobility, opportunities created by advances in information technology which has increased the efficiency in the way data on several aspects of the population is stored, and the need for more timely and up-to-date delivery of

demographic data across the UK, all seem to suggest alternative ways of collecting, organising and disseminating detailed and up-to-date population data ONS (2011). Hence, the UK Statistical Authority (UKSA) “Beyond 2011” programme was set up to examine the feasibility of replacing the traditional census approach with administrative data already being held on the population. A logical starting point for the programme would be to search for similarities in patterns identified by both administrative and census datasets on a wide range of topics. This study therefore, aims to assess the potential of creating area classification models from available administrative data sources with the 2001 OAC Super-Groups as a benchmark. The geodemographic model is chosen for a pilot study because it is one of the most widely used socioeconomic models created from the decennial census for public sector planning and business targeting. The reliability of the classification is tested using independent data sets not included in classification.

The choice of testing the classification against selected census ill-health indicators is informed by the evidence in literature on the use of geodemographics to explain health inequalities health across England, The work of Dedman et al. (2006) demonstrated that geodemographic systems can be used to classify areas according to health needs by clearly showing where high and low illness rates might be expected. Shelton et al. (2006) developed a geodemographic characterisation of mortality patterns in England. Using sex and age standardised mortality data for nine causes of death, he calculated SIRs and found patterns of mortality to reflect socioeconomic circumstances with the more deprived areas suffering poorer health outcomes. Petersen (2009) also found that health inequality can be illustrated based on social area types. Thus, efforts have been made to show that strong relationships exist between population health and area types. However, no attempts at comparing geodemographic classifications created from administrative data sources as potential replacements of the census-based area classification were uncovered. This study represents one of the first efforts directed at filling this research gap. It examines the feasibility of constructing a geodemographic classification for small areas in England from administrative data sources. It examines how the classification compares with the census-based Super groups and whether the new classification can be used to predict geographical patterns of inequalities in health across England.

## **2.0 Data and Methods**

The primary scale of analysis chosen for this study are the 32,482 Lower Super Output Areas (LSOAs) of England being the geographies for which small area administrative statistics are published regularly to enable analysis of patterns over time (Neighbourhood-Statistics, 2004). The GIS boundary data for LSOAs are obtained from the United Kingdom Baseline Reference Database for Education and Research Study (UKBORDERS) available at EDINA (2012). It is worthy of note that the choice of variables for this study is highly limited by the availability of administrative statistics. All the datasets are derived from 100% administrative data sources and are a product of the National

Statistics. They are produced to a high statistical standard and accuracy. All datasets found to be negatively skewed were log-transformed ( $LN(Data + 1)$ ) to near normal distributions which has been found to be well adapted to socioeconomic count data (Rogerson, 2010). With the exception of council tax band as a proxy measure of housing, the strength of relationships between equivalent pairs of variables are shown to be strong in table 1. The classification of LSOAs into six socioeconomic groups as measured by available administrative data is created using the functionality of the SPSS k-means iterative clustering algorithm. Please refer to (Birkin and Clarke, 1998, Birkin and Clarke, 2009, Vickers and Rees, 2006, Vickers and Rees, 2007) for comprehensive details on creating geodemographic classification of areas. The alternative classification created from administrative data sources in this study is generally labelled Geo-Social Area Classification (GSAC). The profile names and pen portraits are derived from most dominant variables in each cluster.

JSAC	1									
Lone Parent	.674	1								
Council Tax Band	.537	.469	1							
Incapacity Benefit	.726	.641	.682	1						
Pension Claimants	-.127	-.079	.119	.104	1					
Census Unemployment	.865	.758	.551	.755	-.055	1				
Census Lone Parents	.649	.835	.555	.657	-.130	.709	1			
Census Rented	.614	.645	.512	.557	.067	.680	.658	1		
Census Pensioners	-.187	-.112	.057	.054	.970	-.093	-.194	-.025	1	
Census LLTI	.453	.435	.578	.746	.614	.543	.385	.424	.589	1
	JSAC	Lone Parent	Council Tax Band	Incapacity Benefit	Pension Claimants	Census Unemployment	Census Lone Parents	Census Rented	Census Pensioners	Census LLTI

**Table 1: Correlation matrix of the main census and administrative variables**

The correlations between equivalent pairs of deprivation variables are highlighted in beige. All correlations are significant at  $p=0.00$ . Incapacity Benefit (IB), Limiting Long Term Illness (LLTI); Job Seekers Allowance Claimant (JSAC);



### 3.0 Analysis

#### 3.1 Administrative data-based Area Classification of England

The classification created from administrative data is generally named ‘Geo-Social Area Classification (GSAC)’. Tables 2 shows profile labels and pen portraits of the six area types labelled after the dominant variables shaping the social character of the clusters. The map of the six clusters and the ONS Supergroups is shown in Figure 1.

Social profiles			
Area Type	Dominant Characteristics	Area	Dominant Characteristics
1: Struggling Families	Persons paying low council tax (bands B & C) Pension Claimants Aged 50 and over Paid care givers Incapacity benefit claimants	4: Suburbia	Resident population mainly elderly aged 60 and over Medium to high council tax payers (Bands D – G) Pension claimants
2: Typical Urban Living	Middle-aged resident population mainly older adults (aged 25-49) Paying council tax D- E High Population Density Lone Parents Job seekers	5: Outer Urban	High council tax payers Residents mainly aged 50-59
3: Deprived Communities	Persons paying low council tax (bands A) High children population (0-15) Job seekers Lone parents Incapacity benefit claimants Paid carers	6: Young Urban Families	Young resident population aged 16-24 High population density Low council tax (Band A)

**Table 2: Cluster labels and the variables defining the socio-economic characteristics of LSOAs in England**

A visual comparison of the maps in Figure 2 shows some spatial similarities between the ONS Super-Groups and the alternative GSAC. This inequality is well defined by both classifications in regions like the North-East, South Yorkshire, North-West and the Midlands. The ONS ‘Country-side’ and the GSAC ‘Suburbia’ which contain affluent LSOAs and the elderly population reflect more suburban distributions. The ethnic dimension is largely missed in the GSAC classification of urban areas due to the lack of readily available small area administrative data on ethnicity for inclusion in the area classifications at the time of this study. Overall, the GSAC appears to identify areas of socioeconomic disadvantage more distinctively. This is as expected since the main administrative variables available for inclusion in the classification relate to various types of economic deprivations. The ONS classification demonstrates a smoother pattern of socio-economic structuring of the population. This pattern is examined statistically by relating the classifications to independent observations to examine how they perform in reflecting socio-economic stratification.

### **3.2 Cross Tabulation-Based Comparisons of Geodemographic Classifications**

Tables 3a and 3b show the results of the cross tabulations of the ONS Super-Groups and the GSAC clusters which is a widely used objective method of determining the ecological equivalence of area classifications (Voas and Williamson, 2001, Webber and Butler, 2007). Each cell on the table shows the proportion of the total population share of LSOAs common to the subclasses of both classifications. The suburbia neighborhoods is classified as an equivalent of the Super-Group country side and urban fringe area types. The GSAC deprived communities are found to be similar to the ONS multicultural and disadvantaged groups. These clusters and most urban areas appear under-represented in the GSAC. Table 3b contains the index scores which quantify the degree of appropriateness of these proportions. An index score of 0 shows lack of representation and absence of the target cluster in the benchmark classification. 50 means that the target cluster is half represented as expected, 100 depicts equal representation on both classifications, an index value greater than 100 indicate an over-representation and 200 means that the target cluster is twice as represented in that order (Boyle et al., 2004). The indices shown in Table 3b indicate higher ecological correspondence between the urban clusters of both classifications. The GSAC appears to more clearly, distinguish areas of socio-economic disadvantage than the ONS Supergroups.

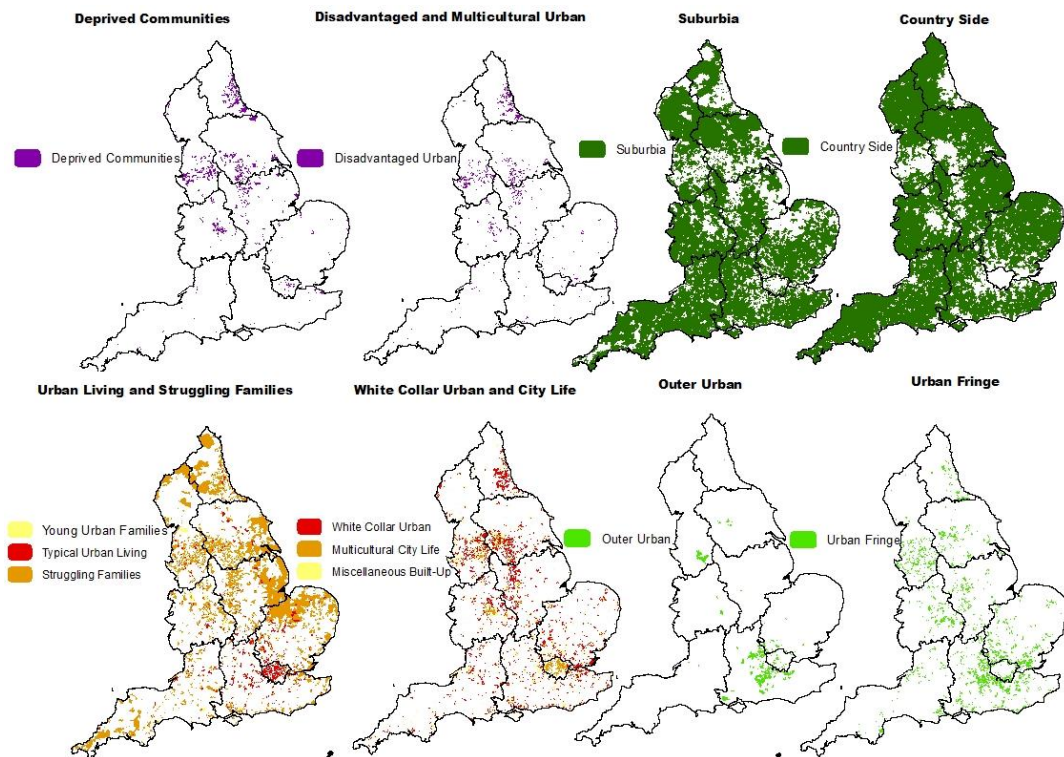


Figure 1: Clusters of ONS LSOA-level geodemographic classification Super-Groups compared with the alternative 'Geo-Social Area Classification' (GSAC)

COUNTS OF LSOAs		ONS SUPERGROUP TYPES FOR LSOAs							Total Number of LSOAs	% LSOAs
GEO-SOCIAL AREA CLASSIFICATION (GSAC)		1 Country Side	2 Professional City Life	3 Urban Fringe	4 White Collar	5 Multicultural City Life	6 Disadvantaged Community	7 Miscellaneous Built-Up		
	1: Struggling Families	780	212	609	3840	259	1063	3802	10565	32.5
	2: Typical Urban Living	65	1411	959	873	1789	98	850	6045	18.6
	3: Deprived	11	53	0	86	1457	3236	898	5741	17.7
	4: Suburbia	3120	387	3449	1440	73	1	536	9006	27.7
	5: Outer Urban	59	202	308	0	0	0	1	570	1.8
	6: Young Urban Families	16	400	28	7	47	0	57	555	1.7
Total Number of LSOAs		4051	2665	5353	6246	3625	4398	6144	32482	100.0
% LSOAs		12.5	8.2	16.5	19.2	11.2	13.5	18.9	100.0	

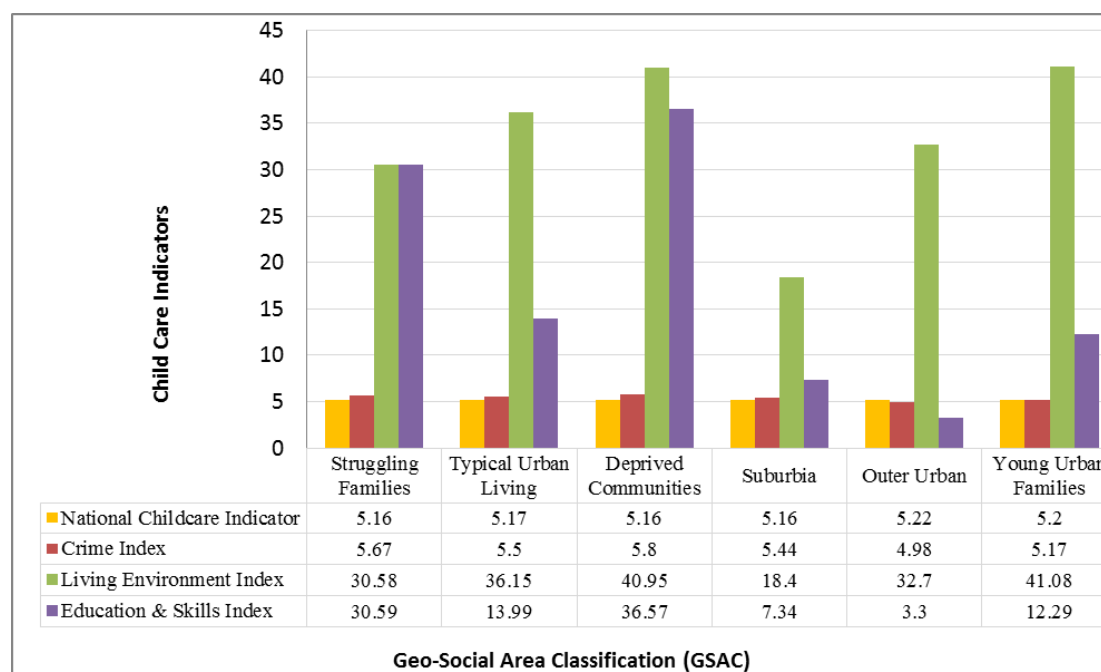
INDICES		1	2	3	4	5	6	7	
	Population Share	Country Side	Professional City Life	Urban Fringe	White Collar	Multicultural City Life	Disadvantaged Community	Miscellaneous Built-Up	
	1: Struggling Families	0.325257	59 (10.0)	24 (4.1)	35 (5.9)	189 (32.8)	22 (3.7)	74 (12.5)	190 (32.0)
	2: Typical Urban Living	0.186103	9 (1.1)	284 (34.9)	96 (11.8)	75 (9.3)	265 (32.5)	12 (1.5)	74 (9.1)
	3: Deprived	0.176744	2 (0.2)	11 (1.5)	0 (0)	8 (1.0)	227 (30.4)	416 (55.7)	82 (11.1)
	4: Suburbia	0.277261	277	52 (7.7)	232	83 (12.1)	7 (1.1)	0 (0)	31 (4.6)
	5: Outer Urban	0.175482	8 (9.8)	43 (51.1)	32 (38.9)	0 (0)	0 (0)	0 (0)	0 (0)
	6: Young Urban Families	0.170864	2 (2.2)	88 (82.2)	3 (2.9)	1 (0.6)	8 (7.1)	0 (0)	5
	Overall Performance		129	100	105	89	138	143	90

Table 3 Cross tabulation results

Table 3A: is Cross tabulations of the Clusters of ONS LSOA-level geodemographic classification Super-Groups and those of the alternative 'Geo-Social Area Classification' (GSAC) of England; 3B shows the index scores derived from the cross tabulation results. An index score of 100 means that the a cluster cell has the same population share of LSOAs, 0 means cluster is absent and 50 signifies a cluster is half as present and 200, twice as present, in that order

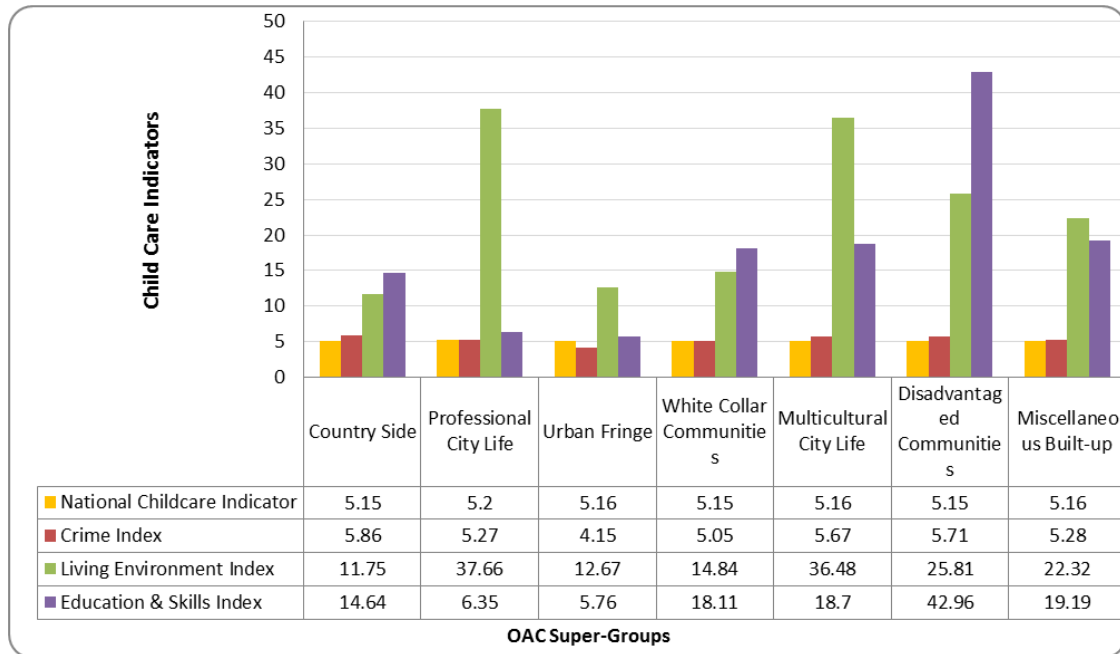
### 3.3 Validating the Classifications with Independent Datasets

The GSAC and the ONS Super-Groups classification performance were examined in relation to the averages of the National Childcare Indicator data (NCI), three indicators used for the construction of the 2010 IMD namely Crime, Living Environment and Education and Training for each area type in London GOR. The results in Figures 2 and 3 show a clear stratification of these indicators along socioeconomic lines. In both classifications, higher proportions of low income groups in urban centres are more likely to take up the formal childcare element of the working tax credit (Gregory, 2009) in comparison with families in more rural areas and elderly populations in London. As expected, crime incidence is demonstrated to be relatively higher in socially disadvantaged and multicultural LSOAs of London located in communities such as Tower Hamlets and Westminster compared with suburban Greenwich and Barnet neighbourhoods. Poorer environmental conditions, educational qualification and professional skills are found to increase with area-level deprivation. Families in suburban communities are shown to live in better socioeconomic conditions. Though similar patterns of socioeconomic stratification are identified by both classifications, the ONS Super-Group which is constructed from a wider range of variables appears to illustrate small area-level socioeconomic stratification of areas more distinctively.



**Figure 2: Geo-Social Area Classification (GSAC) Clusters cross tabulated with other socioeconomic data not included in the classification**

Lower values of National Child care indicator implies low take up of formal childcare tax credit and higher values refer to higher take up. Lower values of crime, living environment and education mean better conditions and higher values are indicative of poorer conditions



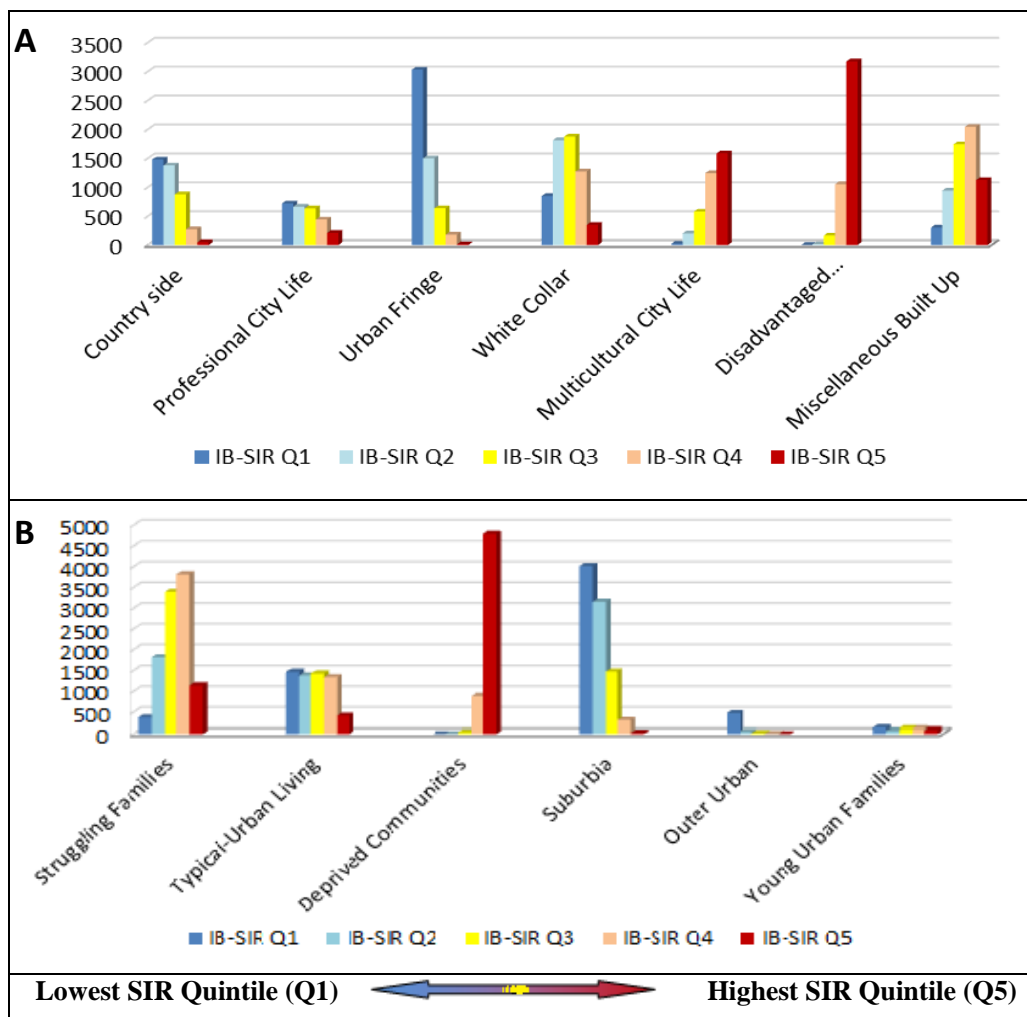
**Figure 3: AC-Super Groups cross tabulated with other socioeconomic data not included in the classification**

Lower values of National Child care indicator implies low take up of formal childcare tax credit and higher values refer to higher take up. Lower values of crime, living environment and education means better conditions and higher values are indicative of poorer conditions

### 3.4 Geodemographic Classifications and Health Inequality

The created classifications was also related to quintiles of health indicators from the 2001 census to examine how well health inequalities in England could be identified. The results show health outcomes to reflect the social characteristics of neighbourhoods and the people who live within them at the LSOA level. Table 4 shows the area correspondence of ONS Super-Groups and GSAC Clusters with LLTI and IB health measures for the year 2001. The cells with values highlighted in blue represent LSOA types with high proportions of good health outcomes. The cells in red are area types demonstrating high levels of ill-health. Areas highlighted in yellow represent the proportions of typical urban areas. The distribution of the quintile of health ratios across geodemographic clusters is better visualized in Figure 4, which clearly shows the more deprived/disadvantaged/multicultural areas as having poorer health outcomes (these are represented with red bars) compared with the affluent suburban groups with lower proportions of ill-health, depicted with blue bars. The presence of rural-urban differentials in health patterns is also clearly seen Core urban areas (ONS White Collar Communities and GSAC Typical Urban

Living groups) appear to reflect higher inequality. Most LSOAs in urban areas contain a complex mix of the first four illness quintiles almost in equal proportions compared to more rural ones. Note that the proportions of LSOAs within core urban neighbourhoods in the highest illness quintile (Q5) are relatively small compared with other neighbourhoods. These areas have high concentrations of younger professional adults who are less likely develop critical health issues. Overall, the results suggest that geodemographic classifications can be a more practical tool for explaining geographical variations in health.



**Figure 4: Quintiles of health measures by geodemographic typologies in England**

The height of the bars represents the count of area types represented in a particular quintile of health. Blue bars represent better health (Q1 and Q2), red bars depict worse health (Q4 and Q5) and yellow bars are average health (Q2)

#### **4.0 Conclusion**

The challenges of data and methodological limitations of the K-Means clustering algorithm, the findings of the study demonstrate national administrative statistics used for creating the geodemographic classification to be of high performance given the strong associations between the datasets and equivalent census measures. The alternative area classification labelled 'GSAC' was found to stratify LSOAs into similar area types with the ONS Super-Groups. A high level of ecological correspondence was observed between the urban clusters of both classifications. Deprived communities of the GSAC area types appear to be clearly mapped out in a similar fashion with census definitions. This is expected given the heavy reliance of the classification on benefit data and council tax bands. The test of the classification against independent child health indicators for London not used in the both classifications further confirms the similarity of the GSAC with the OAC Super groups. In both classifications, poorer health, worse living environment index and higher crime rates are observed in more disadvantaged LSOAs while more affluent neighbourhoods record better health, improved living environment conditions and much lower records of crime.



## References

- Birkin, M. & Clarke, G. 1998. Gis, Geodemographics, And Spatial Modeling In The Uk Financial Service Industry. *Journal Of Housing Research*, 9, 87-111.
- Birkin, M. & Clarke, G. 2009. Geodemographics. *The International Encyclopaedia Of Human Geography*, 382-89.
- Boyle, P., Norman, P. & Rees, P. 2004. Changing Places. Do Changes In The Relative Deprivation Of Areas Influence Limiting Long-Term Illness And Mortality Among Non-Migrant People Living In Non-Deprived Households? *Social Science & Medicine*, 58, 2459-2471.
- Dedman, D., Hennell, T., Hooper, J. & Tocque, K. 2006. Using Geodemographics To Illustrate Health Inequalities. *Liverpool: North West Public Health Observatory, Liverpool John Moores University*.
- Gregory, I. 2009. Childcare Takeup And National Indicator 118: A Summary Of Learning Funded By Government Regional Offices 08/09. In: Department For Children, S. A. F. (Ed.). London: Government Office For London. . [Accessed 20th August, 2012]. Available From: [http://www.daycaretrust.org.uk/Data/Files/Consultancy/Childcare\\_Take\\_Up\\_And\\_National\\_Indicator\\_118.Pdf](http://www.daycaretrust.org.uk/Data/Files/Consultancy/Childcare_Take_Up_And_National_Indicator_118.Pdf).
- Neighbourhood-Statistics 2004. Super Output Areas (Soas): Frequently Asked Questions. Office For National Statistics. [Accessed 20th July, 2012]. Available From: <http://www.neighbourhood.statistics.gov.uk/Dissemination/Info.Do;Jessionid=Gqcqqlrlnvtgz1tgbflvnlthkvhyOcclhksyjqj8fd1pv1d0gkd!1949496690!1342738827103?M=0&S=1342738827103&Enc=1&Page=Aboutneighbourhood/Geography/Superoutputareas/Soafaq/Soa-Faq.Htm&Nsjis=True&Nsck=True&Nssvg=False&Nswid=1366>.
- Ons 2011. Beyond The 2011 Census Project. Office For National Statistics. [Accessed 13th June, 2012]. Available From: <http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html>.
- Petersen, J. 2009. *Social Marketing And Public Health*. Ucl (University College London).
- Rogerson, P. A. 2010. *Statistical Methods For Geography: A Student's Guide*, Sage Publications Ltd.
- Shelton, N. J., Birkin, M. H. & Dorling, D. 2006. Where Not To Live: A Geo-Demographic Classification Of Mortality For England And Wales, 1981-2000. *Health & Place*, 12, 557-569.
- Vickers, D. & Rees, P. 2006. Introducing The Area Classification Of Output Areas. *Population Trend-London*, 125, 15.
- Vickers, D. & Rees, P. 2007. Creating The Uk National Statistics 2001 Output Area Classification. *Journal Of The Royal Statistical Society: Series A (Statistics In Society)*, 170, 379-403.
- Voas, D. & Williamson, P. 2001. The Diversity Of Diversity: A Critique Of Geodemographic Classification. *Area*, 33, 63-76.
- Webber, R. & Butler, T. 2007. Classifying Pupils By Where They Live: How Well Does This Predict Variations In Their Gcse Results? *Urban Studies*, 44, 1229-1253.

# A geospatial relational database schema for interdependent network analysis and modelling

David Alderson<sup>\*1</sup>, Stuart Barr<sup>†1</sup>, Tomas Holderness<sup>‡2</sup>,  
Craig Robson<sup>^1</sup>, Alistair Ford<sup>#1</sup> and Ruth Kennedy-Walker<sup>::1</sup>

Corresponding author: David Alderson ([david.alderon@ncl.ac.uk](mailto:david.alderon@ncl.ac.uk))

<sup>1</sup>School of Civil Engineering and Geosciences  
Cassie Building, Newcastle University, NE1 7RU, UK

<sup>2</sup>SMART Infrastructure Facility  
Faculty of Engineering and Information Sciences  
University of Wollongong, NSW, 2522, Australia

## Summary

Services delivered via National Infrastructure (NI) are key to securing economic growth and societal well-being. Spatially complex interdependent networks form an integral component of NI (e.g. energy supply, transport, waste management, clean water supply and dirty water treatment). It is essential that such infrastructure networks and interdependencies can be managed, analysed and modelled in a robust and consistent manner. This paper presents work undertaken to develop a spatial interdependent network model within existing relational database management software that is suitable for national-scale representation of infrastructure network systems and their interdependencies.

**KEYWORDS:** National infrastructure, networks, relational databases, data management, interdependency.

## 1. Introduction

National Infrastructure (NI) networks no longer operate as isolated, stand-alone spatial networks, but interact to form complex relationships between physical assets and networks. Spatial dependencies and interdependencies between NI networks can lead to the propagation of disruptions and failures resulting from man-made and natural hazards. As such, it is pertinent to consider NI as a ‘network-of-networks’, with inherent underlying spatiality. However, such a representation can require large quantities of data from disparate sources to represent physical components of single and multiple networks. Moreover, new data management and analysis tools are required in order to effectively manage and analyse infrastructure systems within the network-of-networks paradigm. In this paper we present a new relational database schema, implemented in PostgreSQL and PostGIS, that has been developed specifically for representing spatially interdependent infrastructure networks. The utility of the database schema is demonstrated via a prototype reporting tool for infrastructure network visualisation.

---

\* [David.Alderson@ncl.ac.uk](mailto:David.Alderson@ncl.ac.uk)

† [Stuart.Barr@ncl.ac.uk](mailto:Stuart.Barr@ncl.ac.uk)

‡ [Tomas@uow.edu.au](mailto:Tomas@uow.edu.au)

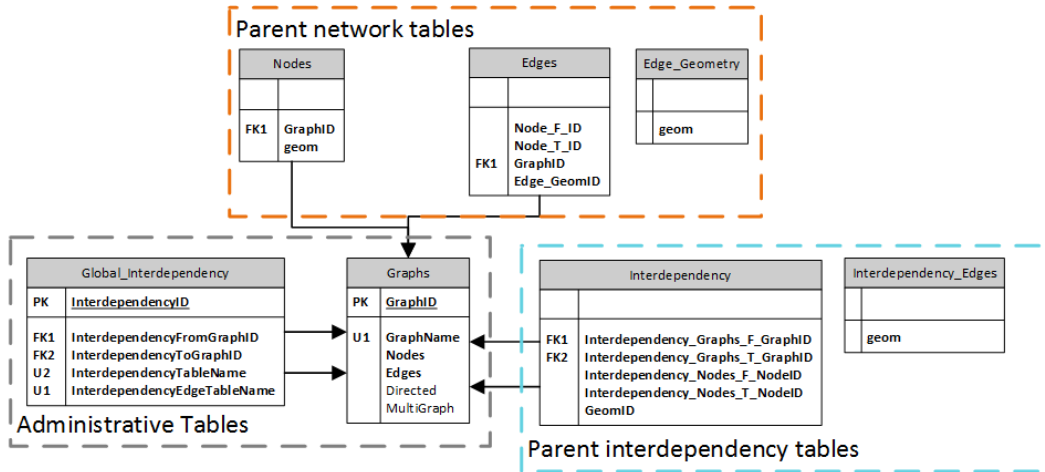
^ [C.A.Robson1@ncl.ac.uk](mailto:C.A.Robson1@ncl.ac.uk)

# [Alistair.Ford@ncl.ac.uk](mailto:Alistair.Ford@ncl.ac.uk)

:: [Ruth.kennedywalker@gmail.com](mailto:Ruth.kennedywalker@gmail.com)

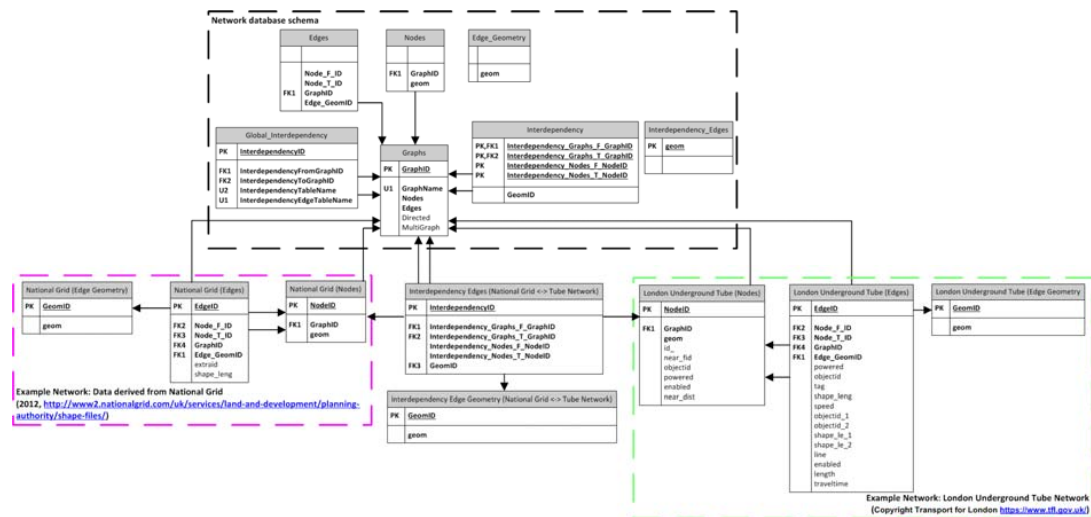
## 2. Network database schema and Python interface

The structure of the network database schema is illustrated in **Figure 1** as an entity-relationship (ER) diagram. The administrative tables, *Graphs* and *Global\_Interdependency* (**Figure 1**) maintain a record of the networks and interdependencies stored within the schema. A network is stored using the concept of table inheritance. A single network comprises three tables, with each inheriting attributes from the *parent* tables (**Figure 1**); ensuring that the attributes of the *parent* tables propagate to the *instance* tables. This same approach is used for the interdependency links between two networks.



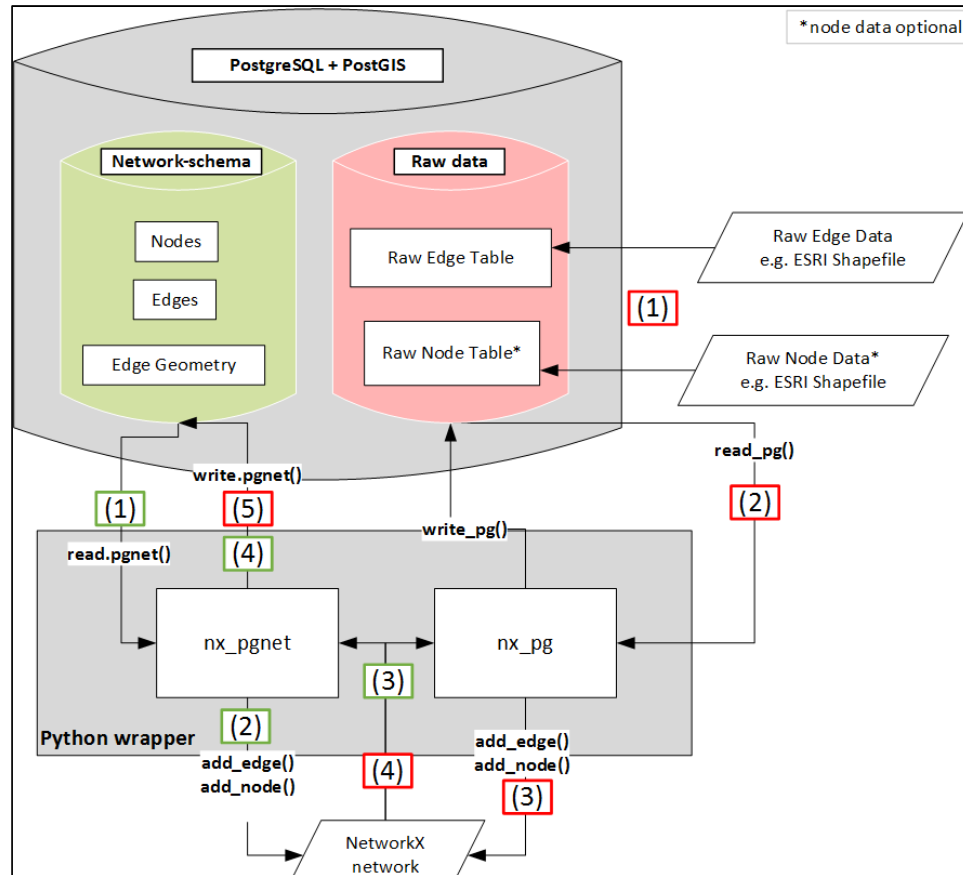
**Figure 1** – ER diagram of the network database schema.

**Figure 2** shows two networks stored within the network database schema; the tables outlined in magenta represent a network derived from National Grid electricity network data while the tables outlined in green represent a network forming the London Underground Tube Network.



**Figure 2** –ER diagram of two dependent networks stored in the database schema.

To construct and store the two networks, two Python-based modules, *nx\_pg* and *nx\_pgnet* have been developed that use NetworkX to build the topology of any network model (**Figure 3**). The *nx\_pg* module is used to convert the raw data into a NetworkX network using the *read\_pg()* function, and then the *nx\_pgnet* module can be used to write this network to a set of schema-enabled tables back to the same database, using the *write\_pg\_net()* function. The process of building from raw data is illustrated in **Figure 3** by the steps highlighted in red, whilst the process of extracting a network from the schema, once built, is highlighted in green. The build process uses the *add\_node()* and *add\_edge()* methods of NetworkX to add the geometry and attributes of the raw edge and node data to a NetworkX network before the final result is then written to the database schema. These modules therefore act as a linking mechanism between the schema used for storage, and NetworkX which is used for network analysis.

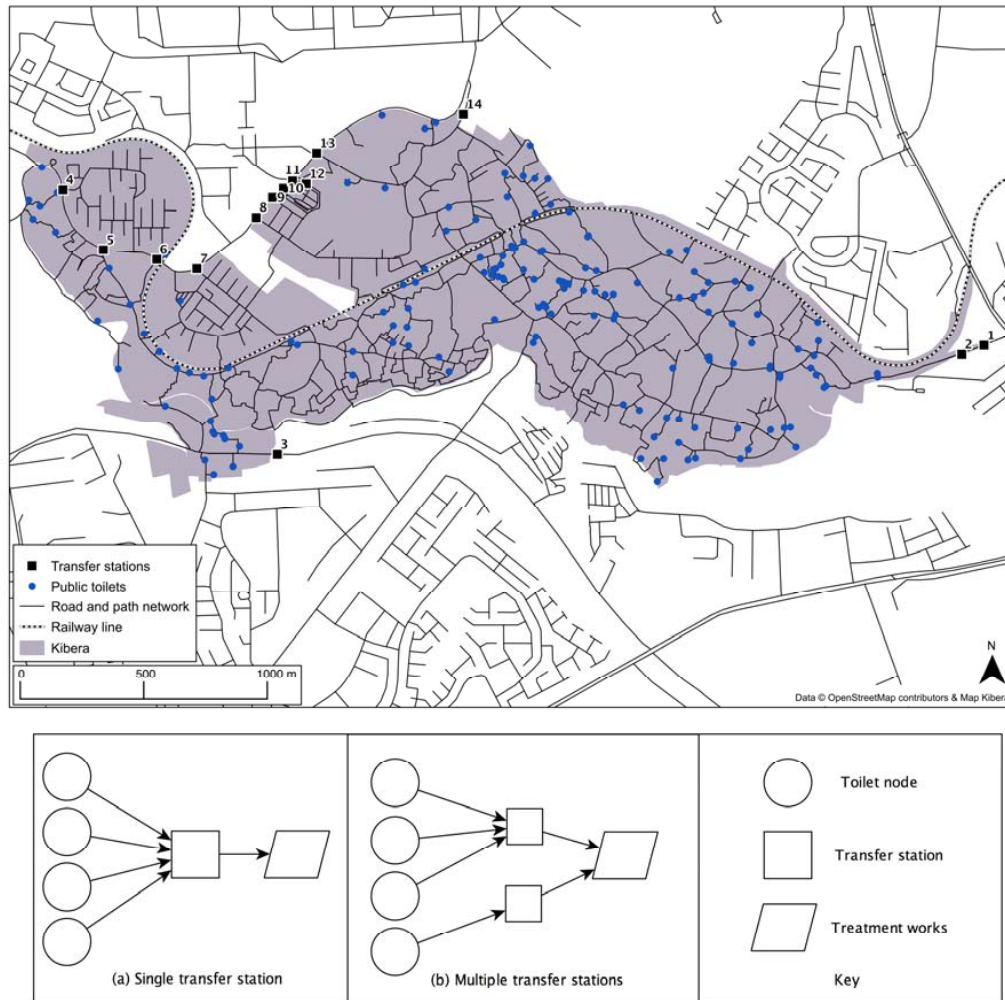


**Figure 3** – Overview of Python modules *nx\_pg* and *nx\_pgnet* that act as an interface between the network database schema and NetworkX.

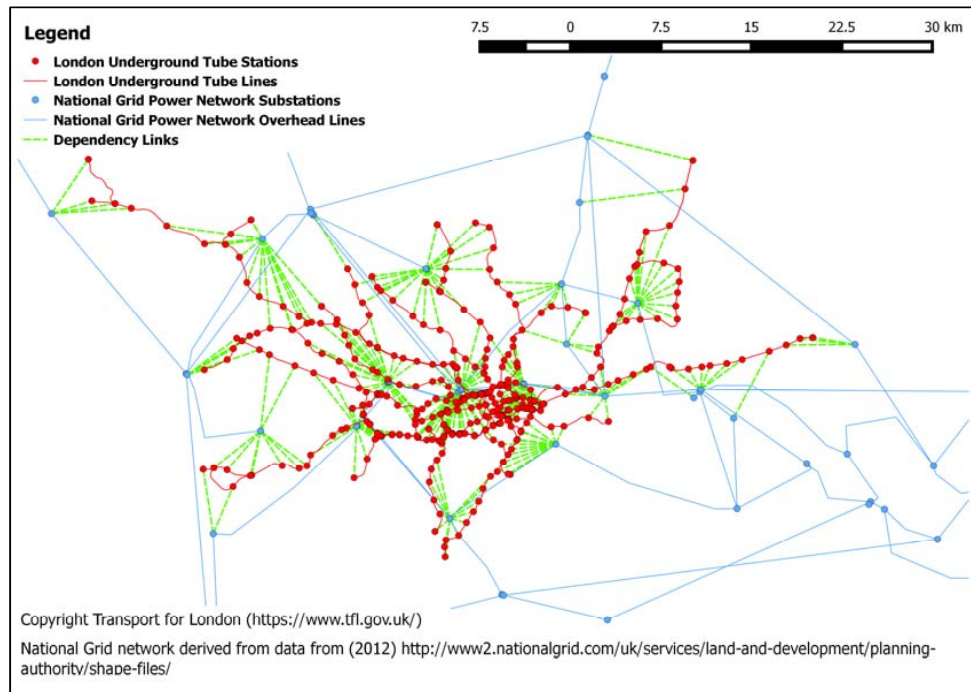
### 3. Enabling network analysis and visualisation

To date, the database schema has been used to facilitate the modelling and subsequent analysis of infrastructure networks at differing spatial scales, for different geographic regions, within different specific application domains. For example, the construction and storage of a road network from informal data sources within the database schema, and subsequent linkage to NetworkX via the Python wrappers, formed a critical component of work assessing optimum faecal sludge removal and disposal over a road network for a suburb in Kibera, Nairobi, Kenya (Kennedy-Walker et al., 2014). This work focussed on assessing two approaches for transferring faecal sludge between public toilets and transfer stations; namely via a single transfer station, or via multiple transfer stations (see (a) and (b) of **Figure**

4), using the network database schema as the data storage and analysis platform linked to NetworkX. This provided the facility to perform cost analyses based around shortest paths across the road network for the two different transfer approaches, between the toilets and transfer stations. In addition to its analytical utility, two database-level functions, *ni\_create\_node\_view*, and *ni\_create\_edge\_view*, create *views* of the network data allowing the schema to be directly linked to Quantum GIS (Quantum GIS, 2014). This functionality is shown in **Figure 5** where the two networks in Figure 2 (National Grid electricity transmission and London underground tube network) are shown along with a representation of inferred spatial dependencies between these two networks.



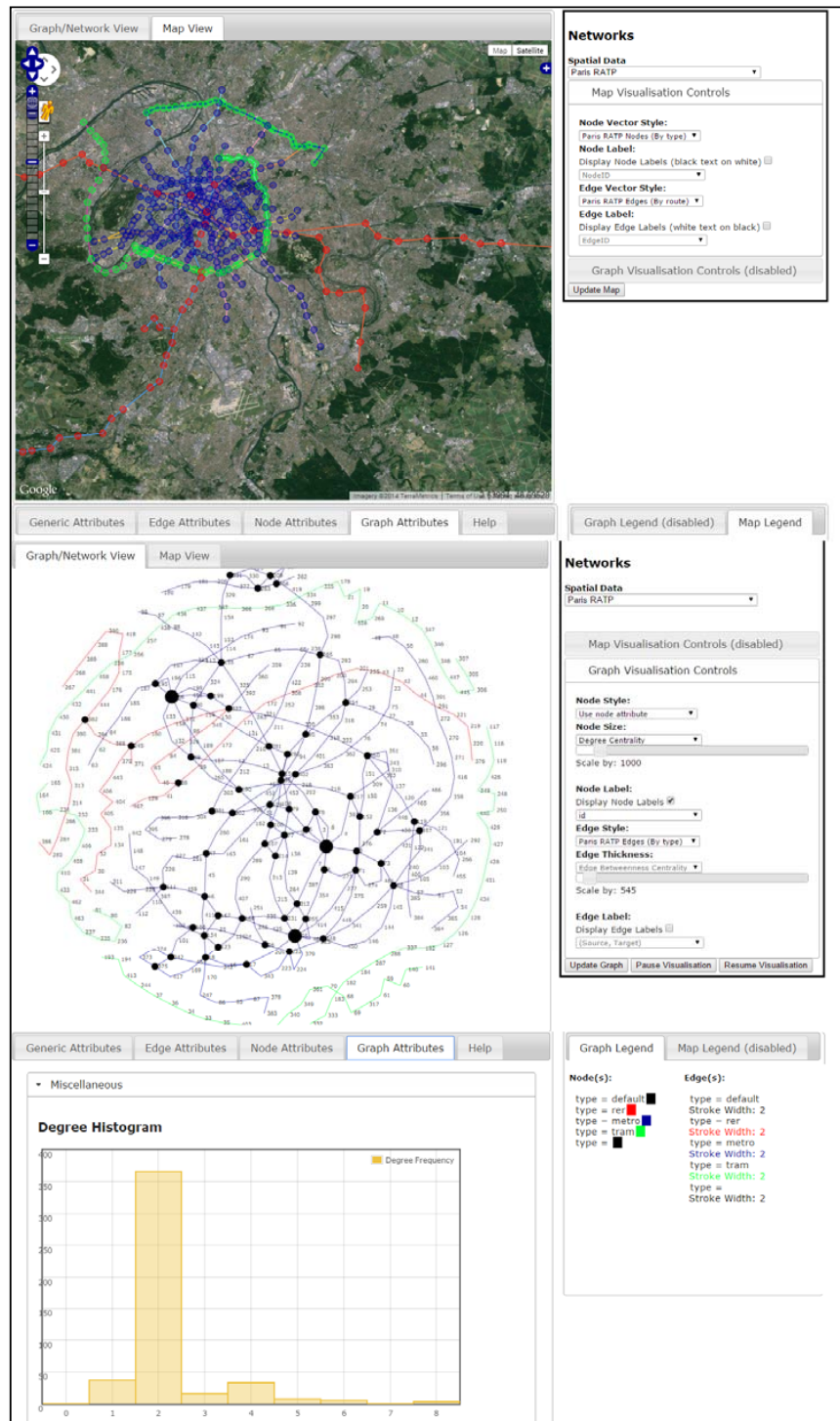
**Figure 4** – Road network for Kibera suburb, Nairobi, Kenya (Kennedy-Walker et al., 2014).



**Figure 5** – QGIS linked view of the two networks and their spatial dependencies shown in **Figure 2**.

As the encoding of a network or set of interdependent networks contains both topographical and topological information a network visualisation and analysis tool has been developed for the database schema using JavaScript Object Notation (JSON) data format that allows both views (topographic and topological) along with analytical metrics to be viewed. **Figure 6** illustrates several components of this visualisation tool. In the tool a user is able to select a particular network its style, and switch between the topological and topographical views. The topographical view is currently delivered using the OpenLayers web mapping library (OpenLayers, 2014a), and uses the OpenLayers Style Map (OpenLayers, 2014b) to encode styles for each network. The topological view was developed using the D3 library (Bostock, 2014), and provides currently a single force-directed layout (Bostock, 2012) to visualise each network. **Figure 6** also illustrates the metric view which is derived by using the functionality of NetworkX.





**Figure 6** – Example of views of the topographic, topological and metric based capabilities of the database schema analysis and visualisation tool. **Top:** RATP Rail Network, **Middle:** topological view, **Bottom:** metric view as a degree histogram.

#### 4. Conclusion

This paper has highlighted the key components of a network database schema developed to facilitate NI network modelling and analysis within a network-of-networks paradigm. By using table inheritance available within PostgreSQL the database schema, coupled with Python wrapping modules, provide the platform through which to conduct infrastructure network modelling and analysis. To address the ability to model interdependencies between networks, a simple mechanism through which interdependencies between two networks can be represented is included within the schema. The utility of the schema has been highlighted through the local-scale assessment of faecal sludge management options via a road-based network analysis. Additionally a range of visualisation and analysis tools that have been developed that can directly interface with the network database schema, enabling rapid visual and analytical analysis of the spatial, topological and metric characteristics of complex spatial infrastructure networks.

#### 5. Acknowledgements

The authors would like to acknowledge funding from the Engineering and Physical Sciences Research Council (EPSRC) grant EP/I01344X/1 to the Infrastructure Transitions Research Consortium (ITRC).

#### 6. Biography

Mr David Alderson received the B.Sc. (Hons) Geographic Information Science from Newcastle University in 2005. He is a Research Assistant in GeoInformatics in the School of Civil Engineering and Geosciences at Newcastle University.

Dr Stuart Barr is Senior Lecturer in Geographic Information Science at Newcastle University.

Dr Tomas Holderness is a Geomatics Research Fellow at the SMART Infrastructure Facility, University of Wollongong, Australia.

Mr Craig Robson is currently studying for a Ph.D. in spatial infrastructure network modelling at Newcastle University.

Mr Alistair Ford is a Researcher in Geomatics at the Newcastle University.

Miss Ruth Kennedy-Walker is currently studying for a Ph.D. in planning and implementation of wastewater collection, treatment and re-use solution in peri-urban areas at Newcastle University.

#### 7. References

Bostock, M. (2012, 11 12). *Force-Directed Graph*. Retrieved 10 30, 2014, from mbostock's block #4062045: <http://bl.ocks.org/mbostock/4062045>

Kennedy-Walker, R., Holderness, T., Alderson, D., Evans, B., & Barr, S. (2014). Using crowd-sourced data for sanitation network modelling in informal settlements. *ICE Municipal Engineer* , 167 (3), 157-165.

OpenLayers. (2014). *OpenLayers.StyleMap*. Retrieved 10 30, 2014, from OpenLayers JavaScript Mapping Library: <http://dev.openlayers.org/releases/OpenLayers-2.13.1/doc/apidocs/files/OpenLayers/StyleMap-js.html#OpenLayers.StyleMap.OpenLayers.StyleMap>

OpenLayers. (2014). *OpenLayers: Free Maps for the Web*. Retrieved 10 30, 2014, from OpenLayers 2: <http://openlayers.org/two/>

*Quantum GIS*. (2014). Retrieved 10 25, 2014, from Quantum GIS: <http://www.qgis.org/en/site/forusers/download.html>



# The Role of Geographical Context in Building Geodemographic Classifications

Alexandros Alexiou<sup>\*1</sup>, Alexander Singleton<sup>♦1</sup>

<sup>1</sup>Department of Geography and Planning, University of Liverpool

November 7, 2014

## Summary

Geodemographic analysis is a methodology that simplifies differentiated patterns of socio-economic and built environment structure for sets of small area geography. A particular issue with many current geodemographic classifications is that these lack any explicit specification of geographic context within the clustering process. Within the broad range of geodemographic applications, current techniques arguably smooth away geographic differences between proximal zones, thus limiting classification sensitivity within local contexts. This research begins to address the issue of geographic context by analyzing and evaluating various local, regional and national extents that can be used as attribute contextual weights.

**KEYWORDS:** Geodemographics, Geographic sensitivity, K-means

## 1. Introduction

Geodemographic analysis is an established methodology that can provide a simplified measure of socio-spatial structure of small area geography. Such classifications have demonstrated utility over a range of public and private sector applications (Longley, 2005; Singleton and Spielman, 2013). Geodemographic analysis typically uses the K-means clustering algorithm of multidimensional socio-economic variables. This methodological framework can capture a wide set of input attributes, taking advantage of the plethora of census variables and other geographically referenced data to generate aggregate multidimensional profiles (Harris et al., 2005).

A particular issue when constructing such classification is the way attributes are used in the clustering process. Due to the aspatial nature of the K-means clustering algorithm, geodemographic classifications account only for similarities in the clustering process and not the geographical context of each area; areas are essentially treated as independent from one another. Arguably, national aggregations could sweep away contextual differences between proximal zones, reducing the local sensitivity of classifications and thus obscuring potentially important patterns. This type of ecological fallacy raises methodological questions regarding the accuracy of geo-classifications, given the inherent loss of within-cluster variation (Voas and Williamson, 2001).

Various approaches use a number of controversial techniques to address these limitations, typically through the implementation of radial buffers for zones, and selecting attribute locational contextual measures. Although there are many national and proprietary classifications available (i.e. the OAC National Classification by the ONS, MOSAIC by Experian and ACORN by CACI), these classifications may not be suitable when assessing local patterns for policy applications. There are indicators that private classifications incorporate locational attribute sensitivity, however, underlying

---

<sup>\*</sup> a.alexiou@liverpool.ac.uk

<sup>♦</sup> alex.singleton@liverpool.ac.uk

techniques are typically obscured and impeding thus impede reproduction, and as such, there are no established tests to their validity (Harris et al., 2005; Longley, 2007). Counter to this argument is that classifications constructed at the national, regional and local extent are effectively built for different purposes, and as such undermines comparison. This is a longstanding debate originating in the earliest of UK classifications (see Openshaw, Cullingford and Gillard, 1980 and Webber, 1980).

## 2. Methodology

This research uses a set of fixed input attributes for Output Area zonal geography to build classifications with different geographic extents. For this purpose, a number of geographical contexts are considered (local, regional, national) to demonstrate the impact on final classification outcome when input variables are kept constant.

Following the methodology of Harris, Sleight and Webber (2005) and Vickers and Rees (2007), a dataset was assembled that includes demographic, economic and housing attributes of England and Wales. The data is assembled in its entirety with 2011 census variables, provided by the Office for National Statistics and aggregated at the Output Area (OA) level. Values were converted into percentages in accordance to their respective denominator (with the exception of *V10: Population Density*). Since k-means clustering is a parametric technique, based on the distributions and correlation levels of the observations certain attributes were discarded (75% cut-off point). The final remaining dataset (Table 1) was normalized using a Box-Cox transformation and converted into z-scores for standardization:

$$z_{i,a} = \frac{x_{i,a} - \mu_S}{\sigma_S} \quad (1)$$

where  $x_{a,i}$  is the attribute value  $i$  of area  $a$  and  $\mu_S$  is the mean and  $\sigma_S$  is the standard deviation of the observations in the dataset  $S$ . In order to measure the contextual differences between the three geographical levels, the mean and standard deviation of the OA observations for the Local, Regional and National datasets  $S_L$ ,  $S_R$ ,  $S_N$  where calculated, and z-scores where adjusted accordingly in equation (1). Each of the three final datasets produced where used for the clustering process in order to measure differences in classification performance.

**Table 1** Final attribute dataset used. Attributes are aggregated per OA code.

<i>Variables</i>	<i>Variable Definition</i>
<i>Demographic</i>	
V1: Age 0–4	Percentage of resident population aged 0–4 years
V2: Age 5–14	Percentage of resident population aged 5–14 years
V3: Age 15–24	Percentage of resident population aged 15–24 years
V4: Age 45–64	Percentage of resident population aged 45–64 years
V5: Age 65+	Percentage of resident population aged 65 or more years
V6: Ethnic Group, Arab	Percentage of people identifying as Arab
V7: Ethnic Group, Black	Percentage of people identifying as black African, black Caribbean or other black
V8: Ethnic Group, Asian	Percentage of people identifying as Indian, Pakistani, Bangladeshi, Chinese or Other Asian
V9: Marital Status, Single	Percentage of population over 16 years who are single
<i>Housing</i>	
V10: Population Density	Number of people per hectare
V11: Rent (Private):	Percentage of households that are private sector rented accommodation
V12: Rent (Public):	Percentage of households that are public sector rented accommodation
V13: Shared	Percentage of households that are shared accommodation.
V14: Flats	Percentage of households which are flats
V15: No central heating	Percentage of occupied household spaces without central heating
<i>Economic Activity</i>	
V16: Working part-time	Percentage of household representatives who are working part-time
V17: Unemployed	Percentage of household representatives who are unemployed
V18: Student	Percentage of household representatives who are full-time students
V19: Low Qualifications	Percentage of people over 16 years with some qualifications but not a HE qualification
V20: Higher Education	Percentage of people over 16 years for which the highest level of qualification is level 4 qualifications and above
V21: NSeC - Managerial	Percentage of households with an HRP with a managerial position

V22: NSeC - Intermediate	Percentage of households with an HRP with an intermediate occupation
V23: Industry, Agriculture	Percentage of population aged 16-74 who work in the A, B and C industry sector
V24: Industry, Manufacture	Percentage of population aged 16-74 who work in the D, E and F industry sector
V25: Industry, Sales	Percentage of population aged 16-74 who work in the G, H and I industry sector
V26: Industry, Technology	Percentage of population aged 16-74 who work in the K, L and M industry sector
V27: Industry, Administration	Percentage of population aged 16-74 who work in the N, O, P, Q, T, and U industry sector
V28: Industry, Art	Percentage of population aged 16-74 who work in the R and S industry sector
<i>Travel Behaviour</i>	
V29: No car household	Percentage of households with no cars
V30: 1 Car household	Percentage of households with 1 car
V31: 3+ Car household	Percentage of households with 3 or more cars
V32: Travel, Public	Percentage of population aged 16-74 who travel to work by public transport
V33: Travel, Foot/Bicycle	Percentage of population aged 16-74 who travel to work on foot or by bicycle

The classification methodology to produce clusters is the iterative allocation–reallocation algorithm, known as the *K-means clustering* detailed in Milligan (1996) and Everitt, Landau and Leese (2001). K-means clustering uses squared Euclidean distance as a dissimilarity function. Essentially, K-means clustering assigns N observations into K clusters in such a way that within each cluster, the average distance of the variable values from the cluster mean is minimized:

$$WCSS = \min_c \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \quad (2)$$

where WCSS is the within-cluster sum of squares for a cluster distribution C with K seeds,  $x_i \in N$  is the data observations and  $\bar{x}_k$  is the k cluster mean. Since the algorithm is dependent on the initial seeds, it must run multiple times in order to obtain optimal results (typically minimizing the WCSS). This research focuses on broad socioeconomic categories (known as the Supergroup hierarchy), and based on the optimum ratio between the number of clusters and their respective total WCSS, 7 clusters were selected in order to carry out the analysis.

Once the optimised sets of K cluster assignments are calculated for each geographic context, clusters within each set are matched in order to determine which cluster ID from one classification fits best to another. Contrary to the typical qualitative way, i.e. cross-tabulation of the within-cluster distribution, an algorithm was developed to analyze the degree of “fitness” of a set of different classifications. The angular cosine similarity measure is used for this purpose, given by:

$$ACS(A, B) = \frac{1 - \cos^{-1}(\cos \theta)}{\pi}, \quad \cos \theta = \frac{\sum_i^n A_i * B_i}{\sqrt{\sum_i^n A_i^2} * \sqrt{\sum_i^n B_i^2}} \quad (3)$$

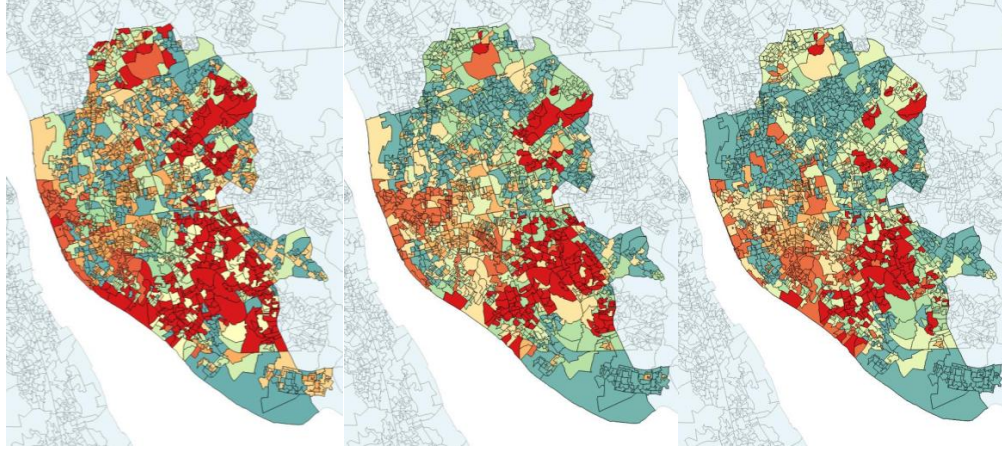
If  $k_i = \begin{pmatrix} \mu_1 \\ \dots \\ \mu_n \end{pmatrix} \in C$  represents a vector with the average attribute values  $\mu$  of cluster  $k_i$  of the set C, then that cluster is more similar to another cluster  $k'_i \in C'$ , given they derive from the same set of observations N, when the ACS is closer to 1 (and 0 if they are completely dissimilar - opposite vectors). Taking this into account, it is possible to find the combination of pairs for which  $C \cong C'$ . If  $C' = \{k'_1, \dots, k'_i\}$  represents the vectors of the attributes means of K clusters, then there is one permutation of  $C' \in {}^K P_K$  for which the similarity between the two classifications is maximized:

$$\sum_{i=1}^n ACS(k_i, k'_i) = \max \quad (4)$$

Finally, this research uses the R programming language in order to perform the analysis and map the output classifications.

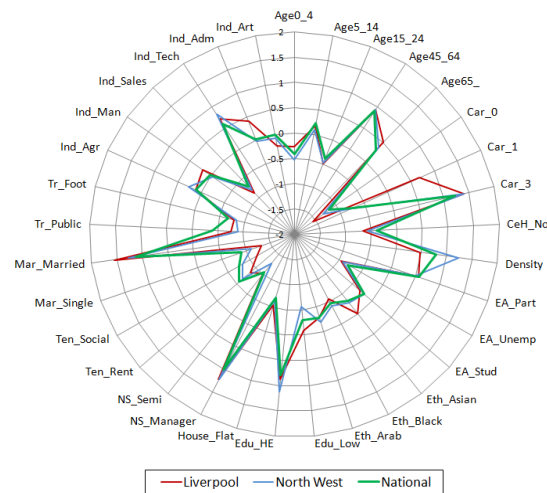
### 3. Preliminary results and future directions

In this particular example, the Local Authority of Liverpool is considered (1584 Output Areas) and is used as a basis to compare different classification outcomes. Figure 1 demonstrates how the local, national and regional classifications are mapped within this context. Between the classifications, there are differences in the emergent cluster patterns, with the local classification appearing to offer the greatest differentiation between areas. The cluster mapped with a red colour represents the most affluent residents (e.g. “White Collar Families”).



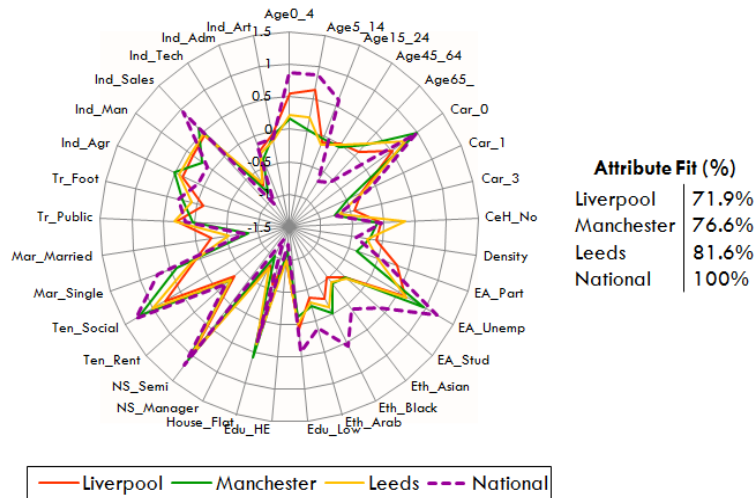
**Figure 1** Differences in cluster patterns in Liverpool, UK. From left to right: Local, Regional and National geographical contexts used to classify Output Areas.

Figure 2 plots the distribution of average attribute values in standard deviations within a cluster portrayed as “White Collar Families”. Further analysis reveals that the number of OAs in the cluster decreases as the attribute extents are scaled. For instance, an affluent family by local standards may not be as affluent by national ones. Since the Liverpool area is considered generally deprived, this number decreases from 234 OAs to 172 in the regional context and 118 in the national one.



**Figure 2** Distribution of average attribute values of cluster “White Collar Families”, mapped red in Figure 1, for local, regional, and national extents respectively.

Although preliminary results show some degree of differentiation, a more extensive analysis is required to explore how these patterns may map between different geographic contexts, for example, how such patterns might differ between Leeds, Liverpool, or Manchester (Figure 3). Furthermore, research is needed to explore how classifications created at local or regional extents can be assembled in a way that national comparisons become possible. A challenge for future research is how these differences can be measured, and how between classifications created for different scales impacts upon the performance of the classifications when used for real world applications.



**Figure 3** Comparison of the distribution of mean attribute values of the cluster “*Hard-Pressed Households*” of the national classification (England), with the local classifications for Liverpool, Manchester and Leeds, along with their respective ACS levels (%).

Finally, for simplicity, administrative definitions of context have been used for this study, however, we recognise that these may not represent true functional regionals or localities, and as such, further work is required about how local or regional extents might be defined, and what impact these geography will have on the final classification. In particular, further work is required to examine how built environment / transport infrastructure can be used to measure geographic extents and how this may impact upon emergent patterns.

#### 4. Acknowledgements

This research is part of a PhD project funded by the ESRC with an Advanced Quantitative Methods (AQM) award at the North West Doctoral Training Centre.

#### 5. Biography

Alexandros Alexiou holds a diploma in Engineering with an MSc (Transport Planning) from the Aristotle University of Thessaloniki and an MPhil (Land Economy) from the University of Cambridge. Alexandros is currently a PhD candidate at the University of Liverpool, with research interests in the creation of new models of urban socio-spatial structure that better account for both geographic context and the dynamics of population.

Alex Singleton is a Reader in Geographic Information Science at the University of Liverpool. His research interests extend a geographic tradition of area classification and have developed a broad critique of the ways in which geodemographic methods can be refined through modern scientific approaches to data mining, geographic information science and quantitative human geography.

## References

- Everitt B S, Landau S and Leese M. (2001). *Cluster Analysis*. 4th edn. London: Arnold.
- Harris R, Sleight P and Webber R (2005). *Geodemographics, GIS, and Neighbourhood Targeting*. Chichester: John Wiley & Sons.
- Longley P A (2005). Geographical information systems: a renaissance of geodemographics for public service delivery. *Progress in Human Geography* 29(1): 57-63.
- Longley, P. A. (2007). Some challenges to geodemographic analysis and their wider implications for the practice of GIScience. *Computers, Environment and Urban Systems* 31(6): 617–622.
- Milligan G W (1996). Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert & G. De Soete (Eds.), *Clustering and Classification* (Vol. 2, pp. 120-125). Singapore: World Scientific Press.
- Openshaw S, Cullingford D and Gillard A (1980). A critique of the national classifications of OPCS/PRAG. *Town Planning Review* 51 (4): 421.
- Singleton A D and Spielman S E (2013). The Past, Present and Future of Geodemographic Research in the United States and United Kingdom. *The Professional Geographer*.
- Vickers D and Rees P (2007). Creating the UK national statistics 2001 output area classification. *Journal of the Royal Statistical Society. Series A. Statistics in society* 170(2): 379-403.
- Voas D and Williamson P (2001). The diversity of diversity: a critique of geodemographic classification. *Area* 33(1): 63–76.
- Webber R J (1980). A response to the critique of the national classifications of OPCS/PRAG. *The Town Planning Review* 51 (4): 440–450.

# **Utilising GIS capabilities to study and analyse the spatial distribution of crimes in Kuwait**

Nawaf Alfadhli, Graham Clarke and Mark Birkin

School of Geography

University of Leeds

April 2015

## **Introduction**

Generally, the level of security is considered to be one of the main indicators of development in any community. Thus, it is essential to maximise public safety and effectively tackle all of the serious types of harm that people face, particularly with regard to crime-related incidents. Naturally, criminal activities are a primary concern, and awareness of them results in a significant negative monetary and psychological impact on people and nations. In every society, crime is classified as a serious state of insecurity, fear and discomfort. Each country determines what constitutes a series of forbidden criminal activities and punishes a criminal for those activities by imposing fines or imprisonment or both. Therefore, there is no permanent or globally determined definition of crime (Henry and Lanier, 2001). One of the primary ways to better understand and recognise patterns of crime and how crime can be effectively handled is to determine the precise geographical location in which the crime is committed (Chainey and Ratcliffe, 2013). Several disciplines, such as sociology, psychology, criminology and geography, have traditionally contributed to fully exploring the study of crime (Georges, 1978).

Hence, decision makers and planners in most developed countries have recently formulated their strategic plans to effectively provide the highest standards of security. Thus, the use of various methods and visualisation tools, particularly the Geographic Information System (GIS) tool, has dramatically increased in modern policing and crime scene investigation. However, in Kuwait, as in most developing countries, police organisations have encountered fundamental problems when attempting to determine, investigate and prevent criminal activities. Surprisingly, modern technologies such as GIS, particularly in the crime investigation field, have not been widely used in these countries to date compared with developed countries, which are greatly concerned with using these visualising tools.

Specifically, the department of the civil defense at the Ministry of the Interior in Kuwait is typically responsible for producing emergency plans. These plans can be provided to the police to respond and deal faster and more effectively with criminal activities. Although Kuwaiti planners are likely to still use the traditional methods in emergency services for various reasons, there are obstacles which should be taken in to consideration, including a lack of public awareness regarding the capabilities that GIS offers in policing services, as well as the kind of physical and spatial data that needs to be available. The Ministry of Interior continues to follow the modern methodologies in monitoring and analysis that are based on the followed procession of the scientific treatment. That treatment is planned to control and limit the negative criminal phenomena of some kinds of behaviours in the society. The goal of the ministry is to manage and participate successfully in taking actions to resist such

behaviours, which are completely out of the steady traditions and customs of the society.

This study basically concentrates on utilising GIS capabilities to study and analyse the spatial distribution of four patterns of crime in the state of Kuwait with the purpose of developing crime reduction strategies and organising the police force. In order to achieve this purpose, two analytical methods have been specifically used: location-allocation modelling and spatial interaction modelling.

The project is carried out based on data provided by 80 police stations covering 89 districts in Kuwait. The crime mapping and crime analysis primarily focuses on determining the types of crimes, such as crimes against the public interest, crimes of defamation and insult, crimes against people (physical body) and property crimes.

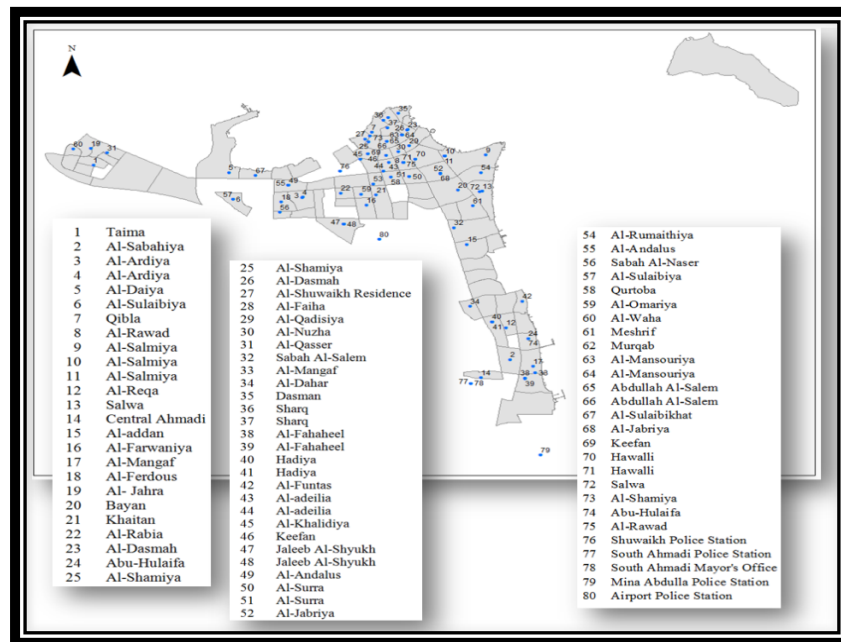


Figure 1: The distribution of existing police stations in Kuwait. Source: Ministry of Interior, 2012

As can be seen in Figure 1, all governorates in Kuwait are relatively covered by the police stations. However, they vary from one governorate to other based the size of population as well as the nature of governorate, administratively and economically. For example, the Capital governorate, which is the capital of Kuwait, is the most important governorate in Kuwait in terms of administrative and economic importance. Moreover, the majority of urban districts, such as Al-Shuwaikh, Abdullah Al-Salem, Al-Rawad and Surra, are located in this governorate. Hence, there are 31 police stations located in this governorate, although the offence rate in 2012 was relatively low in this governorate. There were 864 crimes in a population of 522670. In comparison, the Mubarak Al-Khabeer, which is the latest governorate constructed, has just two police stations established.



## Using GIS Techniques in policing and to prevent crime

A Geographic Information System (GIS) is the result of a combination of traditional types of science, such as geography, cartography and surveying, and modern types of science, such as remote sensing, Global Positioning Systems (GPS) and computer sciences (Gu et al., 2009). Typically, many definitions can be applied to the term Geographic Information System (GIS), and most of those definitions focus on the hardware, software, data and analysis components. However, depending on the nature of the purpose of the system's use and on the user's knowledge, GIS can be easily defined. One way that GIS can be defined is as a computer system that provides several powerful analytical functions, such as capturing, managing, integrating, manipulating, analysing and presenting spatial data and geographic information referenced to the Earth to support informed decision making (Heywood et al., 2002). At its simplest level, GIS is defined as

*"A powerful set of tools for collecting, storing, retrieving, at will, transforming and displaying spatial data from the real world for a particular set of purposes"*  
(Burrough and McDonnell, 1998)

Its rationale is centred on the belief that GIS has significant capabilities to positively help a local police force rapidly allocate, respond to, forecast and, thus, prevent emerging spatial patterns of crime in an informative and efficient way. In addition to its data query capabilities, GIS is a powerful tool that can be prevalently applied to provide a wide range of spatially integrated, comprehensive and referenced data for police organisations so they can have access to specialised crime mapping and conduct specific crime analysis. Moreover, law enforcement organisations can effectively apply the communication component of GIS to identify the detailed relationships between the pattern of offence, the victim and the offender (Neill and Gorr, 2007).

Using the powerful tools of GIS from the primary process of data collection through the investigation and assessment of any criminal activity can fundamentally support policing and crime prevention initiatives. Consequently, strategic decision-making can be supported by utilising the powerful analytical tool that is provided by using GIS (Chainey and Ratcliffe, 2013). It is also helpful in terms of identifying the spatial and social distribution of victims and the demographic characteristics of the offenders (Gaviria and Pagés, 2002).

In the 21<sup>st</sup> century, GIS is considered to be ubiquitous, covering a wide range of aspects of contemporary life, especially in policing and crime prevention. In general, computerized crime mapping can play a positive and vital role in relation to allocating resources, predicting staff needs and assessing crime prevention strategies (Yon, 2003).

A wide range of literature studies conducted in several countries, such as the United States, the United Kingdom, Italy, Germany and some countries in Latin America, have illustrated the relationship between crime and its determinants. They have noted that several socioeconomic, environmental and cultural determinants may positively or negatively impact the crime rate based on the spatial distribution of offences, offenders and victims, specifically age, gender, urbanization, poverty, wages, income inequality, social exclusion, educational level and cultural and family background (Buonanno and Leonida, 2005). Moreover, to profile catchment areas into customer segment types and locate facilities, a mixture of GIS and spatial location models have been increasingly applied throughout the arrival of geodemographic packages (Birkin and Clarke, 1998).

## **Location-Allocation (LA) Models**

In the current century, the most important strategic and operational considerations in metropolitan areas is to optimally determine the location for emergency services facilities such as police stations, fire stations and ambulance stations by maximising the coverage of events throughout the regions and by minimising the response time of dealing rapidly and effectively with emergency events (Li et al., 2011).

Consequently, location-allocation modelling is one of the most prevalent techniques used to determine the optimal location for service provision. These models, which are commonly applied within some GIS systems, particularly ArcGIS, are frequently used by service planners to identify the optimal location for service provision, depending upon a particular set of criteria or identified constraints, such as the number of possible facilities that need to be opened, the number of facility sites to be distributed, the demand points allocated for each single potential facility and the need to minimise the distance travelled, time response or cost of travel from each source (event) to the destination (facility) (Brunsdon and Singleton, 2015). An ideal location plays a critical role in emergency service systems, telecommunication networks, public services, etc. Minimising the costs and decreasing the travel distances are the principal objectives of locating emergency sites (Rahman and Smith, 2000).

Our study is mainly concerned with the redistribution of the locations of police stations in Kuwait. Hence, Location-Allocation models can be applied to help regional planners and decision-makers determine ideal sites for police stations based on critical factors provided by the ArcInfo workstation, considering population size (demand point), the total number of facilities required and the proper travel distance, which is identified as the distance between demand point and the site of police station. We applied three different approaches in more detail in this study to compare the results produced for identifying the optimal locations for 5, 15, 25 and 35 police stations considering population size, crime recorded and crime patterns from a set of 80 existing locations in Kuwait. These approaches can be developed by most of the existing GIS software packages, specifically, by using a fundamental function in ArcInfo WorkStation in ArcGIS (Version 10).

**The optimal locations for 5,15,25 and 35 police stations in Kuwait districts considering the population size, the minimum distance and the number of facilities needed criterion**

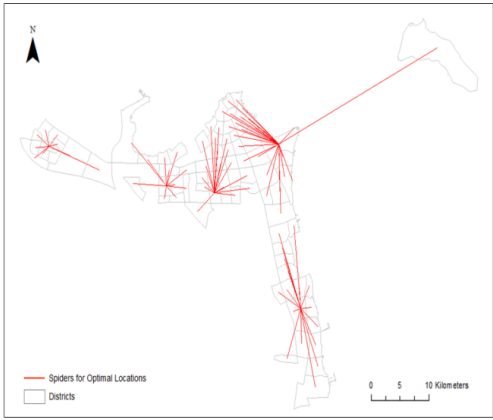


Figure 2: The optimal locations for 5 police stations in Kuwait

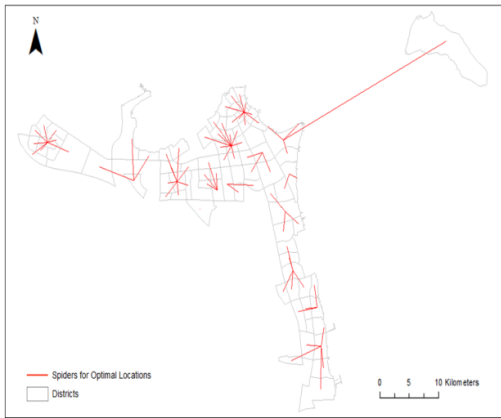


Figure 3: The optimal locations for 15 police stations in Kuwait districts

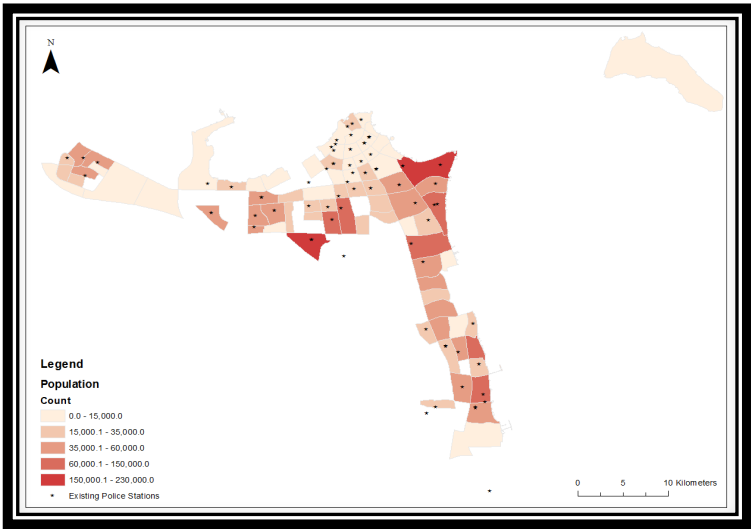


Figure 4: The distribution of population size in the districts of Kuwait



Figure 5: The optimal locations for 25 police stations in Kuwait



Figure 6: The optimal locations for 35 police stations in Kuwait districts

## The Location Set Covering Problem (LSCP)

The dominance concept of the Location Set Covering Problem (LSCP) is to the coverage of a demand district by a police station, which means that the police station can cover the demand area and the ability of police vehicles to reach the demand area from its existing location and position within a stated time or distance standard. It is clear that the location set covering models can benefit critically by the emergency services by providing the total number of police stations and their position, such that all demand areas may require at least one police station which is stationed to cover the demand areas based upon time or distance standard

As previously noted, this approach is principally aimed at covering the most-populated areas with police stations. Moreover, it is capable of determining the total number of police stations needed and their locations based upon population size (Revell, 2009). A further benefit that can be provided by applying this method is the determination of the time needed to reach to an emergency event from an emergency services facility, which calls "Time-Response. "This offers the number of police vehicles that are required to rapidly and appropriately respond to emergencies.

It should be mentioned that many questions can be answered easily by applying LSCP, including:

- How many police stations should be positioned?
- Are there enough to deal effectively with the event of the emergency based on the demand areas and the size of population?
- What is an appropriate site for the police station to be placed, depending on the maximum number of people?
- What is an ideal response time? (In the urban areas, it is approximately 2-5 minutes between police station and demand area).
- How many police vehicles are required or needed?
- Where must they be positioned to provide useful services?

In contrast, the coverer is not necessarily available at all times to attend the demand area when a call comes in. In addition, predictable and rational fashions are considered as the main aspects, which should be included in The Location Set Covering Problem (LSCP)(Revelle, 2009).

Given these benefits, this method is applied in this study through the application of the following formula:

$$\min \sum_j z_j$$

The number of police stations needed is assumed to be thirty-five, despite the fact that the current number of the existing police stations is 80 as can be seen in Figure 1. There are

several reasons to justify the assumption of this proposed number; for example, the majority of cities in all governorates would be covered fairly by police stations. Another important aspect that should be taken under consideration is that there are certain districts that may share one police station spatially and have no dire need to build police station, especially if it is located in a district with a small population. In looking at figure 6, for example, districts like Jahra, Nasseem, Oyoun, Taima, Amgara, Waha, Jahra Industrial, Qasr and Na'eem can be assumed to be covered by police stations located in the adjacent Central of Al-Jahra governorate. Moreover, the Al-Sulaibiya police station could cover three districts fairly: Sikrab, Sulaibikhat and Doha. In terms of the redistribution of police stations in the Al-Farwaniya governorate, seven districts; Andalus, Ghranada, Shuwaikh Health, Reqay, Ardhiya, Sabah Al-Nasser and Ashbeliah could share one police station, i.e. Al-Ferdous police station. However, the designed model, suitably, would not relocate Jaleeb Al-Shiyoukh police station, which is located adjacent to the district that has a significant population (nearly 295,000 people). It is clear that some areas that are positioned in two different governorates may be covered by one police station; for instance, particular districts located in Al-Farwaniya governorate, such as Zahra, Al-Salam, Rai, Omariya, Rabiya, Rehab and Farawaniya, could be covered by Khaitan police station, with Yarmouk and Qortuba located in the Capital governorate. Additionally, the Capital governorate has some areas that are geographically close to some districts in Hawalli governorate. Therefore, it is possible to relocate their police stations as one police station shared between the two governorates, such as the Hawalli police station, which could include the Sha'ab and Jabriya areas located in Hawalli governorate along with Nuzha, Surra, Qadsia and Rawda areas located in the Capital governorate. It should be mentioned that Failaka Island, which has a very small population (117 people) could share Al-Salmiya police station, which is positioned in an overcrowded district. Moreover, the Salwa, Bayan and Meshrif districts in Hawalli governorate could relatively share Salwa police station.

By applying LSCP, most districts in the governorates of Kuwait could be covered by police stations. However, a few of districts that have small population sizes would not be covered by a police station (Figure 6). With flexibility, uncovered districts may utilise the services provided by facilities within covered districts.

## **Spatial Interaction Models**

Spatial interactions, such as migration and airline travel, naturally form a location-to-location network (graph). In the network, a node represents a location (or an area) and a link represents an interaction (flow) between two locations. Locational measures, including both simple measures, such as in-flow, out-flow and net-flow and more complicated measures such as centrality, entropy and assortativity, are often derived to understand the structural characteristics of locations and the roles they play in generating interactions. However, due to the dramatic differences in size (such as population) among locations and the small-area problem, locational measures that are derived with the original flow data often exhibit spurious variations and they may not be able to reveal the true underlying spatial and network structures (Koylu and Guo, 2013).

Spatial interaction models can be grouped under the generic heading of gravity models (Roy and Thill, 2004). They have gained wide acceptance as a reasonable model of spatial interactions between locations (such as regions). Spatial interaction models incorporate a

function characterising the origin,  $i$ , of the interaction, a function characterising the destination,  $j$ , of the interaction and a function characterising the separation between two regions,  $i$  and  $j$ . The model is characterised by a formal distinction that is implicit in the definitions of origins and destination functions on the one hand, and the separation functions on the other hand. Origin and destination functions are described using weighted origin and destination variables, respectively, while the separation functions are postulated to be explicit functions of numerical separation variables (LeSage et al., 2007).

Given these benefits, Gravity method is applied in this study through the application of the following formula:

$$F_{ij} = \frac{A_i S_i D_j}{e^{\beta d_{ij}}}$$

Where:  $F_{ij}$  is the flow from the  $i^{\text{th}}$  district to  $j^{\text{th}}$  police station,  $A_i$  is a balancing factor that takes into account the surrounding police stations of  $j^{\text{th}}$  police station and assumed to be 1,  $S_i$  is the flow from the  $i^{\text{th}}$  district (source of flow),  $D_j$  is the attractiveness of the  $j^{\text{th}}$  police station (destination for the flow) and assumed to be 1,  $\beta$  is a factor that controls the ease and travel distance from the  $i^{\text{th}}$  district to the  $j^{\text{th}}$  police station and is assumed to decrease the flow for close districts and increase it for far districts ( $\beta > 1$  for short distance and  $\beta \leq 1$  for long distance) and  $d_{ij}$  is the distance (aerial) from  $i^{\text{th}}$  district to the  $j^{\text{th}}$  police station.

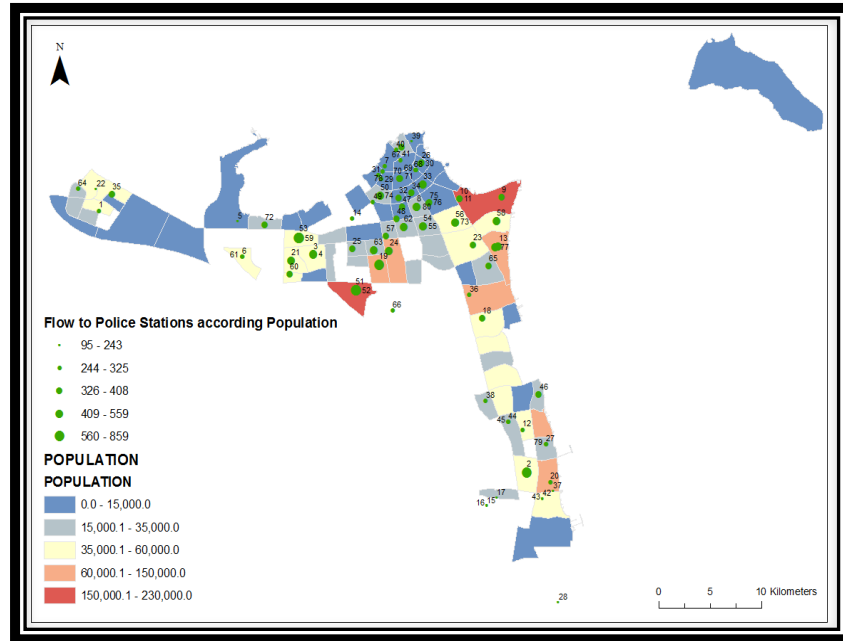


Figure 1: The flow from districts (89 districts) to plice stations (80 police stations) considering the population

## References

- BIRKIN, M. & CLARKE, G. 1998. GIS, geodemographics, and spatial modeling in the UK financial service industry. *Journal of Housing Research*, 9, 87-111.
- BRUNSDON, C. & SINGLETON, A. 2015. *Geocomputation: A Practical Primer*, SAGE.
- BUONANNO, P. & LEONIDA, L. 2005. Criminal activity and education: evidence from Italian Regions.
- BURROUGH, P. A. & MCDONNELL, R. A. 1998. *Principles of GIS*. Oxford University Press, London.
- CHAINEY, S. & RATCLIFFE, J. 2013. *GIS and crime mapping*, John Wiley & Sons.
- GAVIRIA, A. & PAGÉS, C. 2002. Patterns of crime victimization in Latin American cities. *Journal of Development Economics*, 67, 181-203.
- GEORGES, D. E. The geography of crime and violence: A spatial and ecological perspective. 1978. Association of American Geographers Washington, DC.
- GU, W., WANG, X. & GENG, L. 2009. GIS-FLSolution: A spatial analysis platform for static and transportation facility location allocation problem. *Foundations of Intelligent Systems*. Springer.
- HENRY, S. & LANIER, M. 2001. *What is Crime?: Controversies Over the Nature of Crime and what to Do about it*, Rowman & Littlefield.
- HEYWOOD, J., CORNELIUS, S. & CARVER, S. 2002. An introduction to GIS. *Pearson Education Asia*, 13.
- KOYLU, C. & GUO, D. 2013. Smoothing locational measures in spatial interaction networks. *Computers, Environment and Urban Systems*, 41, 12-25.
- LESAGE, J. P., FISCHER, M. M. & SCHERNGELL, T. 2007. Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects\*. *Papers in Regional Science*, 86, 393-421.
- LI, X., ZHAO, Z., ZHU, X. & WYATT, T. 2011. Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, 74, 281-310.
- NEILL, D. B. & GORR, W. L. 2007. Detecting and preventing emerging epidemics of crime.
- RAHMAN, S.-U. & SMITH, D. K. 2000. Use of location-allocation models in health service development planning in developing nations. *European Journal of Operational Research*, 123, 437-452.
- REVELLE, C. 2009. Siting ambulances and fire companies: New tools for planners. *American Planning Association. Journal of the American Planning Association*, 57.
- ROY, J. R. & THILL, J.-C. 2004. *Spatial interaction modelling*, Springer.
- YON, H. Impact of crime mapping on the crime analysis approach of the Ankara Police Department. *FORENSIC SCIENCE INTERNATIONAL*, 2003. ELSEVIER SCI IRELAND LTD CUSTOMER RELATIONS MANAGER, BAY 15, SHANNON INDUSTRIAL ESTATE CO, CLARE, IRELAND, 12-12.

# Characterisation and Classification of Hydrological Catchments in Alberta, Canada Using Growing Self-Organising Maps

Michael Allchin<sup>\*1</sup>

<sup>1</sup>Natural Resources and Environmental Studies,  
University of Northern British Columbia, Canada

January 8, 2015

## Summary

Operational hydrologists are often required to transfer information from well-understood instrumented research basins to ‘wild’ catchments for which few details are available. To do so successfully, the climatological inputs and physiographic processing in both must be sufficiently similar that their resultant flow regimes will also be comparable. This is challenging to determine, because of the wide variety of influences on hydrological response, and the degree of heterogeneity among and within catchments. Pattern recognition – or classification – can help with this. This study explores the application of Growing Self-Organising Maps, a data-mining technique based on unsupervised machine learning, for this purpose.

**KEYWORDS:** Catchment Hydrology; Classification; Data-Mining; Self-Organising Maps

## 1. Introduction: The Need for Classification

Most hydrological research is driven by the dependence of human and ecological systems on freshwater resources, and the potential impacts of their surfeit or deficit. The fundamental spatial unit adopted for many studies is the catchment or drainage basin, conceptualised as a topographic funnel which converts spatially-distributed precipitation into streamflow at a single outlet (Wagener *et al.*, 2007). Processes hosted by the basin thus modulate meteorological inputs, acting as complex spatio-temporal filters which control the pathways and rates of water transmission (Woods, 2002).

While landscape attributes evolve through mutual interaction, so catchments possess a degree of self-organisation (Sivapalan, 2005; Ehret *et al.*, 2008), their physiographies are also highly heterogeneous (Troch *et al.*, 2008): every catchment is essentially unique (Beven, 2000). It follows that the accuracy with which a catchment’s transformation of climatic inputs to streamflow outputs – its ‘hydrological response’ – may be modelled, depends largely on the resolution at which these properties are represented. However, operational hydrologists require generalised, practical representations of links between climate, landscape and flow regimes for predictive purposes. One approach is *regionalisation*, which seeks to transfer information describing the behaviour of instrumented basins to ungauged ‘wild’ catchments. This in turn depends on the recognition of signature spatio-temporal patterns of climate and landscape, and their association with different hydrological responses (Sivapalan, 2005; Beven, 2000). Pattern recognition – *the association of an infinite, continuous set of inputs with a finite variety of outputs* (András, 2008) – implies the identification and labelling of components based on their distinguishing characteristics, or ‘classification’.

There have been repeated calls for the development of objective and rigorous methods for catchment classification (Wagener *et al.*, 2007; McDonnell and Woods, 2004; Sivakumar *et al.*, 2013). Such a framework should integrate physiographic *Form* and climatological *Forcing* with hydrological

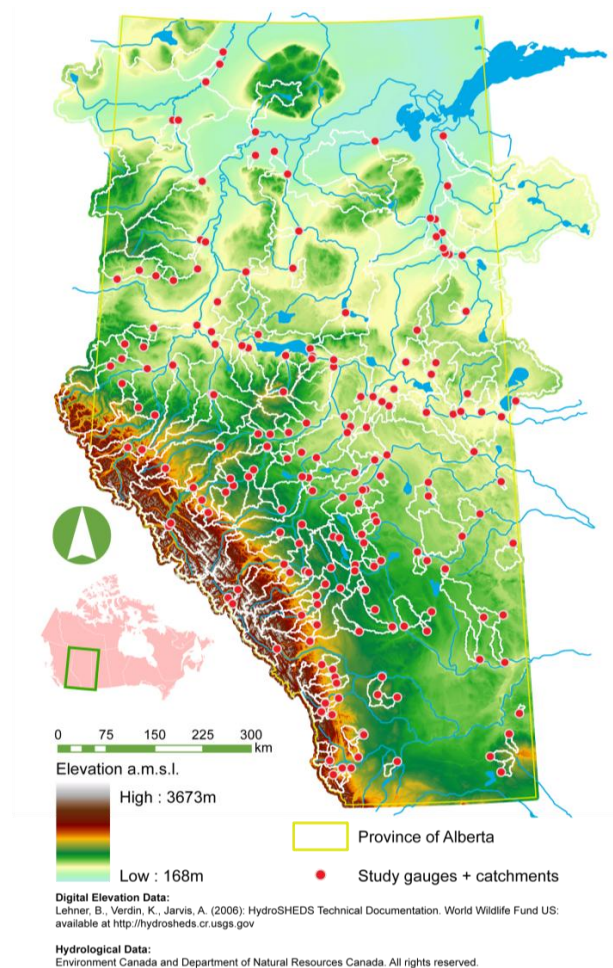
---

\* allchin@unbc.ca



response or *Function*, as manifested in synoptic hydrometric records (Wagener *et al.*, 2007). Whilst the diversity of influencing factors makes this a daunting challenge, this is the type of task for which data-driven techniques such as the Self-Organising Map (SOM: Kohonen, 1982, 1990) have been developed. SOMs have so far been applied only rarely for this purpose (Kalteh *et al.*, 2008), but increases in computational power now permit their operation on mainstream platforms, and more suitable datasets have been made available from credible sources.

This study applied a SOM variant to generate classifications of *Form*, *Forcing* and *Function* for approximately 200 catchments across the Province of Alberta, Canada (Figure 1), supporting the identification of associative patterns between climatological inputs, physiographic processing, and hydrometric outputs.



**Figure 1:** The Province of Alberta, showing catchments and gauges included in this study

## 2. Data and Methods

The conventional SOM comprises a fixed grid of neurons, each owning an ordered set of numerical weights. Training is achieved through unsupervised machine learning (Kohonen, 1982, 1990): on its completion, a SOM partitions a dataset into Voronoi Regions, projecting these onto its grid so that spatial relationships between neurons reflect those within the data-space. Individual neurons thus represent a fine-resolution classification, but may also form distinct contiguous clusters. One problem with a fixed SOM is that some idea of the dataset's internal variability is required to size the grid appropriately, but this may not be available. Dynamic SOMs, such as the Growing SOM algorithm

adopted for this study (GSOM: Alahakoon *et al.*, 2000; Amarasiri *et al.*, 2004), therefore begin with a few neurons, and add more as additional variability is encountered in the dataset.

Physiographic and climatological datasets (Tables 1, 2) were sourced for the Province of Alberta (~700,000 km<sup>2</sup>), using criteria of ready and free availability, quality, consistency, credibility of provenance, and spatio-temporal extent and resolution, and summarised at kilometric resolution. Hydrometric data measured at 213 gauges between 1989 and 2009, and corresponding catchment boundaries, were provided by the Water Survey of Canada.

**Table 1:** *Form* descriptors

Group	Metrics	
<b>Slope</b>	Mean Coefficient of variation	
<b>Slope / Aspect</b> (%age cover)	NW-NE NE-SE SE-SW SW-NW	shallow ( $\leq 10^\circ$ ) moderate ( $10^\circ \leq 30^\circ$ ) steep ( $30^\circ \leq 60^\circ$ ) very steep ( $> 60^\circ$ )
<b>Surface Complexity</b>	Represented by coefficient of variation of Beven and Kirkby (1979) Wetness Index	
<b>Solid Geology</b> (%age cover)	Cenozoic	coarse siliciclastic coarse-medium siliciclastic medium siliciclastic medium-fine siliciclastic volcanic
	Mesozoic	coarse siliciclastic coarse-medium siliciclastic medium siliciclastic medium-fine siliciclastic fine siliciclastic carbonates
	Palaeozoic	coarse siliciclastic coarse-medium siliciclastic medium siliciclastic medium-fine siliciclastic fine siliciclastic carbonates evaporates
	Proterozoic	coarse siliciclastic coarse-medium siliciclastic medium siliciclastic fine siliciclastic carbonates plutonic / high-grade metamorphic
	Archaean	plutonic / high-grade metamorphic
<b>Drift Geology</b> (%age cover)	Alluvial deposits Coarse grain Colluvial blocks Colluvial fines Colluvial rubble Colluvial sand Complex Aeolian deposits Fine grain	

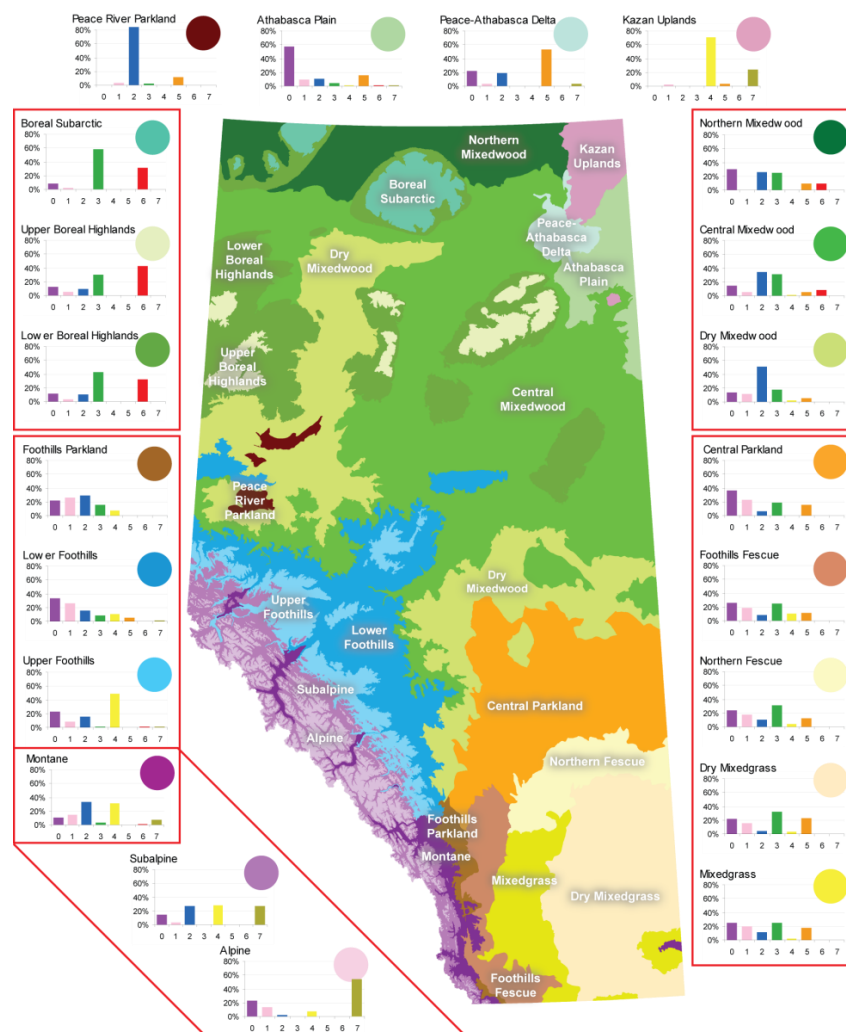
Group	Metrics
	Glaciers Organic deposits Plain sands and gravels Till blanket Till veneer Undivided (implies negligible or absent)
<b>Land Cover</b> (%age cover)	Temperate / sub-polar needleleaf forest Sub-polar taiga needleleaf forest Temperate / sub-polar broadleaf deciduous Mixed Forest Temperate / sub-polar grassland Sub-polar / polar grassland-lichen-moss Sub-polar / polar barren-lichen-moss Wetland Arable Barren Lands Urban / Built-Up Water
<b>Soil Drainage</b> (%age cover)	n/a (no soil) Very poor Poor Imperfect Moderate Good Rapid Very rapid
<b>Permafrost</b> (%age cover)	Isolated patches (0-10%), low (<10%) ground Ice Isolated patches (0-10%), low-nil (0-10%) ground ice Sporadic discontinuous (<10%), low (<10%) ground ice
<b>Total 81 values per data-point</b>	

**Table 2:** *Forcing* descriptors

Description
Mean daily maximum temperature (°C)
Mean daily mean temperature (°C)
Total monthly precipitation (mm)
Degree-days below 0°C
Degree-days above 5°C
Degree-days below 18°C
Degree-days above 18°C
Number of frost-free days
Precipitation as snow (mm)
Hargreaves reference evaporation (mm)
<b>12 monthly values per metric:</b> <b>120 values per data-point</b>

GSOMs were developed to classify representative samples of these descriptions. The physiographic training data comprised 50% of the cells in a chequerboard pattern: given the lower spatial frequency of variation in the climatological dataset, 25% of these were used in this dataset. Prototype GSOMs were generated from increasing training durations, with outcomes judged by the contiguity of neuron-clusters identified by the GSOM software; metrics of central tendency and dispersion of neuron-quantisation errors within clusters; their spatial segmentation when mapped; internal consistency of the underlying descriptors; and comparison with an independent classification, the Alberta Natural Sub-Regions (NSRs: Government of Alberta, 2006).

The second stage identified from these GSOMs the *Form* and *Forcing* class of every 1 km<sup>2</sup> cell in each catchment, and characterised every basin in terms of the fractional cover of each combination. These profiles were used to develop a further GSOM, to identify clusters of catchments with comparable climatological inputs and physiographic processing.



**Alberta Natural Sub-Regions Data:**  
 Natural Regions Committee 2006. Natural Regions and Subregions of Alberta.  
 Compiled by D.J. Downing and W.W. Pettapiece. Government of Alberta. Pub. No. T/852.  
 Available from  
<http://tpr.alberta.ca/parks/heritageinfocentre/naturalregions/>  
 ©2006, Her Majesty the Queen in Right of Alberta, as represented by the Minister of Environment.

**Figure 2:** Distributions of fractional cover of the eight *Form* classes in the NSR polygons

The approach developed to classify the catchments' gauged hydrometric profiles lies outside the core scope of GIS Research: however, a brief description is required. If two catchments exhibit comparable hydrological response (driven primarily by their physical attributes), then with similar climatic inputs in a given year, their annual hydrographs should develop similar shapes. A GSOM was first developed to classify the set of annual hydrographs measured at the available gauges from 1989 to 2009 into clusters with broadly comparable timings and intensities of peak flow. A second GSOM was then developed to cluster gauges based on consistent inter-annual similarities in annual hydrograph shape.

Having thus identified the membership of each catchment in classes of physiographic / climatological *Form-Forcing* and hydrometric *Function*, overlaps could then be explored between these associations, thereby supporting inferences about the probable hydrological behaviour of catchments based purely on readily-available spatial data.

### 3. Results

The *Form* GSOM identified eight physiographic classes across the study area. The disparate metrics involved initially made it challenging to determine whether or not these were meaningful. However, computing their fractional coverages in each of the twenty-one NSR class polygons revealed clear signature distributions, and evident associations with the established NSR categories (*e.g.* Prairie, Mountain / Foothill, Boreal Forest) (Figure 2), indicating that these classes provided useful representations of landscape type.

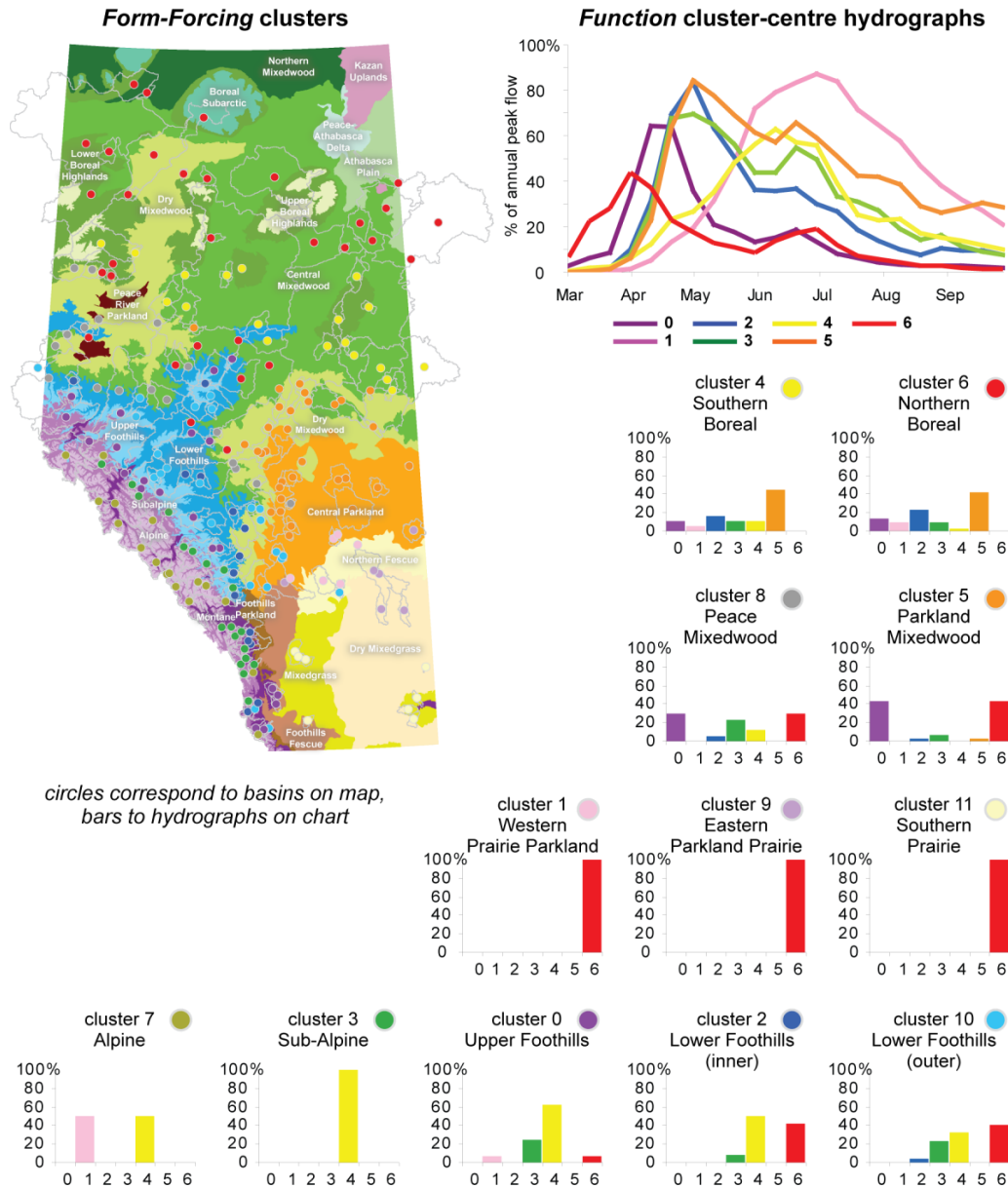
The *Forcing* GSOM identified fifteen climate clusters, which were closely associated with latitude and elevation. It is acknowledged that this was somewhat self-confirming, given that the training dataset had itself originally been generated using the ClimateWNA software (Wang *et al.*, 2012), which downscales the PRISM regional re-analysis for Western Canada (Daly *et al.*, 2002) by spatial interpolation. However, this was the only practical option available.

The percentage spatial cover of each of the 120 theoretically possible combinations of juxtaposed *Form* and *Forcing* classes was computed for every catchment, and this dataset was used to train a further GSOM. Twelve clusters were identified, which were evidently associated with distinct spatial domains across the study area (Figure 3). To confirm that these provided a meaningful representation of physiographic and climatological characteristics, the fractional cover of the twenty-one Alberta NSRs was computed for every catchment, and plotted for the basins in each identified cluster. The rationale here was that the NSRs had been derived largely through qualitative expert judgement (Government of Alberta, 2006) to combine aspects of physiography, climate and ecology, and therefore provided a 'benchmark' of landscape against which to assess this unsupervised, and arguably more objective, classification. The resultant distributions showed clear associations between the two schemes.

The *Function* classification yielded seven clusters of gauges exhibiting broadly consistent similarities in their annual hydrograph shapes from year to year. The long-term mean hydrographs of the members of each cluster also showed distinct characteristics, which may be related to the principal causative influences on their flow distributions, such as snow-melt, rainfall or glacial melt (Pardé, 1933; Lvovich, 1938).

When the membership of the catchments within the *Form-Forcing* and *Function* classes were compared, very clear associations were immediately evident between the two (Figure 3). The similarity of flow-regime distributions in the Boreal Forest, Parkland / Mixedwood and Prairie landscape groups, and the gradual transition in the fractional representation of each *Function* class with diminishing elevation from the high alpine to the lower foothills, are particularly interesting. Note also that although *Function* classes 3 and 5 have quite similar shapes, with a peak relatively late in the spring and slow decline through the summer, they occur in distinct settings: analysis of the *Form* and *Forcing* attributes of the catchments with which they are each associated strongly imply that the former results from higher summer precipitation in the northern foothills, while the latter is

driven by wetland fill-and-spill in the Boreal Forest. The broad summer peaks in the alpine basins of the Rocky Mountains and their higher foothills are inferred to result from a combination of an extended late snow-melt season at these elevations, large amounts of summer precipitation, lacustrine spill-and-fill, and glacial melt.



**Figure 3:** Spatial distribution of the twelve *Form-Forcing* classes (contextualised by NSRs), and fractional representation of the seven *Function* classes within each

#### 4. Conclusion

This paper describes the potential of SOMs to identify relationships between climatological inputs, physiographic processing, and streamflow outputs, using readily-available data from disparate sources. Importantly, it distinguished between catchment clusters possessing similar hydrometric profiles but contrasting physiographic / climatological attributes. While this amounts so far only to

the identification of broad associations between relatively coarse groups of catchments and flow distributions, subsequent refinement is expected to deliver more informative links between catchment descriptions and specific regimes in this geographic context. It has also provided independent support for the association of catchment descriptions based on ecological regions with flow regimes, as suggested by research conducted in the neighbouring Canadian province of British Columbia (Trubilowicz *et al.*, 2011). By extending similar analyses further afield and continuing to develop these techniques, it may be possible to make progress towards rigorous and objective classification schemes which are valid at continental or even global scales.

## 5. Acknowledgements

This paper describes part of a study submitted as a dissertation for the degree of MSc in GIS, supervised jointly by Dr A. Heppenstall (University of Leeds) and Dr J. Leyland (University of Southampton). The study was generously supported by funding made available by Dr A. Anderson of the Foothills Research Institute, Hinton, Alberta.

## 6. Biography

- BSc, Geology, Bristol (1986)
- Developer of 'hydrogeoinformatics' software, primarily for the NERC Centre for Ecology and Hydrology / Wallingford HydroSolutions (1993 – 2013)
- MSc, GIS (Online Distance Learning), Leeds / Southampton (2013)
- Now a (rather elderly) PhD Candidate: University of Northern British Columbia, Prince George, BC, Canada

## References

- Alahakoon D., Halgamuge S.K. and Srinivasan B. (2000) Dynamic self-organizing maps with controlled growth for knowledge Discovery *IEEE Transactions on Neural Networks* 11(3), pp. 601-614
- Amarasiri R., Alahakoon D. and Smith K.A. (2004) HDGSOM: a modified growing self-organizing map for high dimensional data clustering *Proceedings of the Fourth IEEE International Conference on Hybrid Intelligent Systems, December 2004*: pp. 216-221
- András K. (2008) Linguistic pattern recognition *Mathematical Linguistics* Chapter 8: pp 201-217 Springer London
- Beven K.J. and Kirkby M.J. (1979) A physically based, variable contributing area model of basin hydrology *Hydrological Sciences* 24,1,3
- Beven K.J. (2000) Uniqueness of place and process representations in hydrological modelling *Hydrology and Earth System Sciences* 4(2): 203-213
- Daly C., Gibson W.P., Taylor G.H., Johnson G.L. and Pasteris P. (2002) A Knowledge-Based Approach to the Statistical Mapping of Climate *Climate Research* 22: pp. 99-113.
- Ehret U. and 20 others (2014) Advancing catchment hydrology to deal with predictions under change *Hydrology and Earth System Sciences* 18: pp. 649-671
- Government of Alberta Natural Regions Committee (2006) Natural Regions and Sub-Regions of Alberta. Compiled by D.J. Downing and W.W. Pettapiece. Government of Alberta Pub. No. T/852. URL: [http://www.albertaparks.ca/media/2942026/nrsrcomplete\\_may\\_06.pdf](http://www.albertaparks.ca/media/2942026/nrsrcomplete_may_06.pdf) (Retrieved 8<sup>th</sup> January 2015)
- Kalteh A.M., Hjorth P. and Berndtsson R., (2008) Review of the Self-Organizing Map (SOM) Approach in Water Resources: Analysis, Modelling and Application *Environmental Modelling and Software* 23: pp. 835-845
- Kohonen T., (1982) Self-Organized Formation of Topologically Correct Feature Maps *Biological Cybernetics* 43: pp. 59-69
- Kohonen T. (1990) The Self-Organizing Map *Proceedings of the IEEE* 78(9): pp. 1464-1480

- Lvovich M.I. (1938) Opyt Klassifikatsii Rek SSSR *Trudy GGI*, Vyp. 6, Leningrad
- McDonnell J.J. and Woods R. (2004) Editorial: On the need for catchment classification *Journal of Hydrology* 299: 2–3
- Pardé M. (1933) (sometimes quoted as a later edition, 1955: 5<sup>th</sup> edition 1968) *Fleuves et Rivières* Paris: Armand Colin
- Sivakumar B.M., Singh V.P., Berndtsson R. and Khan S.K. (2015) Catchment Classification Framework in Hydrology: Challenges and Directions *Journal of Hydrologic Engineering* 20(1): SPECIAL ISSUE: Grand Challenges in Hydrology, A4014002
- Sivapalan M. (2005) Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale *Encyclopedia of Hydrological Sciences, Chapter 13* (Anderson M.G. – ed.) 2008 John Wiley & Sons, Ltd.
- Troch P.A., Carrillo G.A., Heidbüchel I., Rajagopal S., Switanek M., Volkmann T.H.M. and Yaeger M. (2008) Dealing with Landscape Heterogeneity in Watershed Hydrology: A Review of Recent Progress toward New Hydrological Theory *Geography Compass* 2 (2008)
- Trubilowicz J.W., Moore R.D. and Buttle J.M. (2013) Prediction of stream-flow regime using ecological classification zones *Hydrological Processes* 27(13): pp. 1935-1944
- Wagener T., Sivapalan M., Troch P. and Woods R. (2007) Catchment Classification and Hydrologic Similarity *Geography Compass* 1(4): pp. 901–931
- Wang T., Hamann A., Spittlehouse D. and Murdock T.N. (2012) ClimateWNA - High-Resolution Spatial Climate Data for Western North America *Journal of Applied Meteorology and Climatology* 61: pp. 16-29
- Woods R. (2002) Seeing catchments with new eyes: Review of Spatial Patterns in Catchment Hydrology: Observations and Modelling (Grayson R. and Bloesch G. – eds) *Hydrological Processes* 16: pp. 1111–1113



# New approaches to measure the spatial structure(s) of cities

Dani Arribas-Bel <sup>\*</sup>1 and Emmanouil Tranos<sup>†</sup>1

<sup>1</sup>School of Geography, Earth and Environmental Sciences, University of Birmingham

January 18, 2015

## Summary

This paper uses mobile phone data for the city of Amsterdam to study the distribution of activity over space and time. The extent to which we can empirically learn about the spatial structure of cities is limited by the technology and data available at given point in time. Using new sources of data that did not exist only a few years ago and recent statistical approaches that exploit them in a fuller fashion, we are able to obtain a representation of the changing spatial structure of the city over the course of a year, a week and a day.

**KEYWORDS:** Urban spatial structure, space-time statistics, mobile phone data, big data, urban form.

## 1 Introduction

Our understanding of the spatial structure of cities has been shaped by the type of data available at each moment. Traditionally, researchers looking at the spatial distribution of activity within cities have relied on official sources. These datasets have a high degree of accuracy and representativeness but a low temporal resolution, usually being collected once every ten years in censuses. This degree of coarseness likely hide many patterns of relevance that are lost in-between observations, limiting how much we can learn about human activity within cities. In recent years, several technological advances have given rise to multiple new sources of data that promise to fill many of the gaps left by traditional datasets (Arribas-Bel, 2014). In this paper, we use mobile phone data for the city of Amsterdam to study the distribution of activity over space and time. We begin with the analysis presenting insights that one would expect to obtain from traditionally aggregated data to then move on to much finer disaggregation of mobile phone usage. Taking advantage of these new data also require modification in the methodological approaches and, to this end, we adopt not only traditional tools from spatial analysis but more modern space-time approaches. This additional layer of detail allows us get insights about the changing shape of activity within the same city,

---

<sup>\*</sup>D.Arribas-Bel@bham.ac.uk

<sup>†</sup>E.Tranos@bham.ac.uk

within a day and a within a week, that would have been missed if only a traditional dataset was available.

There is a longstanding literature in quantitative geography and urban economics focusing on the measurement and study of the spatial structure of cities (Anas et al., 1998). Most of this works relies almost exclusively in some form of official data, be it census or transportation datasets, provided by official agencies at a spatial and temporal aggregated level. Initial approaches such as that outlined in Giuliano and Small (1991) were highly influential and, although relied on simple and often ad-hoc measures, seeded the way for more sophisticated analysis. In the early 2000's, McMillen (McMillen, 2001, McMillen and Smith, 2003, McMillen, 2004) significantly advanced the field by including more advanced methods based on non-parametric techniques such as geographically weighted regression (GWR). In more recent years, the efforts have been split between sophisticating the methods further (e.g. Redfearn, 2007) and applying them in empirical contexts that substantially broaden the scope of the areas covered (e.g. Lee, 2007, Arribas-Bel and Sanz-Gracia, 2014).

Urban analysis based on data from mobile phone operators or other big data sources provides new opportunities to urban analysis as it enables researchers to model and gain a deeper understanding of the pulse of the city (Batty, 2010). Analysis using the above data do not focus on the physical form of cities, but on human activity and most importantly, on how citizens use cities. Researchers have now the ability to utilise such data due to pervasiveness of digital technologies which resulted to huge pools of human behavioural data. Urban analysts are now able to use such data as a tool to understand the structure of cities (Louail et al., 2014). The value added of such data is related with their granularity both in terms of space and time. The latter enables researchers to study the dynamics of urban structure and how cities are perceived and used by citizens over time. The results of such research can support urban planning and generate new opportunities for the management of cities. For instance, according to Ahas and Mark (2005) geo-located data from mobile phone operators can be utilized in monitoring the usage of transport infrastructure, in studying and quantifying the temporal dimensions and the dynamics of urban space, and in planning and designing transportation and transport infrastructure.

This paper draws upon a fast developing research domain which utilises data from mobile phone operators in analysing and modeling cities. The main lesson from this research strand is data from mobile phone operators offers the possibility to study micro- and macro-behaviors and truly reflect human behavior given the fact that data is becoming more and more available (Calabrese et al., 2014, p. 25:4). In other words, such data reflect the collective behaviour of people (Calabrese et al., 2010). In another paper, Reades et al. (2009) identified a strong relationship between human activity and aggregated mobile phone usage using the city of Rome as a case study. More recent, Sevtsuk and Ratti (2010) used data for mobile phone activity as a proxy to model population distribution over time and space and Jacobs-Crisioni et al. (2014) employed such data to assess the impact of land-use density and mix on urban activity patterns.

## 2 Data

The data used for this paper has been provided by one of the major mobile phone operators in The Netherlands. It is an aggregated dataset of individual mobile phone activity and includes telecommunication counts at the level of the GSM (Global System for Mobile Communications) zones on an hourly basis for 2010. 815 such zones are included in the analysis which represent the coverage areas of GSM antennas. These zones are represented by irregular polygons which vary in shape and size, the design of which supports the function of the GSM network. For instance, smaller GSM zones can be observed in the centre of Amsterdam as this is a busier area and therefore GSM antennas accommodate smaller areas. In regards to the telecommunications counts, the main focus of the paper lies on the number of Erlangs. This is a measure of telecommunication activity: 1 Erlang can consist of either one phone call of 60 minute duration or of two 30 minute phone calls (Sevtsuk and Ratti, 2010).

## 3 Methods

More detailed data both in space and time do not automatically translate into more detailed insights. To leverage the full power and advantages of mobile phone data, it is necessary to include methodologies that recognize such degree of detail and are able to cope with it. To show this in an intuitive way, we adopt an incremental approach that begins with cross-section spatial methods applied to a completely time-aggregated version of the dataset. In particular, we use the widely adopted local indicators of spatial association (LISAs, e.g. Anselin, 1995). This stage is meant to show what a traditional analysis would be able to find and thus represents a benchmark against new insights will be compared. We then begin to disaggregate the data over more fine-grained temporal scales. First we calculate similar LISA maps for hourly slices over a day and over a week. These allow to start thinking about the idea of an evolving spatial structure, hidden to traditional approaches. Finally we embrace the space-time nature of mobile phone data by adopting some of the more recent statistical techniques such as Kulldorff's scan statistics (Kulldorff et al., 2005, Kulldorff, 2014).

## 4 Acknowledgements

The authors would like to acknowledge the support of John Steenbruggen and the Dutch Ministry of Infrastructure and the Environment (RWS) regarding data acquisition.

## 5 Biography

**Dani Arribas-Bel** is lecturer in Human Geography at the University of Birmingham. He is interested in quantitative spatial methods and urban analysis. In particular, he is interested in how the computational advances and explosion of data witnessed in the last decades can contribute to the

understanding of the spatial structure of cities. He is also involved in the open-source community, mostly as a core developer of the scientific library of advanced spatial analysis PySAL.

**Emmanouil Tranos** is lecturer in Human Geography at the University of Birmingham. He is an economic geographer focusing primarily on digital geographies. He has published on issues related with the spatiality of the Internet infrastructure, the economic impacts that this infrastructure can generate on space and the position of cities within spatial, complex networks. Recently, he has been focusing on the use of big data of high spatio-temporal resolution (e.g. mobile phone data) in urban and regional analysis.

## References

- Ahas, R. and Mark, Ü. (2005). Location based services new challenges for planning and public administration? *Futures*, 37(6):547–561.
- Anas, A., Arnott, R., and Small, K. (1998). Urban spatial structure. *Journal of Economic Literature*, 36(3):1426–1464.
- Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis*, 27(2):93–115.
- Arribas-Bel, D. (2014). Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities. *Applied Geography*, 49:45–43.
- Arribas-Bel, D. and Sanz-Gracia, F. (2014). The validity of the monocentric city model in a polycentric age: Us metropolitan areas in 1990, 2000 and 2010. *Urban Geography*, 35(7):980–997.
- Batty, M. (2010). The pulse of the city. *Environment and Planning B: Planning and Design*, 37(4):575–577.
- Calabrese, F., Di Lorenzo, G., and Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 312–317. IEEE.
- Calabrese, F., Ferrari, L., and Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys (CSUR)*, 47(2):25.
- Giuliano, G. and Small, K. (1991). Subcenters in the Los Angeles Region. *Regional Science and Urban Economics*, (21):163–182.
- Jacobs-Crisioni, C., Rietveld, P., Koomen, E., and Tranos, E. (2014). Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: an exploratory study using mobile phone data. *Environment and Planning A*, 46(11):2769–2785.
- Kulldorff, M. (2014). Satscan v 9.3: Software for the spatial and space-time scan statistics. Technical report.

- Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., and Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine*, 2(3):e59.
- Lee, B. (2007). “Edge” or “Edgeless Cities”? Urban Spatial Structure in US Metropolitan Areas, 1980 to 2000. *Journal of Regional Science*, 47(3):479–515.
- Louail, T., Lenormand, M., Cantú, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J. J., and Barthélemy, M. (2014). From mobile phone data to the spatial structure of cities. *arXiv preprint arXiv:1401.4540*.
- McMillen, D. (2001). Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3):448–473.
- McMillen, D. (2004). Employment densities, spatial autocorrelation, and subcenters in large metropolitan areas. *Journal of Regional Science*, 44(2):225–244.
- McMillen, D. and Smith, S. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics*, 53(3):321–338.
- Reades, J., Calabrese, F., and Ratti, C. (2009). Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836.
- Redfearn, C. L. (2007). The topography of metropolitan employment: Identifying centers of employment in a polycentric urban area. *Journal of Urban Economics*, 61(3):519–541.
- Sevtsuk, A. and Ratti, C. (2010). Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60.

# Big Data Analysis of Population Flow between TfL Oyster and Bicycle Hire Networks in London

N. Sari Aslam<sup>a</sup>, J. Cheshire<sup>b</sup>, T. Cheng<sup>c</sup>

<sup>a b c</sup> University College London(UCL), Department of Civil, Environmental and Geomatic Engineering, Gower St, London, UK

10 March 2015

## Summary

This study seeks to undertake an initial analysis of the likely flow of people between the Tube to bicycle hire network in London. Data for the two networks were extracted for a month (April and June 2012) in order to establish the strength of the relationship between them. The results quantify the extent to which Tube commuters impact the capacity utilization of the bicycle network. We expect this research to have implications in the expansion and maintenance of bicycle hire in London and similar schemes around the world.

**KEYWORDS:** Big Data, Oyster Data, Bicycle share system, Time Series Analysis, Regression Analysis

## 1. Introduction

In London, 24 million journeys are completed on Transport of London's (TfL's) network each day (Transport for London, 2012). The vast majority of these are by bus and Tube but a growing number of travellers make use of London's bicycle hire (O'Brien et al. 2014). The nature of a bicycle hire network is completely unique compared to other transport networks, because the service providers have little or no control over the key resource i.e. bicycle. The challenge is to optimise this network resource utilization based on usage behaviour of the bicycle users.

The rapid pace of technological advances and the availability of huge amounts of data from transport networks have made it possible to analyse population flows in great detail. Billions of rows of continuous and non-invasive data with spatial and temporal dimensions is now available in the public domain (Beecham & Wood, 2013; Blythe & Bryan, 2007; Kusakabe, Iryo, & Asakura, 2010; Lathia, Ahmed, & Capra, 2012; Pérez, Trépanier, & Morency, 2011). Given the huge volumes of data now available, it has become challenging to undertake analysis using conventional statistical software (Blackwell & Sen, 2012). It is therefore often necessary to either subset the data or perform some kind of aggregation to reduce data size.

---

<sup>a</sup>n.aslam.11@ucl.ac.uk,

<sup>b</sup>james.cheshire@ucl.ac.uk

<sup>c</sup>tao.cheng@ucl.ac.uk

## 2. Data Description

The users of London transport network whose usage behaviour is repetitive and can be modelled are the main focus of this study. Integrating and analysing the data from train and bicycle hire networks provide an opportunity to understand the usage behaviour and allow efficient allocation of the resources.

Tube to bicycle: Because of the relative size of the networks, a large influx of Tube users can impact significantly on the capacity utilization of a bicycle hire network. To show the strength of this relationship, the analysis considered exit (Tube stations) to exit (docking stations) data.

Bicycle to train: To gauge the strength of the relationship in the reverse direction, i.e. from bicycle on to train to establish if users coming into the station on bicycle are continuing with their commute via trains. The analysis will consider entry (docking stations) to entry (train stations) data.

The cycle hire data is for the individual journeys from one docking station (origin) to another (destination). In order to focus on the specific time windows it was decided that journeys should be aggregated into 15 minute time intervals. It resulted in two records per docking station per 15 minute period. One record for aggregate 'entry' terminating at the station and the second for aggregate 'exit' from a docking station. For the purpose of this analysis the multiple docking station data have been aggregated based on proximity to the station.

Oyster data available for this analysis was aggregated at 15mins intervals and were provided by TfL. In order to match the bicycle data, all the journeys terminating at a given station within a period were aggregated into one 'entry' record. All the journeys starting from a station within a period were aggregated in one 'exit' record.

The processed data could be classified as below, and were used for the following analysis:

- Aggregate Tube exists
- Aggregate Tube entries
- Aggregate bicycle docking station exits
- Aggregate bicycle docking station entries
- Capacity of the bicycle docking station (the number of bicycles available for use/exit and return/entry)

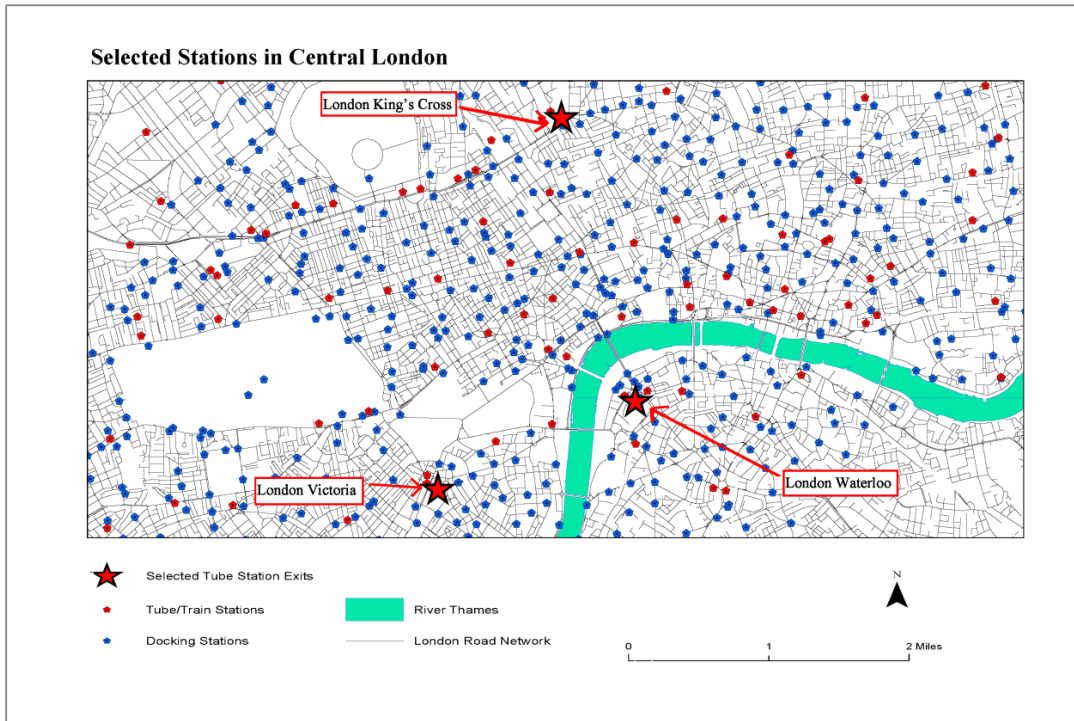
## 3. Methodology

Central London provides the focus for this study. It sees the largest peaks in commuter flows as well as being the primary destination for tourism and leisure.

The study started with the trend analysis of the time series of the two networks, to understand obvious patterns in the data. To look further into this relationship, the Pearson correlation coefficient was calculated for time series data for the two networks, this quantified the strength of the relationship between the two networks. This was followed by linear regression to show the trend lines using docking station data as the dependent variable. The study started with a daily, followed by weekly and monthly analysis of the three selected Tube stations and their corresponding docking stations.

### 3.1. Network Analysis

The population flow between the Tube and Bicycle hire network depends upon the proximity of the bicycle docking station to the Tube station. 'Closest facility' network analysis was conducted to find the docking stations in close adjacency to the Tube stations for the available data (747 Docking stations and 163 Tube Stations) and to identify the shortest path between them. A maximum of five docking stations were selected within 300m (walking distance) of the selected Tube stations.



**Figure 1:** Study Area of the selected stations of London Waterloo, Victoria and King's cross.

In order to undertake more detailed analysis, three Tube stations - London Waterloo, London King's Cross and London Victoria - were selected (shown as stars in Figure 1) and their details described below.

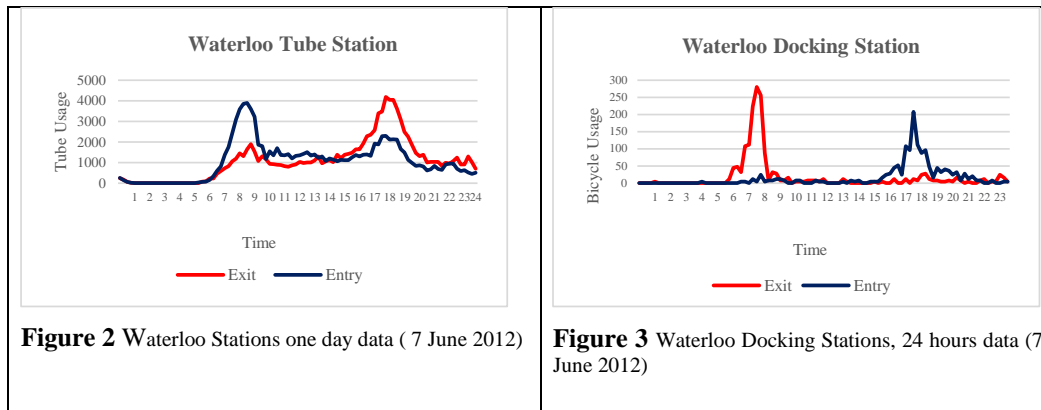
**Table 1:** The list of the Tube and docking stations

Name	Station Exits	Bicycle Hire Docking Stations
Waterloo	Shell Gates Main Gates Auxiliary Gates W&C Validators Jubilee Gates	Waterloo Station 1, Waterloo Waterloo Station 2, Waterloo Waterloo Station 3, Waterloo
Victoria	District Gates Main Gates	Ashley Place, Victoria Cardinal Place, Victoria
Kings Cross	Met Main Entry Gates Tube Gates Thameslink Gates Northern Ticket Hall	St. Chad's Street, King's Cross Belgrove Street, King's Cross Northdown Street, King's Cross

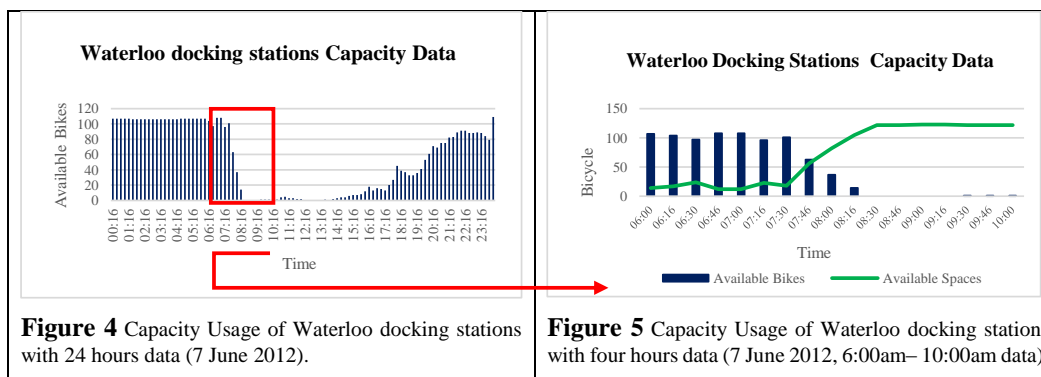
### 3.2. Daily Data (7 June, 2012)

To understand the daily pattern of people flow between the two networks, data points are plotted for the three selected stations in Central London. From Tube station to docking station AM exit data (6am-10am) has strong relationship for all three stations. From Docking stations to Tube station PM entry data (5pm-9pm) has strong relationship, but not as significant as in AM peak. Capacity utilization graphs also support the same results.



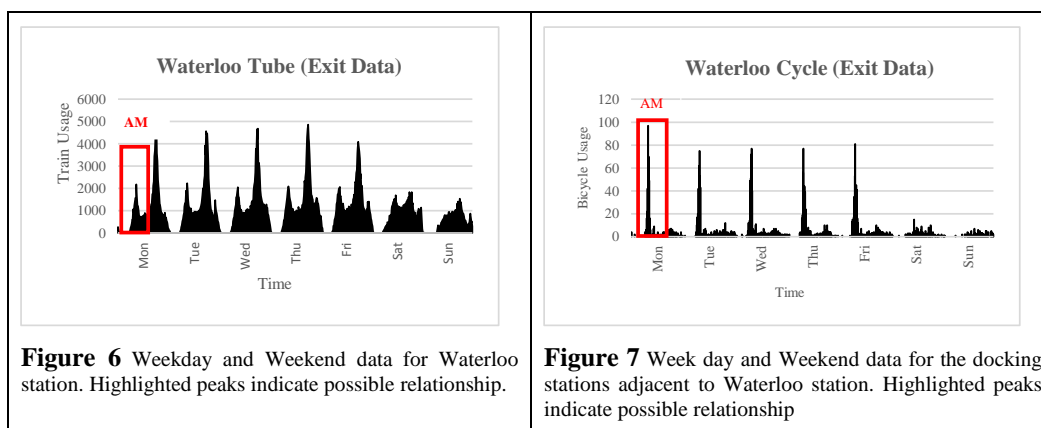


Capacity utilization of the bicycle hire network, as shown in Figures 4 and 5, follows the trend of the Tube network rush hours, but also highlights how the lack of capacity impact the relationship (adversely).



### 3.3. Weekly Data

Results plotted for the same three stations from the 18 June to the 24 June show the separation between the weekend and weekday trends at 15 minute intervals. Figure 6 shows two peaks for Tube in both AM and PM and PM peak is higher than the AM. Figure 7 highlight the AM peak data for bicycle users. The PM peak is less obvious for the bicycle population flow as it may be due to national rail commuters using Tube for the first leg of their journey. The relationship is AM peak between two networks.



### 3.4. Monthly Data (April, 2012)

The analysis for the month was carried out to highlight the relationship between docking station and train station over a longer duration. The results for the four weeks period further emphasised the weekday and weekend trends visible in weekly data.

After investigating the daily, weekly and month trends, further insight can be gained through calculating the Pearson correlation coefficient (defined in Equation 1).

$$\text{Correlation}(r) = \frac{\text{Cov}(x,y)}{\text{std.dev}(x) \cdot \text{std.dev}(y)} \quad (\text{Equation 1})$$

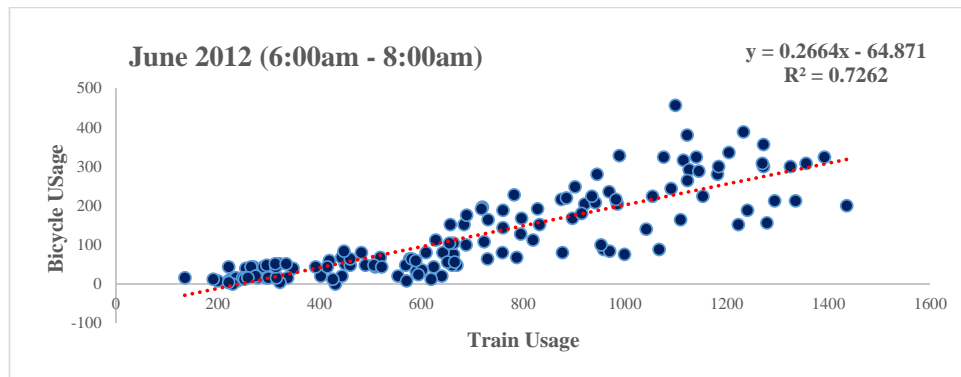
where x is the number of population using train stations at an hour interval (train usage), and y is the number of population using bicycle docking stations at an hour interval (bicycle usage).

**Table 2:** Correlation coefficients using month AM peak data all days (6:00am - 10:00 am) excluding weekends in Waterloo Tube/Train Stations and Docking Stations

Morning Peak/Morning Activities				
Station to Docking S.	0600-0700	0700-0800	0800-0900	0900-1000
Waterloo (Exit)	0.88534	0.90249	0.90056	0.39092
Evening Peak/Evening Activities				
Station to Docking S.	1700-1800	1800-1900	1900-2000	2000-2100
Waterloo (Exit)	0.23922	0.42573	0.27059	0.41861

Description	Range
Very Weak	0.01 – 0.19
Weak	0.20 – 0.39
Modest	0.40 – 0.69
Strong	0.70 – 0.89
Very Strong	0.90 – 0.99

Linear regression was conducted by assuming docking stations time series data as the dependent variable and train station data as the independent variable. The chart (Figure 8) shows two-hour intervals over a period of months excluding weekends. It shows a good fit with a coefficient of determination of 72% explaining all the variability of the data around its mean.



**Figure 8** Linear regression applied to Waterloo AM Exit (0600-0800)

#### 4. Conclusions

This paper was set out to study the relationship between the bicycle hire network and the TfL Oyster network. The results contain few surprises with the strength of the relationship closely linked to the rush hour commuting patterns.

There is a dip in the correlation after the rush hours, initial perception was that it is due to the drop in the number of commuters from the Tube, but it has been observed that the lack of available bicycles also adversely impact the relationship as shown in figure 4. Although the current report only focuses on three major Tube stations, the methods are easily expanded to cover the entire London Underground network.

Cycling as a government policy always has a positive impact on the environment, health and economy. Journeys made from bicycles instead of other modes of transport makes cities less congested, cut transport and delivery costs, reduce illness-related expenditure and make people fitter and the environment cleaner (“Position Paper of the European Cyclists’ Federation,” n.d.). The analysis in this paper provides an insight into the user behaviour of the bicycle hire network and it will allow future infrastructure investment decision to be made in an informed manner.

#### 5. Acknowledgements

I am grateful to the Economic and Social Research Council for funding my studentship at UCL. Furthermore, I owe a special thanks to Prof Paul Longley for his support during the paper preparation.

#### 6. Biography

Nilufer Sari Aslam completed MSc Geographic Information Science at UCL in 2013. Currently she is in MRes at Urban Sustainability and Resilience at UCL to contribute how big data can be analysed using statistical approaches. Nilufer’s research interests are big data analysis, spatial temporal analysis, network analysis and demand modelling.

#### 7. References

- Beecham, R., & Wood, J. (2013). Exploring gendered cycling behaviours within a large-scale behavioural data-set. *Transportation Planning and Technology*, 37(1), 83–97. doi:10.1080/03081060.2013.844903
- Blackwell, M., & Sen, M. (2012). Large Datasets and You: A Field Guide. *Manuskript*, 14627, 1–8. Retrieved from <http://www.mattblackwell.org/files/papers/bigdata.pdf> npapers3://publication/uuid/9B8CCBB4-E848-4D7E-9F6E-14AE5A96FC0C
- Blythe, P., & Bryan, H. (2007). Understanding behaviour through smartcard data analysis. *Proceedings of the ICE - Transport*, 160(4), 173–177. doi:10.1680/tran.2007.160.4.173
- Kusakabe, T., Iryo, T., & Asakura, Y. (2010). Estimation method for railway passengers’ train choice behavior with smart card transaction data. *Transportation*, 37(5), 731–749. doi:10.1007/s11116-010-9290-0
- Lathia, N., Ahmed, S., & Capra, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, 22, 88–102. doi:10.1016/j.trc.2011.12.004
- O’Brien, O., Cheshire, J., & Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 34, 262–273. doi:10.1016/j.jtrangeo.2013.06.007

Páez, A., Trépanier, M., & Morency, C. (2011). Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards. *Transportation Research Part A: Policy and Practice*, 45(7), 640–652. doi:10.1016/j.tra.2011.04.002

Position Paper of the European Cyclists ' Federation. (n.d.).

Transport for London Transport for London. (2012), 1–3.

# Researching long-run trends in South East England 1931-2011 for the European Union and Greater London Authority

Paula Aucott<sup>\*1</sup> and Humphrey Southall<sup>†1</sup>

<sup>1</sup>Department of Geography, University of Portsmouth

June 7, 2015

## Summary

This paper describes the sources, methods and preliminary results of two related projects on historical census data funded by government bodies for policy purposes. Both required data for diverse historical reporting areas to be redistricted to a single set of modern units. All redistricting is done by a simple vector overlap method, but this requires boundary data for both the modern and the historical units; and, as far as possible, that the historical units be more detailed than the modern ones. Even for recent periods, locating boundary maps is often much harder than locating statistics.

**KEYWORDS:** historical GIS, policy relevance, redistricting, census

## 1. Introduction

Predicting long-term trends decades ahead requires data stretching back decades. Through its censuses since 1801, the UK has been gathering just such data, but the analytic potential is poorly exploited: “modern” census research generally looks just one or two censuses back, while “historical” census research focuses on the period 1851-1911 when reporting geographies were consistent.

This paper presents two highly applied census projects. The first, funded by the European Union, has created time series 1951 to 2011 for the total populations of the 8,941 Wards of Great Britain as used by the 2011 census. The second, funded by the Greater London Authority (GLA), is redistricting a diverse data from censuses 1801-1961, creating consistent data for the current London Boroughs, for London’s Central Activity Zone (CAZ) as defined by the GLA, and for the overall GLA area; but we focus here on constructing consistent data on London’s industrial structure, converting diverse historical classifications to Standard Industrial Classification 2007 (SIC).

Note that the statistical results presented here are preliminary and subject to revision.

## 2. Estimating historical populations of 2011 Wards

Earlier work for *Vision of Britain* ([www.VisionofBritain.org.uk](http://www.VisionofBritain.org.uk)) included redistricting to modern units, but only to Britain’s 408 districts. Redistricting to a geography twenty times more detailed requires more detailed input data, and different sources are used for each census year.

2011 and 2001 are not our concern, data being available from the Office of National Statistics. For 1991 and 1981 population counts and vector boundaries are available for 103,419 (1991) and 105,598 (1981) English Enumeration Districts, plus Welsh and Scottish EDs, so redistricting to 8,941 British

---

<sup>\*</sup> Paula.Aucott@port.ac.uk

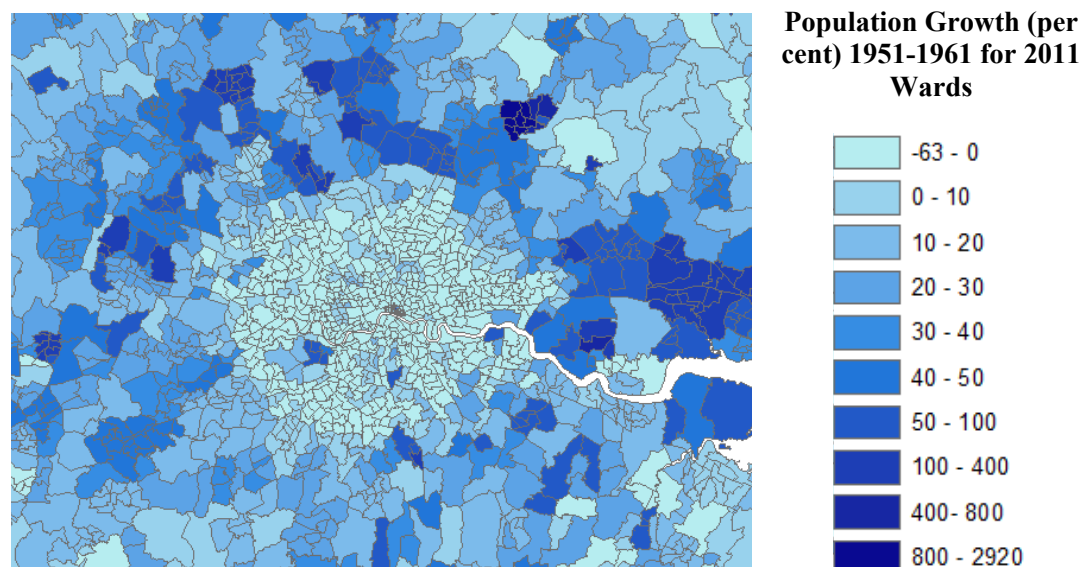
<sup>†</sup> Humphrey.Southall@port.ac.uk

Wards is unproblematic. The historical statistics are associated with the boundaries of their contemporary administrative units and converted to the required output projection (where necessary). The historical polygons are spatially joined to the boundaries of the modern polygons, the population figures are interpolated and finally the boundaries are dissolved to give modern units with weighted statistics.

1971 is somewhat more complex. After significant investigation no digital mapping of real ward or enumeration district boundaries for this year has been found to exist. Instead we use the closest alternative, Thiessen polygons for Wards which were created by aggregating Census Enumeration Districts which themselves were generated from a centroid point dataset. The same processing method as above was employed, although the artificial boundaries meant some mismatching had to be manually corrected.

1961 was the first census year that many urban areas were “unparished”, meaning that unless we map Ward boundaries, cities as large as Birmingham, with over a million people, are single polygons. We addressed this issue by seeking ward maps from the modern councils, archives and local studies libraries. The GLA provided a paper map covering all London Boroughs and Figure 1 shows the much higher density of polygons for wards in central London in comparison to the larger civil parish polygons in adjacent counties.

For towns where we could not obtain boundary maps, we located Wards as points then built Thiessen polygons. Overall, this is the most detailed mapping of the 1961 census ever done, as 1960s analyses either mapped Wards as points (Ministry of Housing and Local Government, 1966) or dealt only with parishes (e.g. Osborne, 1966). Although the current analysis covers only total populations, this GIS could map the 1961 census microdata being restored by a project at Essex University.



**Figure 1** The Wards of Greater London in 1961, plus Civil Parishes in adjacent areas.

Table 1 provides a very simple summary of our calculations of population densities for constant geographical areas over sixty years, grouping wards by their distance from Nelson’s Column; we were able to include 1951 because the 1961 Census listed 1951 populations of 1961 wards and parishes. The distance zone of most rapid growth is highlighted in bold, and the zone of least growth, or contraction, is italicized. During the first four decades the zone of highest growth is clearly moving outwards through and beyond the Green Belt. The innermost zone experienced the most rapid decline up to 1991, but then very strikingly became the zone of maximum growth.

**Table 1** Mean Decennial Population Growth within a 50 mile radius of central London

Distance (Miles)	Growth 1951-61	Growth 1961-71	Growth 1971-81	Growth 1981-91	Growth 1991-2001	Growth 2001-11
5	-5.22	-16.50	-17.61	-9.61	<b>22.75</b>	<b>16.54</b>
10	-4.07	-4.96	-6.45	-4.95	12.42	14.16
15	12.23	5.67	-2.69	-4.48	7.28	8.67
20	33.88	8.20	-1.50	-3.71	9.09	9.42
25	<b>41.21</b>	19.98	4.95	-1.14	7.02	6.32
30	32.43	<b>31.73</b>	9.44	-0.01	8.67	8.41
35	17.46	23.66	10.14	3.64	8.08	5.64
40	10.03	21.95	12.16	3.61	11.33	8.88
45	10.76	4.75	<b>13.61</b>	<b>5.43</b>	12.72	11.20
50	12.89	27.29	7.82	5.32	11.57	9.84

### 3. Longer-run trends in industrial structure

A wide range of census variables are available via the Small Area Statistics from 1971 onwards, and similar data are becoming available from transcriptions of individual level census returns via Essex's Integrated Census Microdata system (<http://icem.data-archive.ac.uk>). However the latter is legally limited to data over a hundred years old, so long run overall perspectives must draw on the published census reports.

The Greater London Authority need data on industrial structure, so our work is drawing on four pre-digital censuses: 1841, 1881, 1931 and 1951. However, work on the first two is incomplete while we can include comparable data for 1971 and 2011. For 1931 and 1951 the historical geography is simpler than in the population analysis, data being reported for the pre-1974 system of local government districts, but instead of simple population counts we have to work with a different industrial classification for each date, and re-classify each to the Standard Industrial Classification (2007) used by the 2011 Census.

Despite working with 438 detailed categories in 1931 and 160 in 1951, they describe manufacturing in much greater detail than services and some SIC "Sections", such as "Information and Communication", cannot be identified at all. Once data are re-classified, they are redistricted using a Geography Conversion Table derived from the parish-level table from the relevant census, our equivalent Civil Parish GIS coverage, and the available 2011 Local Authority District boundaries.

**Table 2** Changes in Industrial Structure (SIC 2007) 1931-2011

			C:	G:	H:	K:	O:
			Manuf-	W'sale	Transport	Finance	Public
Area	Year	Total	acturing	+Retail	+ Storage	+Insur.	Admin-
				Trade			istration
Greater London	1931	3,688,129	1,141,278	649,325	372,833	112,622	188,862
	1951	4,145,021	1,443,653	633,725	418,131	185,295	311,408
	1971	3,921,180	1,010,450	612,620	422,210	251,670	332,270
	2011	4,500,481	142,654	550,529	242,411	409,904	265,069
City of London	1931	9,534	2,354	1,716	573	473	548
	1951	322,052	66,130	57,767	51,957	90,940	7,822
	1971	336,490	45,350	27,380	51,050	134,640	11,520
	2011	356,706	2,864	14,242	4,832	163,425	7,105

Table 2 shows preliminary results for five selected SIC 2007 Sections, which have in fact been calculated for all the individual boroughs. Note that the 1931 data are unavoidably based on place of residence rather than work, with large consequences for the City of London data.

#### 4. Conclusion

The Great Britain Historical GIS project has been working for over twenty years with historical census data. Our experience shows it is quite possible for historical GIS research to achieve significant non-academic “impact”, and to draw on a wider range of funding, but significant adaptations in approach are necessary. Firstly, research must come up to the present, and achieving long runs of data requires new skills: disinterring obscure statistical datasets; locating even more obscure boundary maps; even negotiating copyright.

Secondly, both policy makers and the general public have more need for local time series than for maps. Further, policy makers almost always require time series for modern units, even though it is generally easier to redistrict modern small area data to less detailed historical units. This means that once the above ingredients are assembled GIS techniques must be used to redistrict these diverse data sets to a single constant geography.

Despite the extensive research into complex redistricting algorithms, we use simple vector overlay – “cookie cutter” – methods for two reasons (Simpson, 2002). Firstly, the improvement in accuracy from more complex methods is limited, and finding more detailed historical datasets uses our time better. Secondly, these simpler techniques are more easily explained to non-technical audiences. Our focus is on real world use, both the public bodies funding the work described here and the general public accessing our web site: *Vision of Britain* had 1.65m. “unique visitors” in 2014 and the most accessible statistical content is redistricted data for modern districts.

#### 5. Acknowledgements

This research was funded by the European Union and the Greater London Authority. We are also grateful to the many local authorities who have provided us with 1961 ward boundary maps.

#### 6. Biography

Paula Aucott is a Senior Research Associate and doctoral candidate in the Department of Geography, University of Portsmouth; she has Masters degrees in English Local History and in GIS, and is the project manager for the Great Britain Historical GIS. Humphrey Southall is Professor of Historical Geography at the University of Portsmouth and director of the Great Britain Historical GIS.

#### References

Ministry of Housing and Local Government (1966) *Population change 1951-1961 by wards and civil parishes: compiled from the 1961 census. Scale 1:625,000*. Ordnance Survey, Southampton.

Osborne R H (1966) *Atlas of Population Change in the East Midland Counties 1951-1961*. Department of Geography, University of Nottingham.

Simpson L (2002). Geography conversion tables: a framework for conversion of data between geographical units. *International Journal of Population Geography*, 8(1), 69-82.



# Understanding Spatio Temporal Patterns of Crime Using Hotspot AND Coldspot Analysis

Bates E<sup>\*1</sup> and Mackaness W<sup>†2</sup>

<sup>1</sup>AQMeN, University of Edinburgh

<sup>2</sup>Institute of Geography, University of Edinburgh

9 January 2015

## Summary

This paper argues that we need to think as much about where crime does not happen as where it does. The use of hotspot maps is a widely accepted practice in policing. These maps highlight areas with high concentrations of crime but tell us less about areas with medium or low concentrations of crime. Understanding what makes a 'low crime place' may provide lessons for reducing crime. This paper proposes techniques which use a mixed method approach, combining LISA, Group Trajectory Analysis and Focus Groups, to give us a more nuanced and detailed understanding of crime at the neighbourhood level.

**KEYWORDS:** *Hot and Cold Spots; Mixed Method Spatio Temporal Analysis; Vandalism*

## 1. How best do we view neighbourhood crime?

Robert Sampson argues that:

“[Ne]ighborhoods vary in size and complexity depending on the social phenomenon under study and the ecological structure of the larger community...”(Sampson,2012,54).

Thus there is no one single scale that represents a neighbourhood. The Modifiable Areal Unit Problem (Openshaw, 1984) tells us that our choice of scale will affect the results of our analysis. This presents a particular challenge for representing neighbourhood level crime using maps; what scale(s) should we use? Weisburd and colleagues reveal considerable variation in trajectories of crime at the street segment level (Weisburd et al, 2012). Taylor argues for the use of multiple scales, taking both a top down and a bottom up view of crime (Taylor, 2010).

## 2. How do patterns of crime change over time?

Routine activity theory (Cohen and Felson, 1979) suggests that concentrations of crime are related to individual's routine behaviours. Routine behaviours can change over time across months, years and decades. Each of these activity changes can also impact on local levels of crime. This means we need to reflect temporal change in our crime maps too.

## 3. The Case Study

Here we assessed whether different local areas had consistently high or low levels of vandalism across time. Vandalism was chosen as it has rarely been the focus of research in recent years but remains a volume and signal crime impacting negatively on local communities. Focusing on a specific crime potentially reduced the complexity of possible explanations of crime concentrations.

### 3.1 The Study Area, Data and Scale

The study area was within the city of Edinburgh, managed by a single senior police officer, with a mix of physical and socio-economic environments. The data used were individual locations of recorded crimes of vandalism across a six year period from 1<sup>st</sup> April 2004 to 31<sup>st</sup> March 2010, along with qualitative data generated in focus groups. Recorded crime data were aggregated together into financial

---

\* ellie.bates@ed.ac.uk

† william.mackaness@ed.ac.uk

years (1<sup>st</sup> April to 31<sup>st</sup> March) at 100m by 100m grid square, data zone and output area.

### 3.2 What methods could allow us to explore hot and cold spots of crime across time?

This paper used three methods to explore levels of crime in local neighbourhoods:

- *'Talking to the map'* - Community police officers and their senior officer were invited in three separate focus groups to annotate maps. They were asked to shade maps based on their local knowledge and share their understanding and experience of the characteristics of crimes of vandalism. This technique was developed based on previous work which had asked police officers to designate areas of high crime (Ratcliffe & McCullagh, 2001; Craglia et al, 2005, Haining & Law, 2007) and using participatory GIS to elicit resident perceptions (Cinderby, 2009).
- *Local Indicators of Spatial Association (LISA)* – two indicators were used: Local Moran's I (Anselin, 1995; Assunção and Reis, 1999) and  $G_i^*$  (Getis & Ord 1992; Ratcliffe & McCullagh, 1999). Data were assessed for recorded crimes over six financial years.
- *Group Trajectory Analysis* using aggregate geographic data – This technique was adapted from work done by Groff et al, 2010. Two types of technique were used at two different scales. At 100m grid a negative binomial count model was used. At output area a categorical model was used to assess whether areas had constantly average levels of vandalism, around average or below average groups across time.

The *Talking to the Map* approach had the advantage that it did not rely on the choice of any predefined scale and allowed participants in the exercise to create their own boundaries. It enabled commentary by participants to be closely linked to places and specific locations. Limitations: 1) participants were constrained as they were asked to shade a pre-existing map; 2) the areas labelled as high or low crime areas relied upon police officer's personal knowledge and experience so may have been subject to bias and 3) the total number of focus groups was small.

Using *LISA* had the advantage that it provided a method of visualising statistically significant both hot and cold spots simultaneously; the method took direct account of spatial autocorrelations and results were relatively easy to present and explain to police officers. Limitations: 1) data was subject to the small number and multiple testing problems (Haining, 2003) which could only be partially resolved and 2) it did not provide a method of summarising crime trajectories.

Using *Group Trajectory Analysis (GTA)* enabled distinct trajectories across time to be assessed. It was possible to identify high, medium and low concentrations of crime at two scales, with two different methods. Limitations: (1) the technique did not control for spatial autocorrelation, (2) the small number problem was only partially resolved.

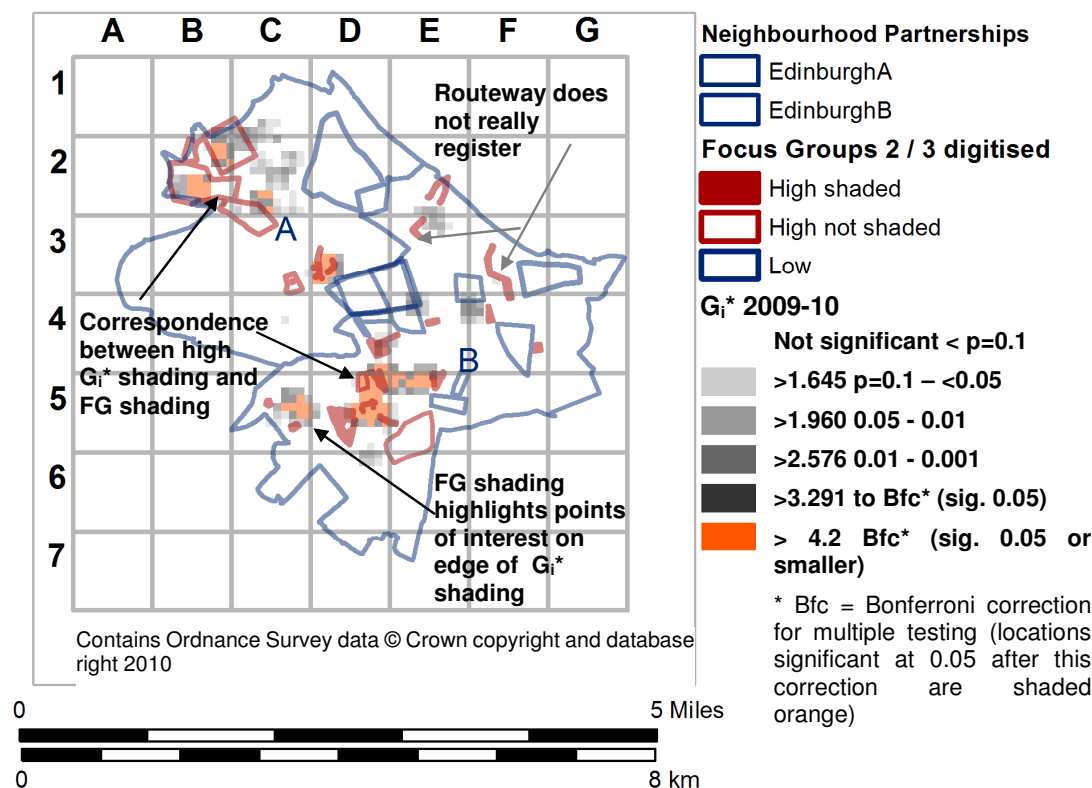
Additionally these techniques were combined in two key ways which added value and allowed triangulation of results:-

- *Combining officer knowledge with LISA maps.*
- *Comparing results of office focus groups, LISA and GTA.*

Combining these techniques created a rich set of results which provided far more insights than could have been achieved by one method alone and enabled rich insights into possible drivers of change across time.

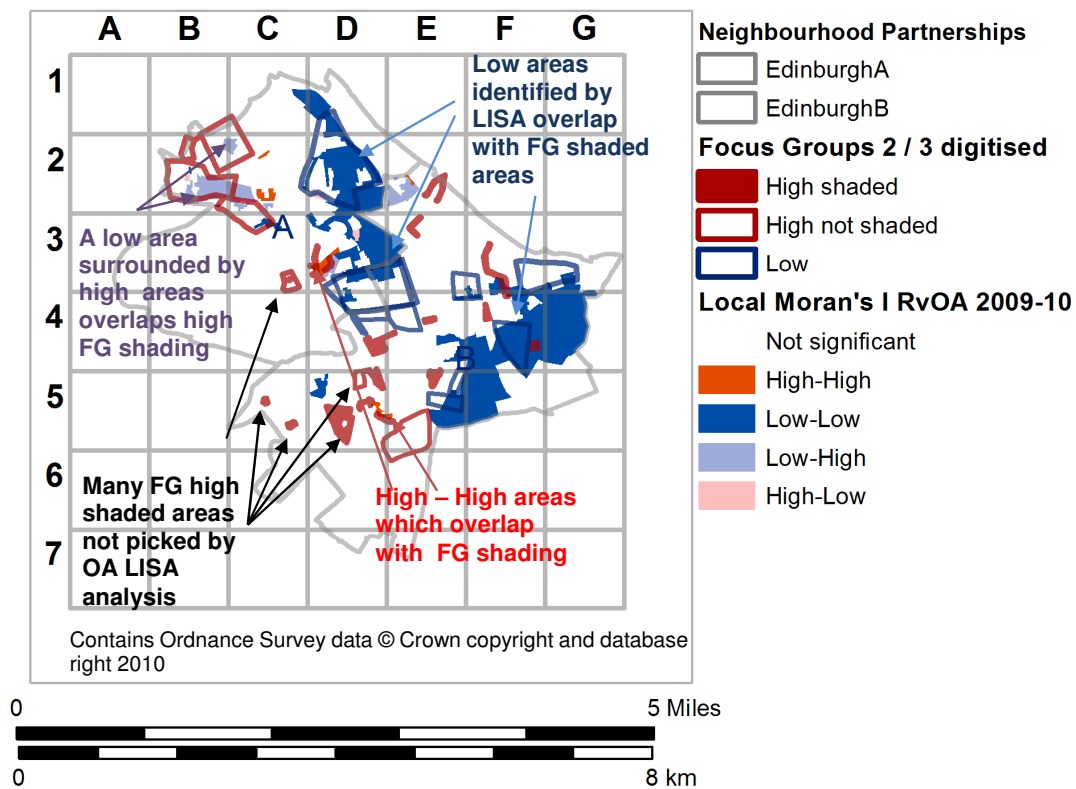
#### 4. Key results

Choice of Scale has a notable impact of processes observed in visualisation. Certain scales appear to show different types of crime concentrations more clearly. This was apparent by comparing high and low concentrations of crime identified by focus groups with LISA analysis. Officers in Focus groups used a combination of fill in shading and outlines for high crime concentration areas (pink), and primarily outlines only for low areas (blue). The types of high area highlighted, varied from single roads and building complexes, to whole estates, low areas were often larger (Figure 1; Figure 2)<sup>‡</sup>. This raises the possibility that different processes may be operating at different scales.



**Figure 1: G<sub>i</sub>\*results for a single year of aggregate crime at 100x100m Grid scale compared to digitised focus group (FG) shaded maps**

<sup>‡</sup> Figure 1 is based on Figure 4.21 (Bates, 2014, p145), Figure 2 is based on Figure 4.22 (Bates, 2014, p146) used with permission.

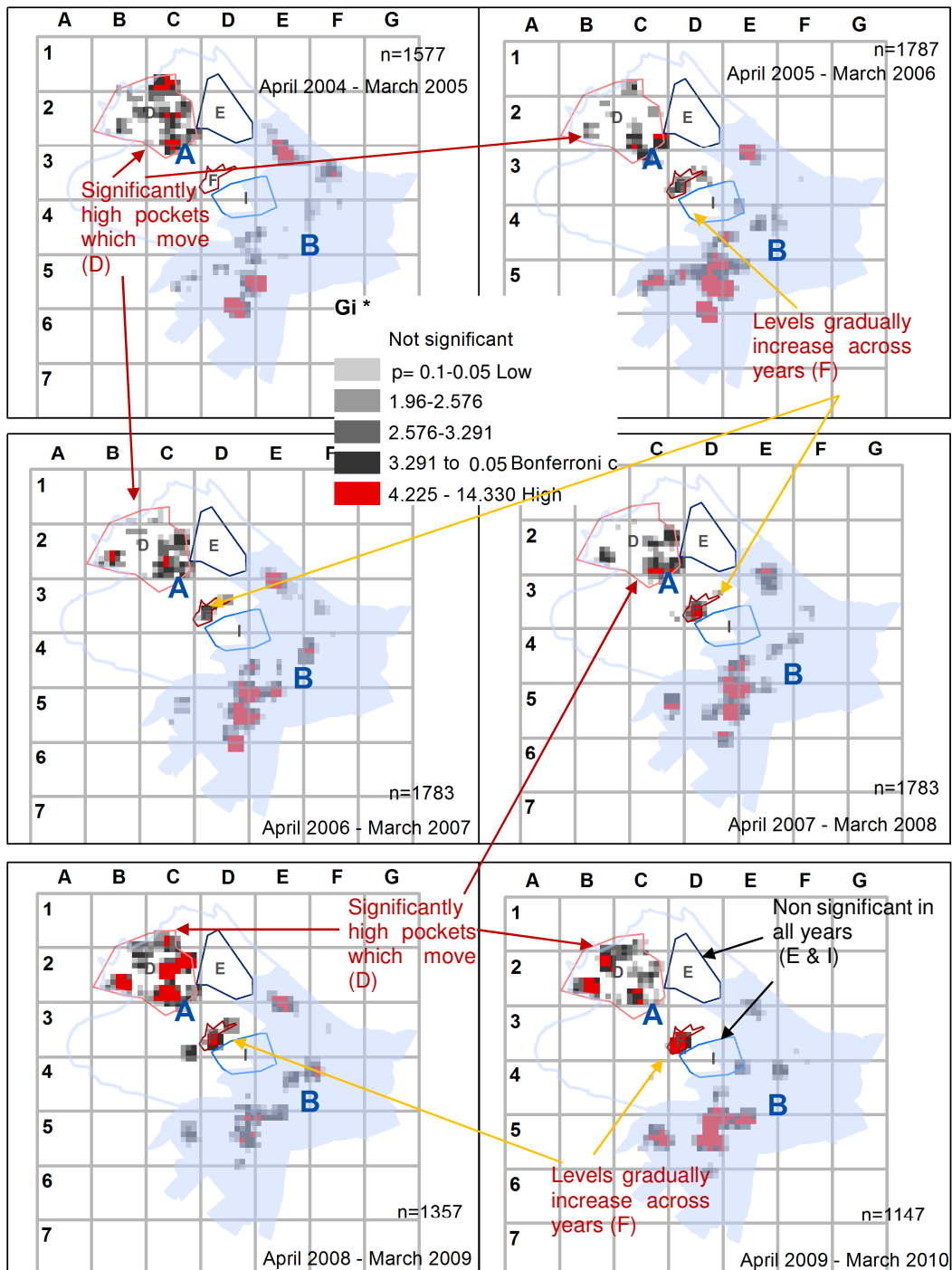


**Figure 2: Local Moran's I analysis for a single year at Output Area (OA) scale compared to digitised focus group (FG) maps**

GTA provided increased descriptive and potential explanatory power, especially when combined with results of spatial analysis (LISA) and information from focus groups. This was important in identifying high, low and 'drifting areas' – pockets that move over time. Figure 3<sup>§</sup> is an example of Gi\* analysis with these identified areas overlaid – D – Delta is Drifting High Area, Indigo I a Drifting Low Area; F – Foxtrot a High Area and E- Echo a consistently low area.

Figure 4 shows complementary of Group Trajectory results with these same areas overlaid. GTA found 4 distinct groups of areas with similar trajectories of changing levels of crime across time were identified at OA level and 6 groups at Grid level (just the 3 high groups are shown here).

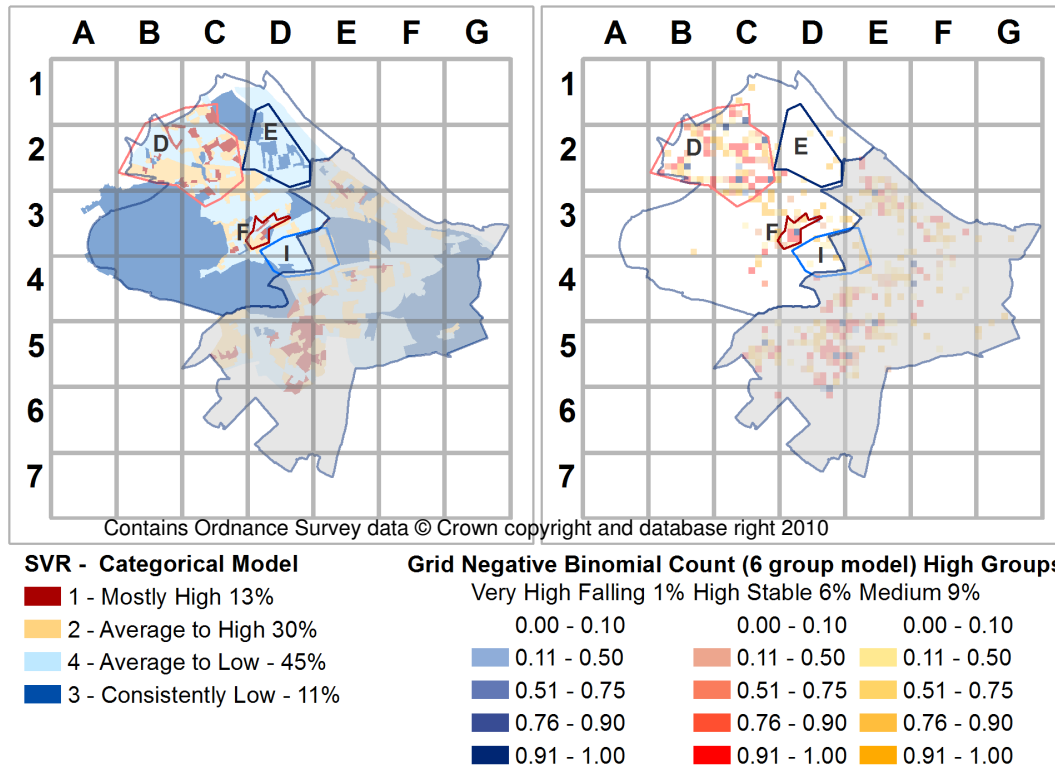
<sup>§</sup> Figure 3 is based on Figure 6.7 (Bates, 2014, p218) used with permission.



### All Vandalisms - Vandalism, Malicious Mischief and Fire-raising

Contains Ordnance Survey data © Crown copyright and database right 2010

**Figure 3: Vandalism in micro-neighbourhoods across the study period in EdinburghA using  $G_i^*$  (row standardised z-scores – Bonferroni correction – significant at 0.05 areas shown in red)**



**Figure 4: Results of Group Trajectory Analysis with areas identified from combined analysis D,E,F and I overlaid \*\***

### 5. Key outstanding challenges

High, Low and Drifting areas had distinct and interesting characteristics; this suggests analysts need to move to using techniques which allow them to identify not only crime hotspots but coldspots and the areas in between.

This analysis suggests we ideally need to analyse data at multiple scales (and clearly state limitations when we use a single scale). More profoundly different results, observed at differing scales, may reflect the impact of differing external processes on high, low and drifting concentrations of crime. Should we model at two or more scales simultaneously? If so, how do we visualise the results of this multi-level analysis?

Modelling this type of data is challenging. No single technique presented here fully takes account of common methodological issues. Combining techniques assists but each still has limitations in one or more areas. More complex modelling strategies may be needed; a further key challenge will be developing the skill base for complex models to be understood and interpreted in a way that is accessible to a lay audience.

\*\* Figure 4 is based on Figure 6.11 (Bates, 2014, 225) used with permission

## Biography

Ellie Bates is an AQMeN research fellow in criminology interested in crime and place. William Mackaness is a senior lecturer specialising in visualisation methodologies and GIS.

## Acknowledgements

Lothian and Borders Police for supply of data; the Scottish Centre for Crime and Justice Research and University of Edinburgh for funding and supporting the PhD this research is based on.

## References

- ANSELIN, L. (1995) Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27, 93-115
- ASSUNÇÃO R. & REIS E., (1999) A New Proposal to Adjust Moran's I for Population Density. *Statistics in Medicine*, 18, 2147-2162
- BATES, E (2014) *Vandalism a Crime of Place*, (PhD Thesis) University of Edinburgh.
- CINDERBY, S. (2009) How to reach the 'hard-to-reach': the development of Participatory Geographic Information Systems (P-GIS) for inclusive urban design in UK cities. *Area*, 1-13
- COHEN, L. E. & FELSON, M. (1979) Social Change And Crime Rate Trends : A Routine Activity Approach . *American Sociological Review*, 44, 588-608
- CRAGLIA, M., HAINING, R. & SIGNORETTA, P. (2005) Modelling high-intensity crime areas: comparing police perceptions with offence/offender data in Sheffield. *Environment and Planning A*, 37, 503-524.
- HAINING, R. & LAW, J. (2007) Combining police perceptions with police records of serious crime areas: a modelling approach. *Journal of the Royal Statistical Society A*, 170, 1019-1034.
- GETIS, A. & ORD, J. K. (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24, 189-206
- GROFF, E. R., WEISBURD, D. & YANG, S.-M. (2010) Is it Important to Examine Crime Trends at a Local "Micro" Level?: A Longitudinal Analysis of Street to Street Variability in Crime Trajectories. *Journal of Quantitative Criminology*, 26, 7-32
- OPENSHAW, S (1984) The Modifiable Areal Unit Problem, *Concepts and techniques in modern geography*; no.38, Norwich
- HAINING, R. (2003) *Spatial Data Analysis Theory and Practice*, Cambridge, Cambridge University Press.
- RATCLIFFE, J. H. & MCCULLAGH, M. J. (2001) Chasing Ghosts? Police Perception of High Crime Areas. *Brit. J. Criminology*, 41, 330-341.
- TAYLOR, R. B. (2010) Communities, Crime, and Reactions to Crime Multilevel Models: Accomplishments and Meta-Challenges. *Journal of Quantitative Criminology*, 26, 455-466.
- WEISBURD, D., GROFF, E. & YANG, S. M. (2012) *The Criminology of Place Street Segments and our Understanding of the Crime Problem*, Oxford, Oxford University Press

# TravelOAC: development of travel geodemographic classifications for England and Wales based on open data

Nick Bearman<sup>\*1</sup> and Alex D. Singleton<sup>†1</sup>

<sup>1</sup>Department of Geography and Planning, School of Environmental Sciences, University of Liverpool, Gordon Stephenson Building, Liverpool, L69 7ZQ, United Kingdom

## Summary

This paper develops a custom geodemographic classification for travel in England and Wales. Travelling is an important factor in many life decisions, including home and work life. Variables for transport (distance to nearest airport, rail station, ferry station, tram stop and bus stop, number of cars owned, and mode of travel to work) and demographics (gender, age and social class) for each Output Area in England and Wales are used to create eight clusters of different transport characteristics. The characteristics of the different clusters are discussed, along with future improvements to be implemented in the classification method.

## KEYWORDS:

Open geodemographics, travel, public transport, Output Area classification, open data

## 1. Literature Review

Geodemographics are "the analysis of people by where they live" (Sleight, 1997) and started in the early 1900s with Charles Booth's poverty map of London, a spatial representation of different social classes (Booth, 1902; LSE, 2014). The academic side of geodemographics developed through the 1950s and 1960s, when the commercial application of geodemographics also developed. These commercial developments in the 1970s were primarily led by Richard Webber (1977), involved in the creation of MOASIC and ACORN, two commercial classifications that are still in use today. This (and many other) commercial classifications are a "black box" process - there was little publicly known about the data sets used or how the classification was constructed.

Recent developments have brought geodemographic classification generation back into the academic sector, with the open geodemographics developing a type of classification that is academically rigorous and with the methods used open to inspection. This was partly driven by the availability of open data, particularly data for small areas related to the census. The OAC (output area classification) was developed for the 2001 and 2011 censuses (Vickers and Rees, 2007) and it is the 2011 OAC and 2011 Census data that this work builds upon.

Both the OAC classification and a number of commercial geodemographic classifications were built with a range of applications in mind. However, the validity of a geodemographic classification for a generic application has been questioned, as the factors influencing someone to pick a specific holiday destination are likely to be different to their opinions on private healthcare (Singleton and Longley, 2009). Additionally, it is likely that for a specific application (e.g. travel) that there is additional open data available which could contribute useful information to the geodemographic classification. From an analytical point of view, it has been found that "differences between [geodemographic] classes are generally smaller than the differences found within any particular class" (Voas and Williamson, 2001). These factors, alongside the fact that while generating a specific geographic used to be a complex undertaking, it is now easier as a result of the maturing of spatial data technologies (Adnan et al., 2010; Singleton and Longley, 2009), create a compelling argument to develop application specific classifications. Below, the area of travel and transport is introduced, and this will be followed

---

<sup>\*</sup> n.bearman@liverpool.ac.uk

<sup>†</sup> alex.singleton@liverpool.ac.uk



by the creation of a travel geodemographic classification.

Everyone needs to travel for a variety of reasons (work, school, shopping, etc.) and the factors behind the choice of a specific mode of travel for a specific journey are varied and complex. A travel geodemographic classification will show how transport provision and usage varies across the country, and highlight any relationship it has with other demographic factors such as gender, age and SES (socio-economic status). It could also be used to target transport improvements that are particularly important to increase uptake of public transport and reduce reliance on private car use. Understanding the geodemographics of travel accessibility and travel use will help development of transport options and contribute to the task of reducing transport CO<sub>2</sub> emissions. Transport accounts for about 28% of the UK's total CO<sub>2</sub> emissions (Hickman and Banister, 2007) so understanding more about this issue can help reduce carbon emissions.

## 2. Methods

The theoretical framework adopted for this work assists with selecting variables by considering the domains that this work is interested in, the concepts within each domain and then the variable within each concept. Table 1 shows how this is applied to the travel geodemographic, and also shows which census variable was used. Gender, age and social class were included in the clustering process to represent a proxy for income (social class) and different transport needs (e.g. working vs. non-working).

**Table 1.** Variables included in the classification

Domains	Concepts	Variable	Census table used
Demography	Gender	Gender	KS101 Usual resident population
	Age	Age groups	KS102EW Age structure
	Social class	National Statistics socio-economic class	KS611EW NS-SeC
Transport	Travel to work	Mode of usual travel to work	QS701EW Method of travel to work
	Ease of access to car	Car ownership	KS404EW Car or van availability
	Ease of access to public transport	Distance to closest bus/train/ferry/airport stop	NA ( <i>distance calculated from NaPTAN data</i> )

The transport stop data was generated using stop locations for England and Wales from NaPTAN<sup>‡</sup>, with the distance to each type of nearest stop calculated for each Output Area population centroid. OA population centroids were used because these more appropriately reflected the location of the population within an OA, and the use of point data allowed a much simpler spatial calculation to take place (measuring distance points to points) than calculating distances from points to polygons (Output Areas).

Walking routes were modelled in Routino for distances from each OA centroid to the nearest rail station, tram stop and bus stop using previously discussed methods (Bearman and Singleton, 2014). Straight-line distances were used for airport and ferry ports because in the vast majority of cases a walking route to an airport or ferry port would not make sense. The transport types chosen were based on the existing categories in NaPTAN, and will have different weightings in the classification according to the number of different stops of that type to reflect their different levels of importance.

A k-means classification was performed using the data described above. Both an initial k-means cluster analysis and a clustergram analysis (Galili, 2010; Schonlau, 2002, 2004) were run to establish the number of clusters, and based on these results a classification was run for eight clusters.

<sup>‡</sup> NaPTAN, National Public Transport Access Nodes, <http://data.gov.uk/dataset/naptan>

### 3. Results & Clusters

The output from the cluster analysis shows interesting characteristics, with each of the eight clusters showing distinct characteristics. A number of parameters within each category showed similar relationships in the initial analysis, so were collapsed together in the final analysis. Table 2 and Figure 1 shows the clusters derived from this analysis, along with their variables and spatial distribution. A number of the clusters (e.g. 3 and 5) have strong patterns within the variables, whereas others have a weaker relationship.

**Table 2.** Clusters and data patterns for the transport geodemographic classification

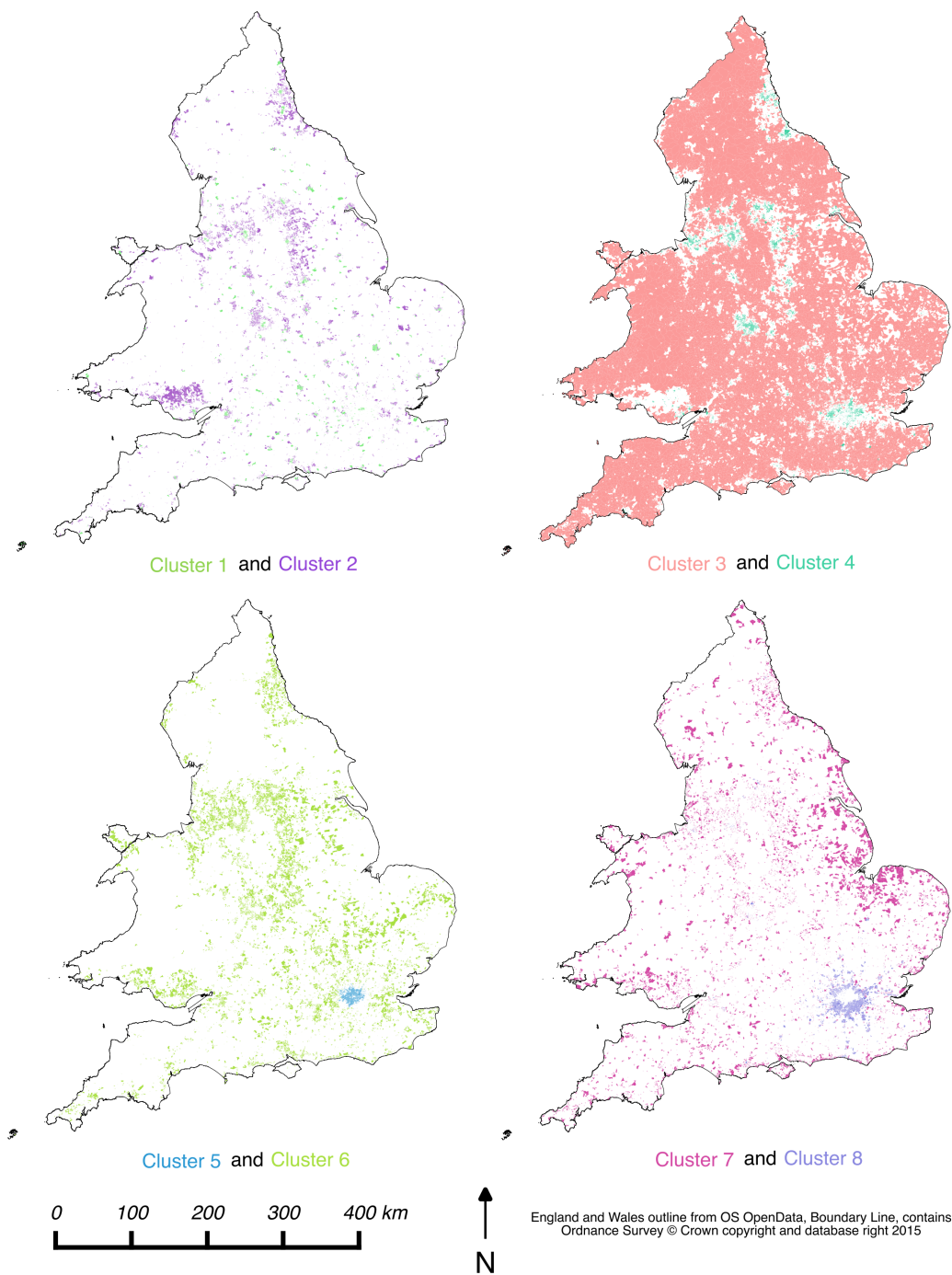
Cluster	1	2	3	4	5	6	7	8
N (total N = 181,408)	8,755	38,634	24,076	21,161	9,480	42,445	23,541	13,316
Distance to public transport	Mid	Far	Far	Close	Close	Mid	Far	Far
Car ownership	0 car	1 car	2+ cars	0 car	0 car	2+ cars	1 car	1 car
Gender	Male	Mix	Mix	Mix	Slightly male	Mix	Female	Slightly female
Age	15-44	Under 45	45+	Under 45	15-44	Mix	45+	Slightly <45
National Statistics Socio-economic classification <sup>#</sup>	1							
	2							
	3							
	4							
	5							
	6							
	7							
	8							
Usual mode of travel to work	Cycle / Walk	Passenger	Home work / drive	Bus & taxi	Tram /bus	Drive	Slight car	Train

<sup>#</sup> 1. Higher managerial, administrative and professional occupations 2. Lower managerial, administrative and professional occupations, 3. Intermediate occupations, 4. Small employers and own account workers 5. Lower supervisory and technical occupations, 6. Semi-routine occupations, 7. Routine occupations, 8. Never worked and long-term unemployed

Table 2 summarises the analysis of the results, showing the different characteristics of the different clusters from this analysis. The full table (including variable index scores) is available in the appendix. Clusters 1 and 5 are professional workers, more likely to be male and in management. Cluster 5 is predominantly located in London, with a higher use of public transport, whereas cluster 1 is located in other urban areas, with a higher rate of cycling and walking. Clusters 2 and 3 are located in rural areas, with cluster 3 in higher management with a higher income and more likely to have 2 or more cars, and cluster 2 with lower incomes, with 1 car, in lower supervisory, technical and routine occupations. Cluster 4 is located in urban areas, with an emphasis on the more deprived areas in the north of England and London, with a combination of low wages and a reliance on bus and taxis. Cluster 6 is primarily distributed in suburban and rural areas, with a mixed demographic, higher income and multiple cars. People within clusters 7 and 8 are more likely to be female and to have a lower income. Cluster 8 has a London commuter focus and much exhibits higher train usage, whereas cluster 7 covers rural areas in the rest of England with a reliance on a car.

### 4. Discussion

These results show the potential benefits of generating a travel geodemographic classification. There are both travel patterns associated with age and income (SES), as well as additional ones associated with access to and use of different modes of transport. In addition, there are particular geographic patterns, both for rural and urban locations, which also reflect income and age distributions. Gender is a strong factor in two clusters (cluster 1 and 7, and to a lesser extent cluster 5) and is likely to reflect income and social make up differences. Socio-economic classification also features strongly in the classification, and we use this data as a proxy for income levels.



**Figure 1.** Spatial variation of a selection of the clusters from the travel geodemographic classification.

This data indicates particular clusters which might benefit from targeting to reduce CO<sub>2</sub> emissions, either by promoting more public transport use, or acting on reasons for low levels of public transport usage. In addition it highlights areas that have low public transport provision as well as low public transport usage levels, which could assist in targeting new public transport provision.

There are a number of refinements to be made to the classification, both in the data included and the processing of the classification. Currently, distances to most forms of public transport are calculated as walking distances. However this will vary between individuals, and is also not appropriate for some types of multimodal transport (for example, when someone drives to a train station to catch a train). Additionally, this work does not take into account the frequency of public transport or the routes followed, which are important factors when considering transport accessibility. In addition, data from the National Travel Survey could provide more contextual information for the analysis of the cluster behaviour patterns.

## **5. Acknowledgements**

The authors of this paper would like to acknowledge support from the ESRC for this research, as well as acknowledging the open data sets used in this work, including 2011 Output Area Classification, OS Open Data Boundary Line, 2011 Census data and National Public Transport Access Nodes. The authors would also like to thank the reviewers for their useful comments.

## **6. Biography**

Nick Bearman is a Research Associate and University Teacher at the University of Liverpool, previously working at the University of Exeter Medical School. He is interested in the use of GIS to solve novel problems, particularly in the areas of secondary data reuse, health and big data.

Alex Singleton is a Professor in Geographic Information Science; his research concerns how the social and spatial complexities of individual behaviour can be represented and understood within a framework of quantitative social science and computer modelling, extending from a geographic tradition of area classification.

## **References**

- Adnan, M., Longley, P.A., Singleton, A.D., and Brunson, C. (2010). Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases. *Trans. GIS 14*, 283–297.
- Bearman, N., and Singleton, A.D. (2014). Modelling the school commute for 7.5m students over 4 years using data from the school census. In GISRU2014, (University of Glasgow, UK),.
- Booth, C. (1902). *Life and Labour of the People in London* (London: Macmillan).
- Galili, A.T. (2010). Clustergram: visualization and diagnostics for cluster analysis (R code).
- Hickman, R., and Banister, D. (2007). Looking over the horizon: Transport and reduced CO<sub>2</sub> emissions in the UK by 2030. *Transp. Policy 14*, 377–387.
- LSE (2014). Poverty maps of London (Charles Booth Online Archive).
- Schonlau, M. (2002). The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *Stata J. 2*, 391–402.
- Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Comput. Stat. 19*, 95–111.
- Singleton, A.D., and Longley, P.A. (2009). Geodemographics, visualisation, and social networks in applied geography. *Appl. Geogr. 29*, 289–298.
- Sleight, P. (1997). *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business* (Henley-on-Thames, Oxfordshire England: NTC Publications).

Vickers, D., and Rees, P. (2007). Creating the UK National Statistics 2001 output area classification. *J. R. Stat. Soc. - A* 170, 379–403.

Voas, D., and Williamson, P. (2001). The diversity of diversity: a critique of geodemographic classification. *Area* 33, 63–76.

Webber, R.J. (1977). *An Introduction to the national classification of wards and parishes* (London: Centre for Environmental Studies).

### Appendix: Index Scores of Cluster Variables

**Table 1.** Clusters and variable index scores (0 = typical) for the transport geodemographic classification

Cluster		1	2	3	4	5	6	7	8
N (total N = 181,408)		8,755	38,634	24,076	21,161	9,480	42,445	23,541	13,316
Distance <sup>#</sup>	Rail	-0.35	-0.11	<b>0.58</b>	-0.20	-0.17	-0.05	0.26	-0.37
	Tram	0.04	0.24	0.08	-0.29	<u>-0.74</u>	0.04	0.28	<u>-0.50</u>
	Bus	-0.36	-0.10	0.38	-0.28	-0.20	-0.14	<b>0.53</b>	-0.05
	Airport	-0.32	0.21	0.44	-0.47	<u>-0.63</u>	-0.06	0.46	<u>-0.61</u>
	Ferry	0.05	0.17	0.07	0.10	<u>-1.02</u>	0.25	-0.09	<u>-0.72</u>
Car / van ownership	None	<b>0.75</b>	0.30	<u>-1.00</u>	<b>1.41</b>	<b>1.44</b>	<u>-0.74</u>	-0.14	-0.20
	1	0.07	0.44	<u>-0.96</u>	<u>-0.57</u>	-0.46	0.07	0.47	<b>0.61</b>
	2 or more	<u>-0.78</u>	<u>-0.51</u>	<b>1.46</b>	<u>-1.13</u>	<u>-1.22</u>	<b>0.71</b>	-0.08	-0.10
Age	0 - 4	-0.48	0.34	<u>-0.62</u>	<b>0.99</b>	0.18	-0.01	<u>-0.88</u>	0.34
	5 - 14	<u>-1.31</u>	0.22	0.06	<b>0.80</b>	<u>-0.61</u>	0.20	<u>-0.82</u>	0.07
	15 - 44	<b>2.14</b>	0.15	<u>-0.87</u>	0.42	<b>1.42</b>	-0.09	<u>-1.06</u>	0.24
	45 - 64	<u>-1.45</u>	-0.22	<b>1.08</b>	<u>-0.77</u>	<u>-1.01</u>	0.34	0.39	-0.19
	65+	<u>-0.90</u>	-0.21	0.46	<u>-0.58</u>	<u>-0.81</u>	-0.20	<b>1.59</b>	-0.28
National Statistics Socio-economic classification	SES1&2 Higher / middle management	0.17	<u>-0.81</u>	<b>0.89</b>	<u>-1.18</u>	<b>1.05</b>	0.35	-0.07	<b>0.77</b>
	SES3 Intermediate management	<u>-0.83</u>	-0.26	-0.12	<u>-0.80</u>	<u>-0.77</u>	<b>0.74</b>	0.28	0.49
	SES4 Self-employed	<u>-0.86</u>	-0.44	<b>1.39</b>	<u>-0.62</u>	-0.14	-0.08	0.31	0.12
	SES5 Lower supervisory	<u>-0.76</u>	<b>0.76</b>	<u>-0.72</u>	-0.06	<u>-1.01</u>	0.14	0.40	<u>-0.71</u>
	SES6&7 Semi-routine / routine	<u>-0.82</u>	<b>1.05</b>	<u>-0.91</u>	<b>0.83</b>	<u>-1.03</u>	-0.30	0.21	<u>-0.84</u>
	SES8 Never worked	-0.22	0.21	<u>-0.65</u>	<b>1.98</b>	0.31	<u>-0.55</u>	-0.40	-0.20
Usual mode of travel to work	Car (Driver)	<u>-1.00</u>	-0.01	<b>0.60</b>	<u>-1.16</u>	<u>-1.92</u>	<b>0.97</b>	0.11	-0.48
	Car (Passenger)	-0.46	<b>0.87</b>	-0.44	-0.15	<u>-1.37</u>	0.24	-0.11	<u>-0.79</u>
	Tram	-0.22	-0.31	-0.27	0.07	<b>3.44</b>	-0.29	-0.31	0.44
	Train	0.04	-0.38	-0.09	-0.19	<b>0.62</b>	-0.19	-0.37	<b>2.34</b>
	Bus	0.37	0.09	<u>-0.82</u>	<b>1.23</b>	<b>1.20</b>	-0.31	<u>-0.51</u>	0.07
	Taxi	-0.03	0.32	-0.41	0.45	0.11	-0.20	-0.21	0.02
	Motorcycle	-0.33	0.09	-0.22	-0.35	<b>0.78</b>	0.08	-0.12	0.31
	Cycle	<b>1.14</b>	0.02	-0.40	-0.23	<b>1.21</b>	-0.09	-0.13	-0.08
	Walk	<b>2.11</b>	0.33	<u>-0.50</u>	0.01	0.03	-0.36	-0.05	-0.22
	Other	0.15	-0.10	0.26	-0.06	0.29	-0.12	-0.01	0.01

**Bold** = more than or equal to **+0.5**, underline = less than or equal to **-0.5**.

<sup>#</sup> For distance, positive values are higher distances than average, and negative values are closer than average.

## **Geodemographics and Big Data – A New Research Agenda**

*Mark Birkin, School of Geography, University of Leeds*

[m.h.birkin@leeds.ac.uk](mailto:m.h.birkin@leeds.ac.uk)

**Acknowledgements:** This work was funded as part of the Consumer Data Research Centre, an ESRC Big Data investment. I am grateful to CallCredit Information Group for permission to use data from its Cameo geodemographic classifications.

### **Introduction**

Geodemographics has been of interest to geographers, market analysis professionals and service planners for several decades. Intellectually, the technique has its roots in the factorial ecology of North American cities (see for example Shevky and Bell, 1955; or Rees, 1968). In simple terms, the ‘factorial’ part involves the use of small area data – usually from a population census – to provide numbers which can be crunched through some statistical technique such as a factor analysis; and the ecology bit means that the outcome is usually a classification which characterises each of these areas as a demographic ‘ecosystem’ through which its similarity or difference to other areas can be assessed. Geodemographics is a nice technique for teaching and for elementary spatial analysis because it is quite intuitive and generates reliable outputs which are easy to understand, but there hasn’t been a lot of research innovation in recent years. It has been popular amongst both marketing professionals and (public) service providers as a straightforward and effective method for segmenting users. Although there is now an open source classification which is available for academic and public use (the output area classification – see Vickers and Rees (2007); Gale et al (2015) there are also a variety of proprietary classifications, and a number of these have adopted newer methods from self-organisation and machine learning e.g. genetic algorithms (e.g. in systems such as Mosaic); and more important perhaps most have incorporated additional data sources which can add extra spatial and socio-demographic refinement, from sources such as shareholder registers, house prices, the electoral roll, and county court judgements (as in CallCredit’s Cameo range, for example).

### **Recent developments**

A question which then arises is whether even more varied and diverse sources of (big) data now sound the death knell for geodemographics, or provides the opportunity for innovative and original research which has been mostly lacking in recent times. We assert that the case for reinvigoration is a strong one, primarily because the rationale for geodemographics has never been predicated on the pure convenience of spatial data. Rather, the case is that there are spatial processes which operate at a neighbourhood scale and render patterns which are distinctive for defined geographical areas which are somehow more than the sum of their component parts (i.e. the individual people and households which they contain). In this context the capacity to use data about individuals in order to construct micro-scale profiles is not necessarily an advance in itself – although some combination of individual and neighbourhood profiling may be an ideal modelling strategy in many circumstances (e.g. Birkin and Clarke, 2011). It could be therefore that a major appeal of ‘big data’ is simply that it promises a much broader and diverse basket of indicators from which appropriate classifications may be constructed.

It will be taken as a given for the present purpose that the advent of big data will impact academic life in some rather significant form. The process by which this might come about is not a material concern. We will content ourselves with two brief observations, the first that current developments such as the instantiation of a Big Data Network of administrative, business, local government and social media data by the Economic and Social Research Council, extended investment in Understanding Society, and the inevitability of fundamental revisions to the shape and content of census statistics, are changing the research landscape in a rapid and dramatic manner. The second is that a strong tendency towards open data also now seems ingrained and will promote easy availability of an extensive range of data, but it is neither likely or necessarily desirable that everything will be made available in this form.

As an illustration of this latter point, some of the ideas which will be tested later will exploit the availability of a fine scale proprietary geodemographic classification for research purposes. The Call Credit Information Group have very generously agreed access to their Cameo classifications at postcode level ; clearly unrestricted access to these data would be less than appealing as a commercial strategy.

Geodemographics has not kept pace with the latest developments. The paper suggests that a qualitative shift has taken place which demands a fundamental reappraisal of methods and perspectives.

### **Opportunities**

A number of possibilities could be evaluated in more detail.

- 1 Geodemographics as a filter for big data
  - Is it possible to eliminate the noise in big data by using data compression methods which are directly analogous to the segmentation and clustering techniques used in geodemographics?
- 2 Is there an optimal scale for geodemographics?
  - Is it possible to devise experiments which explore the explanatory and predictive power of alternative classifications at a variety of spatial scales? And does this say anything about the nature and cohesion of small areas and neighbourhoods? An example of this type is sketched, together with some preliminary results, in Figure 1.
- 3 Can interactions be represented and is this helpful?
  - As more data becomes available, it should be possible to construct new kinds of inputs to the geodemographic process e.g. including those which represent interactions between areas alongside the more usual area-based attributes and characteristics. An example of this type is sketched, together with some preliminary results, in Figure 2.
- 4 Is it possible to aggregate geodemographic groups and how can this be done?
  - The use of geodemographics across scales is a recurring problem – for example, how is it possible to summarise a school catchment area or retail market from the composition of a number of different small area profiles in that area?
- 5 Does new data change the case for variant classifications?
  - Commercial providers have experimented with specialist classifications e.g. for retailing and financial services. Could this idea be extended using the wide variety of data sets that are now available for individual sectors or activity types?

Figure 1. Improvements from a geodemographic model across scales

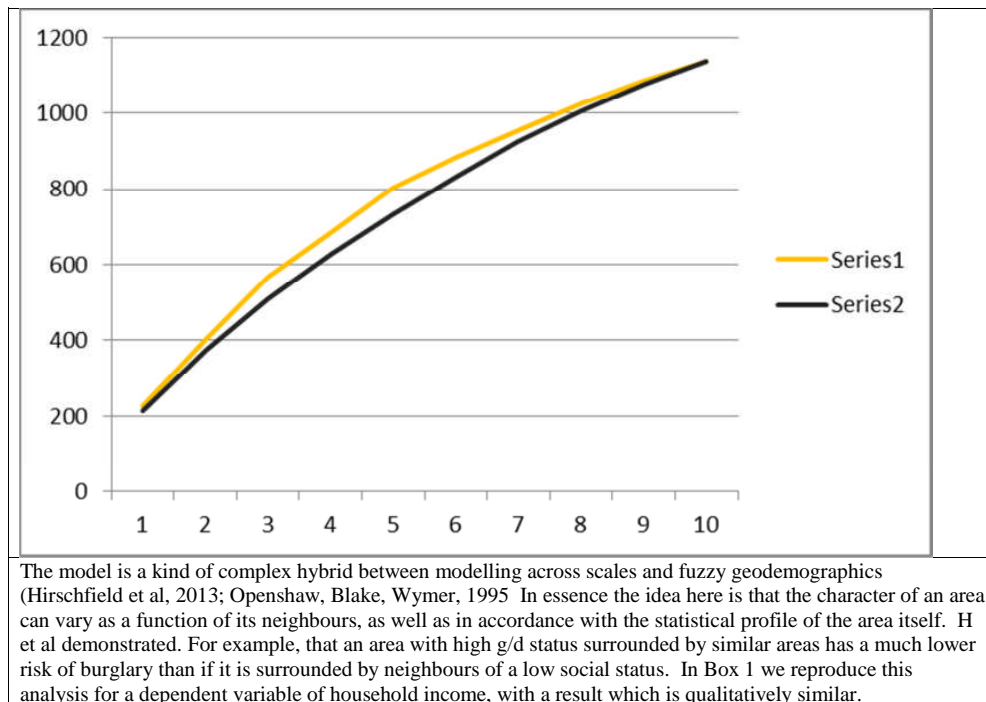
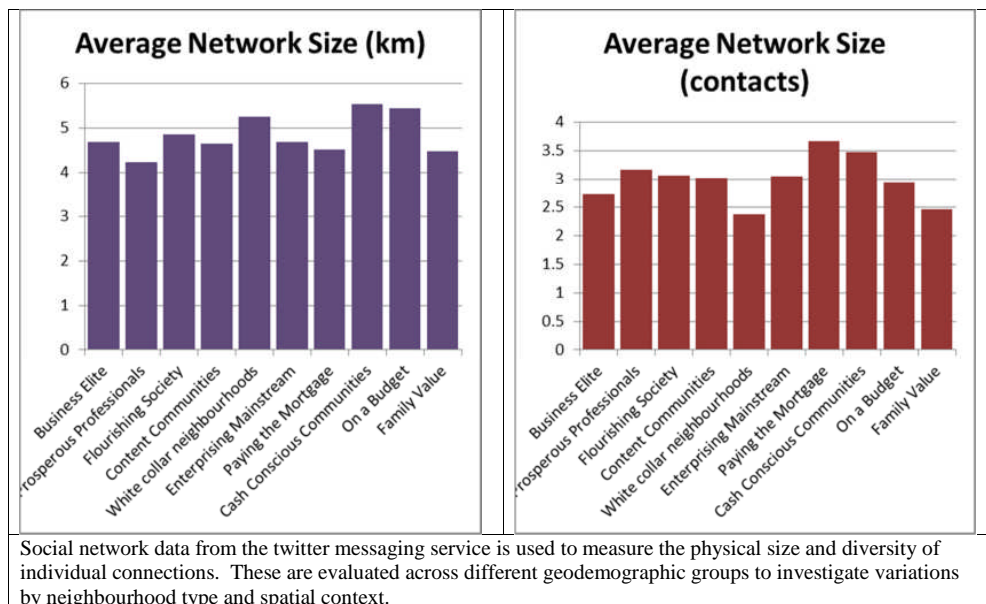


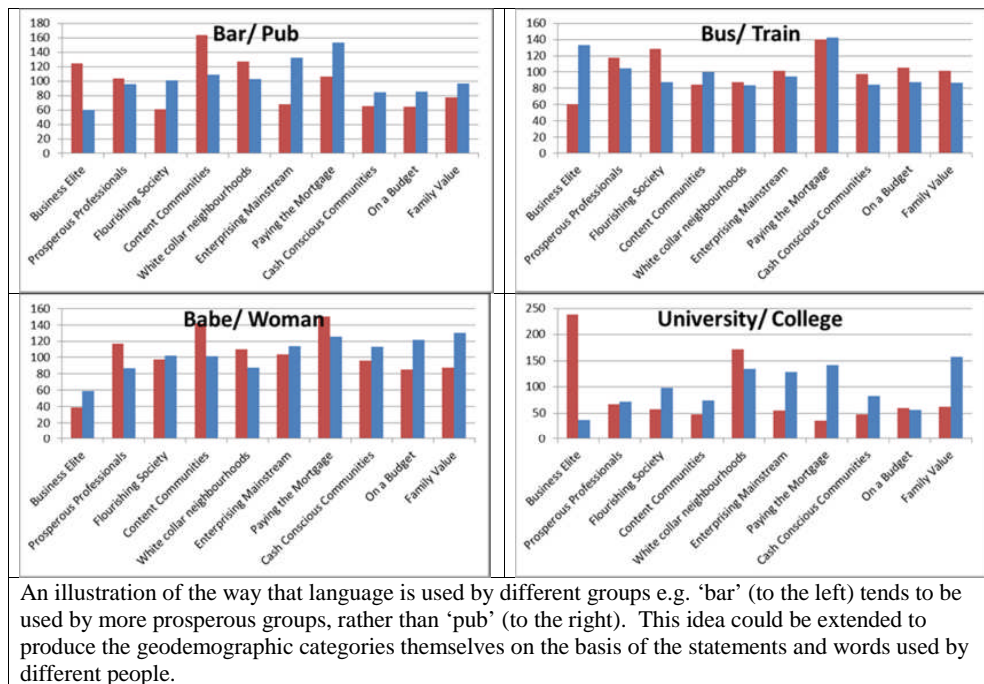
Figure 2. Geodemographic interactions





- 6 Are there certain data sources which add more value than others – should we be more strategic in our approach to their selection?
  - The problem here might be optimise across a wide range of different input alternatives to provide a better performing classification.
- 7 Are there effective ways to exploit the availability of rolling data i.e. updating?
  - A significant weakness of the output area classifications is that they are tied to the decennial census. Commercial providers have been using various data sets and methods to update their classifications for some time. These approaches could be refined further and translated into academic practice.
- 8 Geodemographics in space and time
  - The ultimate extension of the updating concept could be towards something like real-time geodemographic system which reflect diurnal variations in the character of areas (e.g. the city centre as a workplace by day and a playground by night)
- 9 New forms of geodemographics: from you are where you live to you are what you do
  - This could again involve data which are in common use commercially (e.g. lifestyles) but rarely taken into academic use; or newer sources altogether such as social media and tweets. An example of this type is sketched, together with some preliminary results, in Figure 3.

Figure 3. Language as an indicator of activity



## **Discussion**

Our nine questions lead naturally towards a research agenda in geodemographics. A number of next steps are possible – for example to continue to sketch case studies rather lightly (e.g. as figures 1 to 3), to develop more rigorous implementations of one or more of these options; perhaps to eliminate or consolidate these options, or open completely new possibilities.

It could be that the direction of travel is to some extent data driven – for example, real-time data would be needed to support meaningful work on real-time classifications. Priorities should also be determined with due reference to the opinions of both data users and data suppliers, and these two groups may well be overlapping – for example, the data providers have a strong interest in added value outputs which might eventually result from more speculative research. We hope that the active pursuit of such conversations could lead to productive and mutually beneficial interactions e.g. between research groups and business organisations.

## **References**

- Birkin M, Clarke G (2011) The enhancement of spatial microsimulation models using geodemographics, *Annals of Regional Science*, 49(2), 515-532.
- Gale, C., Singleton, A., Bates, A., Longley, P. (2015) Creating the 2011 Area Classification for Output Areas (2011 OAC), JOSIS Discussion Forum, online at <http://josis.net/index.php/josis/article/viewArticle/232>, accessed 5<sup>th</sup> June 2015.
- Hirschfield, A., Birkin, M., Brunson, C., Malleon, N., and Newton, A (2013) How places influence crime: The impact of surrounding areas on neighbourhood burglary rates in a British City. *Urban Studies*, 49(8), 1-16. p22, 39pp.
- Openshaw, S., Blake, M., & Wymer, C. (1995). Using neurocomputing methods to classify Britain's residential areas. *Innovations in GIS*, 2, 97-111.
- Rees, P. (1968) The factorial ecology of metropolitan Chicago, University of Chicago.
- Shevky, E., Bell, W. (1955) *Social Area Analysis*, Stanford, California.
- Vickers, D., Rees, P. (2007) Creating the UK National Statistics 2001 output area classification, *Journal of the Royal Statistical Society, Series A*, 170, 379-403.

# Integrating BIM and GIS :

## Exploring the use of IFC space objects and boundaries

Gareth Boyes<sup>\*1</sup>, Charles Thomson<sup>†1</sup> and Claire Ellul<sup>‡1</sup>

<sup>1</sup>Department of Civil, Environmental and Geomatic Engineering, University College London

January 9, 2015

### Summary

In GIS, understanding the layout of interior spaces has important applications in the analysis of energy efficiency, indoor navigation and atmospheric pollution. However, detailed models of building internals are usually held in engineering or architecture software, including Building Information Modelling (BIM) software such as Autodesk Revit, and are not easily understood by GIS in this format. Additionally, such models contain excessive detail, such as wall thickness, not required for GIS operations such as topological adjacency. This paper describes the process required to convert BIM data into GIS, addressing both conceptual modelling and data format differences.

**KEYWORDS:** Energy & Sustainability – Building Information Modelling – Space Boundaries – GIS and BIM Integration – 3D Topology

### 1. Introduction

There are two complementary geospatial systems used in the Architecture, Engineering, Construction, Owner, Operator (AECOO) community. Readers will already be aware of Geographic Information Systems (GIS) for the collation, storage, analysis and management of information. The AECOO community has also adopted Building Information Modelling (BIM) as a coordinated set of processes for the design, management and sharing of building and infrastructure information (Mott MacDonald, 2014). At the heart of the BIM process is an object-based model that is ascribed throughout with geospatial information.

The UK Government report *Construction 2025* aims to reduce greenhouse gas emissions by 50% in

---

<sup>\*</sup> gareth.boyes.13@ucl.ac.uk

<sup>†</sup> charles.thomson.11@ucl.ac.uk

<sup>‡</sup> c.ellul@ucl.ac.uk

the built environment (HM Government, 2013). Implementation of BIM as standard industrial practice is key to achieving this strategic aim.

Environmental data such as ambient temperature, air and noise pollution are commonly available as GIS datasets. GIS enhances the BIM process with established toolsets for overlaying and querying data. This overlay and query of multiple-source data is a fundamental feature of GIS that permits architects and engineers to make efficient decisions (Cowen, 1988). However, fundamental differences between BIM and GIS exist that preclude successful integration.

The study of spaces is key to understanding the topological layout of buildings. The term *intramural space* is adopted in this paper to refer to the volume of a room contained within the walls, columns, floors and ceilings of the spaces. However, these building elements can overcomplicate the layout in certain situations, e.g. the determination of topological adjacency (Ellul, 2013). The term *contiguous space* is therefore used herewithin to refer to the space bound by wall centrelines that lies between the floor surface and the surface of the floor above. This paper explores the potential of using contiguous space objects and boundaries as a means of simplifying detailed BIM models so that they can then be used in conjunction with GIS methods.

## **2. Literature Review**

### **2.1. BIM**

The concept of 3D object-based building systems, recognisable as a Building Information Model in current parlance, has existed for 40 years (Eastman, 1975). Originally limited by computing power, the concept has since transitioned into reality. There is not one all-defining definition of BIM but great care must be taken to differentiate the BIM process and the product that is the Building Information Model (herewithin "BIM model"). The distinctive characteristics of a BIM model are listed in Table 1 (Isikdag and Zlatanova, 2009 and Lee, Sacks and Eastman, 2006).

**Table 1** Distinctive characteristics of a BIM model

Object-based	Building elements and parametric relationships are object-orientated
Data-rich	All physical and functional characteristics are represented
Three-dimensional	3D as opposed to 2.5D
Spatially-related	Topology of building elements is maintained hierarchically
Semantically-rich	Semantic classification permits objects to inherit properties and behaviours
Generative views	2D and 3D visualisation with annotations
Parametric	Elements are definable from the properties of other objects

Industry Foundation Classes (IFCs) (IAI, 2000) provide the AECOO community with an interoperable format for exchanging BIM data between different software platforms (Laakso & Kiviniemi, 2012). More specifically, IFC is suitable for exporting BIM data to GIS and energy simulation software.

## **2.2. Three dimensional GIS**

GIS has traditionally been constrained to two dimensions. Limited 3D can be implemented using a method known as 2.5D in which height is stored not as a dimension but as an attribute. Advances in computing now permit 3D GIS, opening up analytical methods that were previously impracticable. 3D GIS has many applications in the built environment that include energy performance simulation (Bazjanac, 2008), indoor navigation (Worboys, 2011), urban planning (Stoter et al., 2011), cadastral registration (Stoter et al., 2011), noise propagation (Gröger et al., 2012) and renewable energy modelling (Resch et al., 2014).

Two data structures for storing 3D GIS data are Oracle Spatial and CityCML. Oracle Spatial provides schema and functions for storing and querying spatial features in Oracle relational databases (RDBMS) (Murray, 2013). This SDO\_GEOMETRY format exceeds the functionality of the OpenGIS Implementation Schema (Herring, 2010). CityGML has been developed as the OGC standard information model for describing 3D urban objects as an application schema of the Geographic Mark-up Language (Gröger et al., 2012). Although the model is formatted in XML, it is geometrically compatible with the OpenGIS and Oracle Spatial implementation schema for RDBMSs. CityGML provides the framework for semantic-geometric relationships and has the functionality of representing differing levels of detail (LoDs).

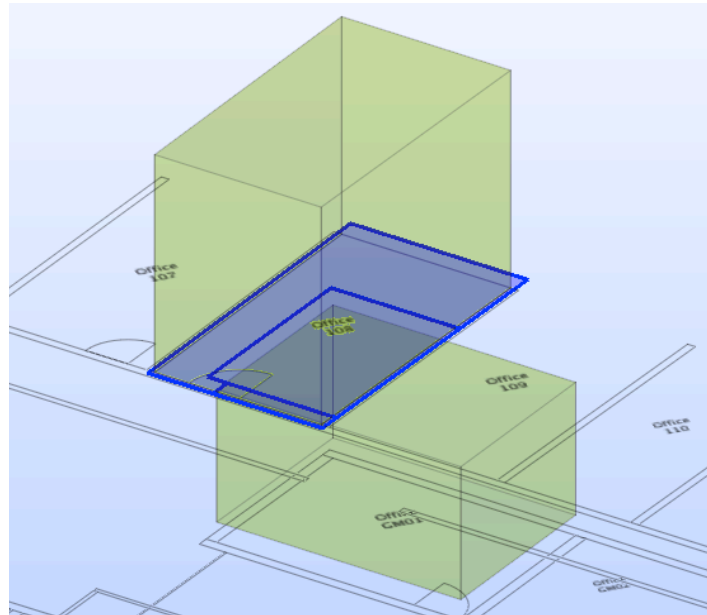
### 2.3. Integration of BIM and GIS

Research continues into GIS and BIM integration. Isikdag and Zlatanova (2009) concluded that manipulating data from one system to the other requires a two-part transformation of both the geometric and semantic datasets. Because the two systems are conceptually misaligned, one dataset cannot be transformed without the transformation of the other. Research has focussed on the unidirectional conversion from IFC to CityGML (Isikdag and Zlatanova, 2009) and this functionality is implemented in software such as FME Workbench (Safe Software, 2012). Unidirectional conversion is inherently wasteful and solutions have been proposed to develop a CityGML Application Domain Extension (van Berlo and de Laat, 2011) and an intermediary Unified Building Model (El-Mekawy, Östman and Hijazi, 2012). Ultimately, any meaningful attempt to integrate BIM and GIS requires a systematic mapping of conflicting semantic data structures (Bittner, Donnelly and Winter, 2006).

### 2.4. Space Objects and Space Boundaries

The geometry of spaces is described as two different entities in IFC, firstly as a space object (*IfcSpace*) and secondly as a space boundary (*IfcRelSpaceBoundary*). *IfcSpace* is geometrically defined as a closed solid. Depending on the complexity of the shape, the space will be represented using profile sweep or boundary representation. The space object contains attributes related to the room such as name, volume and net floor area (buildingSMART, 2014a).

*IfcRelSpaceBoundary* is geometrically described as a collection of planar polygon faces. Each face relates to a particular building element that is stored as a boundary attribute (buildingSMART, 2014b). Revit provides the option to export IFC space boundaries as either 1st or 2nd level space boundaries (Autodesk, 2014). 1st level space boundaries are simple polygon faces representing each building element not including openings for windows or doors (Weise et al., 2011). 2nd level space boundaries describe not just building element geometry but also the geometry of any object on the other side (Weise et al., 2011). Figure 1 shows a screenshot from Solibri Model Checker showing one *ifcSpace* object above another with the adjoining 2nd level space boundaries.



**Figure 1** Space objects and 2nd level space boundaries in Solibri Model Checker

### 3. Data

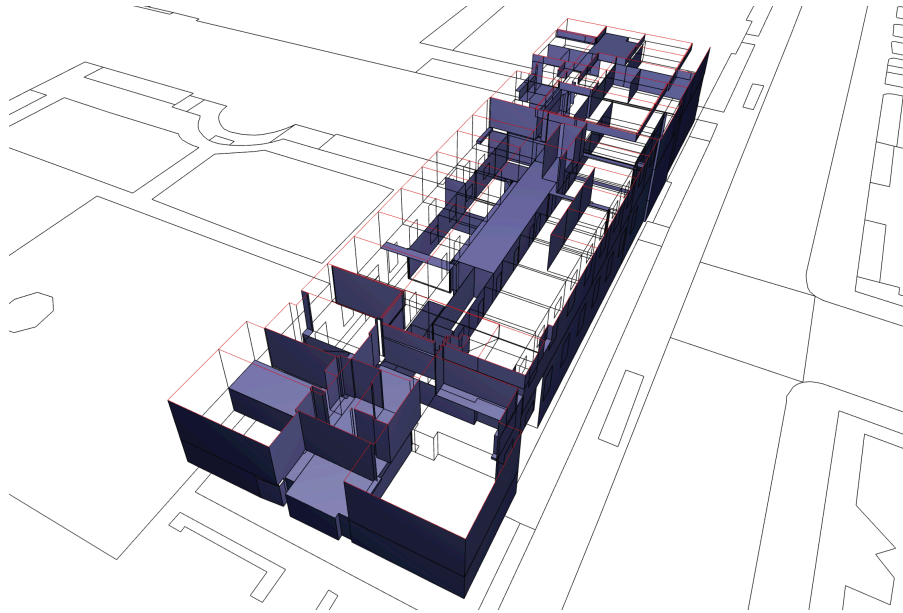
This investigation tested the methods developed on a BIM model of a building on the University College London campus. This model had been developed in Autodesk Revit from point cloud data captured using laser-scanning equipment (Backes et al., 2014). The model uses a coordinate system aligned to the British National Grid with an applied offset.

### 4. Method and Results

Basic building models were created in Autodesk Revit 2015 and exported to FZK Viewer and Solibri Model Checker to test the functionality of exporting space objects and boundaries as IFC files from Revit. The suitability of FME Workbench for transforming the geometry of space objects and boundaries into Oracle Spatial SDO\_GEOMETRY objects was also investigated and found to be incapable of interpreting *IfcRelSpaceBoundary* geometry. Python script was developed from code published on a web-based repository (Mvaerle, 2010) to interpret the *IfcRelSpaceBoundary* geometry as *SDO\_GEOMETRY* objects. IFC class bindings and suitable geometry classes do not exist in Python and script was written to overcome this hurdle.

Building models were exported from Revit with various combinations of the *At Wall Finish / At Wall Centre* and the 1st/2nd Level Boundary settings. To create 3D spaces bounded by room ceilings, it is necessary to change the *Volume Computations* setting in the *Room & Area* drop-down menu to *Areas*

and Volumes. Figure 2 shows contiguous space boundary faces exported from Revit into Oracle Spatial and then visualised as a 3D PDF in Adobe Acrobat superimposed on an OS Mastermap of the surrounding area.



**Figure 2** Space boundary faces of Chadwick Building

## 5. Discussion

The most efficient way to convert intramural spaces is to use FME Workbench to break up the faces of *IfcSpace* objects using a *GeometryCoercer*. Conversely, contiguous space boundaries are best exported from Revit as 2nd Level *IfcRelSpaceBoundary* objects (having selected the *At Wall Centre* setting) and then use the Python script described in the previous section to interpret the geometry and insert into Oracle Spatial.

The key to creating space objects efficiently in Revit is to start with a well-constructed model. Wherever possible, floors should be continuous throughout a particular building level. Splits in floor heights, e.g. mezzanine levels, and stairwells must be contained by a wall or virtual boundary. These methods should stand as good practice for any modellers intent on using BIM data sourced from architectural designs or captured from laser scans of a building.

The outputted 3D model has a number of shortcomings. Firstly, the polygon faces do not contain internal holes corresponding to windows and doors and further script development is in progress to



overcome this issue. Secondly, issues relating to face orientation require further investigation.

## 6. Conclusion

This investigation has shown that it is possible to extract simplified BIM data from Revit in a format suitable for analysis in 3D GIS. Until FME Workbench is capable of transforming the geometry of *IfcRelSpaceBoundary* objects, Python script is suitable for inserting the boundary faces into Oracle Spatial. Further work will focus on extending the functionality of the Python script to include polygon holes and resolve issues relating to face orientation.

## References

- Autodesk 2014. Revit Help - Exporting a project to IFC. [online] Available at: <<http://help.autodesk.com/view/RVT/2014/ENU/?guid=GUID-14037C31-EBAD-41A8-9099-E6DD65BB626E>> [Accessed 5 August 2014].
- Backes, D., Thomson, C., Malki-Ephshtein, L. and Boehm, J. 2014. Chadwick GreenBIM: Advancing operational understanding of historical buildings with BIM to support sustainable use. In: L. Malki-Ephshtein, C. Spataru, L.M. Halburd and D. Mumovic, eds. *Proceedings of the 2014 Building Simulation and Optimization Conference*, 23-24 June 2014. London, UK.
- Bazjanac, V. 2008. IFC BIM-based methodology for semi-automated building energy performance simulation. *Lawrence Berkeley National Laboratory*.
- Bittner, T., Donnelly, M. and Winter, S. 2006. Ontology and semantic interoperability. In: *Large-scale 3D data integration: challenges and opportunities*. Large-scale 3D data integration: challenges and opportunities, pp.139–160.
- buildingSMART 2014a. IFC 2x3 documentation - IfcSpace. [online] Available at: <<http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/ifcproductextension/lexical/ifcspace.htm#definition>> [Accessed 31 July 2014].
- buildingSMART 2014b. IFC 2x3 Documentation - IfcRelSpaceBoundary. [online] Available at: <<http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/ifcproductextension/lexical/ifcrelspaceboundary.htm#definition>> [Accessed 31 July 2014].
- Cowen, D. 1988. GIS versus CAD versus DBMS: What are the differences? *Photogrammetric Engineering and Remote Sensing*, 54(11), pp.1511–1555.
- Eastman, C.M. 1975. The use of computers instead of drawings in building design. *AIA Journal*, [online] 63(3), pp.46–50. Available at: <[codebim.com/wp-content/uploads/2013/06/Eastman\\_1975.pdf](http://codebim.com/wp-content/uploads/2013/06/Eastman_1975.pdf)>.
- El-Mekawy, M., Östman, A., and Hijazi, I. 2012. A unified building model for 3D urban GIS. *ISPRS International Journal of Geo-Information*

- Ellul, C. 2013. Can topological pre-culling of faces improve rendering performance of city models in Google Earth? In: *Progress and New Trends in 3D Geoinformation Sciences*, Lecture Notes in Geoinformation and Cartography. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.133–154.
- Gröger, G., Kolbe, T.H., Nagel, C. and Hafele, K.-H. 2012. *Open Geospatial Consortium OGC City Geography Markup Language (CityGML) Encoding Standard*. 2nd ed.
- Herring, J.R. 2010. *OpenGIS® Implementation Specification for Geographic information-Simple feature access-Part 2: SQL option*. 1st ed. Open Geospatial Consortium.
- HM Government 2013. *Construction 2025. Industrial Strategy - government and industry in partnership*. HM Government.
- International Alliance for Interoperability (IAI) 2000. *IFC Technical Guide, Industry Foundation Classes–Release 2x.*, T. Liebich & J. Wix, eds., International Alliance for Interoperability.
- Isikdag, U. and Zlatanova, S. 2009. Towards defining a framework for automatic generation of buildings in CityGML using Building Information Models. In: *3D Geo-Information Sciences*, Lecture Notes in Geoinformation and Cartography. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.79–96.
- Laakso, M., Kiviniemi, A., 2012. The IFC standard - a review of history, development, and standardization. *Journal of Information Technology in Construction*, 17, pp.134–161.
- Lee, G., Sacks, R. and Eastman, C.M. 2006. Specifying parametric building object behavior (BOB) for a building information modeling system. *Automation in Construction*, 15(6), pp.758–776.
- Mott MacDonald 2014. Building Information Modelling. [online] mottmac.com. Available at: <<https://www.mottmac.com/article/2385/building-information-modelling-bim>> [Accessed 11 Dec. 2014].
- Murray, C. 2013. Oracle Spatial Developer's Guide 11g Release 2.[online] Available at: <[docs.oracle.com/cd/E11882\\_01/appdev.112/e11830.pdf](https://docs.oracle.com/cd/E11882_01/appdev.112/e11830.pdf)>[Accessed 7 July 2014]
- Mvaerle 2010. GitHub : Python-ifc. [online] Available at: <<https://github.com/mvaerle/python-ifc>> [Accessed 11 Dec. 2014].
- Resch, B., Sagl, G., Törnros, T., Bachmaier, A., Eggers, J.-B., Herkel, S., Narmsara, S. and Gündra, H. 2014. GIS-based planning and modeling for renewable energy: challenges and future research avenues. *ISPRS International Journal of Geo-Information*, 3(2), pp.662–692.
- Safe Software 2012. FMEpedia - Converting BIM IFC data to CityGML. [online] Available at: <[http://fmepedia.safe.com/articles/Samples\\_and\\_Demos/Converting-BIM-IFC-data-to-CityGML](http://fmepedia.safe.com/articles/Samples_and_Demos/Converting-BIM-IFC-data-to-CityGML)> [Accessed 8 Jan. 2015].
- Stoter, J., van den Brink, L., Vosselman, G., Goos, J., Zlatanova, S., Verbee, E., Klooster, R., van Berlo, L., Vestjens, G., Reuvers, M. and Thorn, S. 2011. A generic approach for 3D SDI in the Netherlands. Proceedings of the Joint ISPRS Workshop on 3D City Modelling Applications and the th D GeoInfo Conference Wuhan, China.
- van Berlo, L. and de Laat, R. 2011. Integration of BIM and GIS: The development of the CityGML

- GeoBIM extension. In: *Advances in 3D Geoinformation Systems*, Lecture Notes in Geoinformation and Cartography. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.211–225.
- Weise, M., Liebich, T., See, R., Bazjanac, V., Laine, T. and Welle, B. 2011. *Implementation guide: Space boundaries for energy analysis*.
- Worboys, M. 2011. Modeling indoor space. The 3rd ACM SIGSPATIAL International Workshop. New York, New York, USA: ACM Press, pp.1–6.

## **Biographies**

Gareth Boyes is an experienced engineer and non-practising solicitor. In 2013 he turned his attention to geospatial research and started an MSc in GIS at UCL, London and submitted his dissertation on the “Integration of BIM and GIS”. He is now continuing his research on a PhD studentship at UCL.

Charles Thomson is a PhD student at University College London whose research investigates the automation and validation of parametric geometry reconstruction for BIM from point cloud data. He also has interests in the point cloud data collection process, especially in relation to indoor mobile mapping systems.

Claire Ellul is a Lecturer in Geographical Information Science at University College London. Prior to starting her PhD, she spent 10 years as a GIS consultant in the UK and overseas, and now carried out research into the usability of 3D GIS and 3D GIS/BIM integration. She is the founder and current chair of the Association of Geographical Information's 3D Specialist Interest Group.

# Spoilt for Choice? An Investigation Into Creating Gastner and Newman-style Cartograms

Chris Brunsdon <sup>\*</sup>1 and Martin Charlton <sup>†</sup>1

<sup>1</sup>Maynooth University National Centre for Geocomputation

January 8, 2015

## Summary

A number of choices are encountered when creating cartograms using the Gastner and Newman algorithm. Two important ones are the starting map projection, and the resolution of the grid size used to compute the cartogram transform. We experiment with a number of projection and grid size combinations, and define a measure of ‘cartogram success’, and use this, together with a more descriptive assessment, to identify best practice in choosing resolution and initial projection.

**KEYWORDS:** Cartogram, Rasterisation, Visualisation, Cartography, Map Projection

## 1 Introduction

Cartograms are well established map projections used for the creation of statistical maps that take into account the underlying population of geographical reporting units (GRUs). They have the property that the projected area of each areal unit is proportional to its population<sup>1</sup>. There are a number of standard algorithms to compute cartograms - see for example Tobler (1973), Dorling (1996) or Sagar (2013). Most of these begin with a standard map projection, with populations supplied for each GRU, and then compute a ‘warp’ - of the map drawn in this projection to obtain the cartogram. One particular example of this approach is the Gastner and Newman algorithm (Gastner and Newman, 2004). This solves the *diffusion equation* - Equation 1 below, using a pixel-based approximation.

$$\frac{\partial \rho(x, y, t)}{\partial t} = \nabla^2 \rho(x, y, t) \quad (1)$$

This equation describes the flow of fluids with varying density  $\rho$  at time  $t$  and at each location  $(x, y)$ . For cartograms, it is assumed that  $\rho$  is proportional to population density. Study of Equation 1 reveals that fluids

---

<sup>\*</sup>christopher.brunsdon@nuim.ie

<sup>†</sup>martin.charlton@nuim.ie

<sup>1</sup>At least approximately

will diffuse towards an asymptotic state of uniform density. Thus, for each initial location the mapping onto a particle's asymptotic location provides a cartogram transformation.

### 1.1 Spoilt for choice?

A number of observations can be made about this algorithm.

1. The method is based on approximation.
2. It depends on a grid-based estimation of population density.
3. This estimation is derived from a map using a conventional projection. Several such projections exist.

These raise a number of issues - observation 2 suggests that a number of choices must be made prior to running the cartogram algorithm - namely, at what resolution should the grid be created, and what method should be used to estimate the density values in the grid. Observation 3 implies a further choice - that of the starting map projection - and observation 1 leads to an over-arching issue that, although the cartogram algorithm is often presented as a unique process, there are in fact a number of degrees of freedom. The aim of this paper is to investigate the effect of varying these, and hopefully uncover some 'best practice' recommendations for the creation of cartograms.

## 2 Choices for Initial Conditions

The two key choices relate to the initial map projection, and the resolution of the grid used in the numerical solution of the diffusion equation. The Gastner and Newman algorithm begins with a set of initial densities - typically these are obtained from a set of polygons in a 'conventional' map projection with associated population counts, and assigning population density estimates to each polygon by dividing the count by the corresponding polygon areas. The results are then converted to a raster grid before applying the algorithm - each pixel is assigned to a polygon (approximately) and the density associated with that polygon is then assigned to the pixel - at which point the algorithm is applied.

### 2.1 Choice of Initial Map Projection

There are several 'conventional' map projections<sup>2</sup>, and several possible starting configurations for the algorithm. Map projections can be classified in a number of ways, but one helpful approach here is to use the following classes:

- **Equal Area** - These are projections that give polygons whose area is the same as that on the surface of the Earth. One cost of achieving this is that shapes of areas are prone to distortion.
- **Conformal** - These projections preserve local angles, so that the projected angles where curves meet agree with those on the Earth's surface.

---

<sup>2</sup>actually an infinite number if parameters such as the location of parallels are allowed to vary

- **Equidistant** - These preserve distances from some fixed point, or line on the Earth's surface.
- **Compromise** - These do not attempt to preserve area, distance or local angles, but instead aim to strike a balance between the distortions, in order to produce aesthetically pleasing results.

Projections may also be classified by the developable surface onto which the Earth's surface is projected: cones, cylinders and planes are typical examples (Bugayevskiy and Snyder, 1995). This leads to a cross classification of surface and property, which often features in the name: for example, *Lambert's Azimuthal Equal Area* is an area preserving projection based on a plane.

The projections we have chosen for the experiments described in this paper cover a range of types and developable surfaces. They are listed, together with their properties in Table 1.

Table 1: Projections used

Developable Surface	Projection Name	Type	Abbreviation
Cone	Lambert Conformal Conic	Conformal	LCC
	Albers Equal Area	Equal Area	AEA
	Equidistant Conic	Equidistant	EqDC
Cylinder	Robinson Pseudocylindrical	Compromise	Robin
	Mercator	Conformal	Merc
	Eckert Type VI	Equal Area	Eck6
	Mollweide	Equal Area	Moll
	Equidistant Cylindrical	Equidistant	EQC
Plane	Van der Grinten	Compromise	VanDG
	Stereographic	Conformal	Stere
	Lambert Azimuthal Equal Area	Equal Area	LAEA
	Azimuthal Equidistant	Equidistant	AEqD

It could be argued that equal area projections are the most appropriate choice, since these will give the most accurate estimates of density for the Gaster and Newman algorithm. We therefore include a large number of these. However two questions we wish to address are

- How do results differ between different equal area projections?
- How robust are the results to the use of other kinds of projection?

and on this basis we also include a variety of other projections, representing all combinations of type and developable surface.

## 2.2 Resolution of Raster Representation

As discussed above, once a projection is chosen, the next step is to rasterise the density estimates. A second choice influencing the outcome is the resolution of the raster used to do this. Clearly, the greater this is, the more accurate the result – recall that the algorithm is based on a differential equation representing a

continuous system. On the other hand, computation time will increase with resolution - and it will be useful to identify a point at which no notable improvements are achieved, so that unnecessarily long program runs are avoided. Thus, for each map projection, cartograms are created at a number of resolutions. In each case the raster is square, and of size  $n \times n$  where  $n$  is one of  $\{512, 768, 1024, 1280, 1536\}$ .

### 3 Evaluation

The set of map projections and resolutions outlined above were used to compute cartograms of European economic regions. The software used was Brunsdon's *getcartr* package in R<sup>3</sup>. Cartograms are assessed in two ways here. Firstly, an objective scoring system is used. Although based on approximation, the transformed areas in a cartogram should ideally be proportional to their underlying populations. Thus, if a cartogram algorithm has worked effectively, then

$$A_i = kP_i \quad (2)$$

where  $A_i$  is the area of a zone  $i$  (in cartogram space),  $P_i$  is the corresponding population, and  $k$  is some constant value. Thus, for any given cartogram, fitting a least squares regression line *without an intercept* should give an estimate of  $k$ , say  $\hat{k}$ . Perhaps more usefully, looking at the size of the residuals - that is the values of  $A_i - \hat{k}P_i$  gives an indication of how well this linear fit has worked for each area. Squaring these residuals and summing gives an overall measure of the degree of disagreement in the proportionality between area and population and hence a measure of the success of the cartogram. As a further enhancement, the measurement can be standardised by computing  $1 - R^2$  for the fitted, intercept-free model, allowing for different scale metrics in cartogram space. If we call this measure  $\gamma$ , then it may be seen that

$$\gamma = \frac{\sum_i r_i^2}{\sum_i \hat{k}^2 P_i^2 + \sum_i r_i^2} \quad \text{where } r_i = A_i - \hat{k}P_i \quad (3)$$

For an ideal<sup>4</sup> cartogram, this value will be zero. Thus, this quantity will be computed for each cartogram created.

A second approach to assessment will be to assess the cartograms visually - although it is possible that several cartograms may have  $\gamma$  values near to zero, there could be aesthetic reasons why some may be preferable to others. Although this is a subjective matter, visualisations identifying cartograms with different characteristics will be used to investigate any qualitative traits (for example, excessively stretching certain countries) that may be associated with particular initial projection characteristics.

#### 3.1 Assessment via Objective Scoring

Firstly,  $\gamma$  was computed for each of the twelve projections at the five different resolutions stated above - the results are tabulated in Table 2.

<sup>3</sup><https://github.com/chrisbrunsdon/getcartr>

<sup>4</sup>only in the sense defined above.

Table 2:  $\gamma$  values for Cartograms

Resolution	512	768	1024	1280	1536
LCC	0.936	0.028	0.017	0.012	0.008
AEA	0.929	0.043	0.026	0.016	0.013
EqDC	0.706	0.741	0.017	0.011	0.007
Robin	0.946	0.035	0.740	0.014	0.009
Merc	0.826	0.874	0.018	0.012	0.009
Eck6	0.911	0.032	0.018	0.012	0.008
Moll	0.939	0.035	0.019	0.012	0.009
EQC	0.845	0.871	0.027	0.019	0.013
VanDG	0.955	0.790	0.019	0.012	0.008
Stere	0.911	0.024	0.015	0.010	0.007
LAEA	0.918	0.708	0.020	0.013	0.009
AEqD	0.939	0.027	0.017	0.010	0.008

Patterns are more clearly identified when using graphical approaches. In Figure 1 the value of  $\gamma$  is plotted on a log scale against pixel resolution for each projection. It can be seen generally that after some very poor performances for low resolutions,  $\gamma$  decreases exponentially - the reductions are linear when using the logarithmic scale. One interesting observation is that for the Robinson Pseudocylindrical projection, at one resolution (1024) performance is worse than the next lowest one (768).

In Figure 2 the same relationships are shown, shaded by development surface. Here, it seems there is no clear winner.

In Figure 3 the relationships are shown yet again, shaded by projection class. Here, it seems there is no clear winner - although the conformal projections seem to fair better, with none of them in the worst performing projections. Although in theory one would expect equal area projections to be the most appropriate, they do not seem to be outstanding.

### 3.2 Visual Assessment

In Figure 4 each of the original maps of Europe are shown beside the corresponding cartograms. The presentation exploits Edward Tufte’s notion of ‘small multiples’ and possibly owes an apology to Andy Warhol’s estate<sup>5</sup>. The colouring of the maps corresponds to the developable surface (background), and the class of projection (foreground). The colourings are given in Table 3.

In addition, the layout of the individual projections is given in Table 4.

Visually, some cartograms appear more ‘squashed’ along the north/south axis, although this doesn’t seem to be associated with any particular developable surface, or projection type. In particular, the Albers Equal Area and Equidistant Cylindrical projections are very squashed - and the Mercator notably elongated. It is

<sup>5</sup><http://www.oilpainting-repro.com/prod/marylyn-monroe-3x3-warhol-1925,217.html>



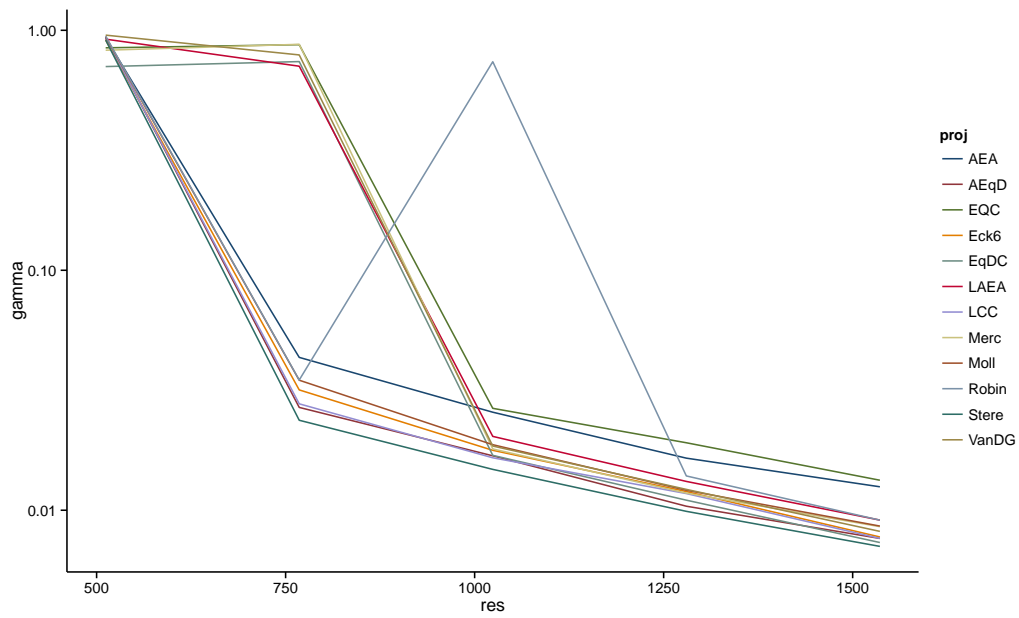


Figure 1:  $\gamma$  vs. Resolution Showing Projection

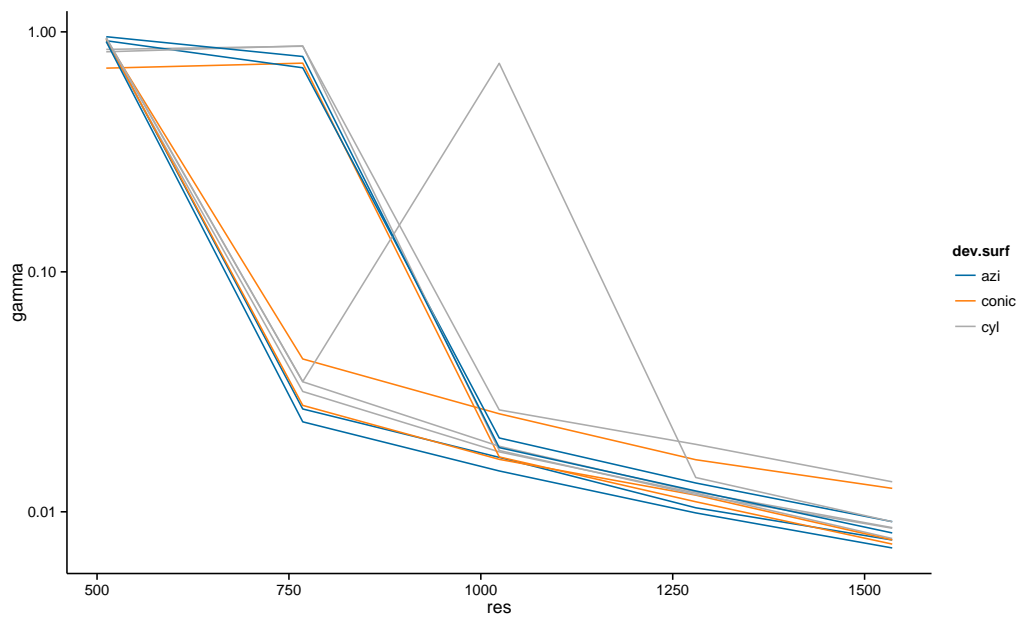


Figure 2:  $\gamma$  vs. Resolution Showing Development Surface

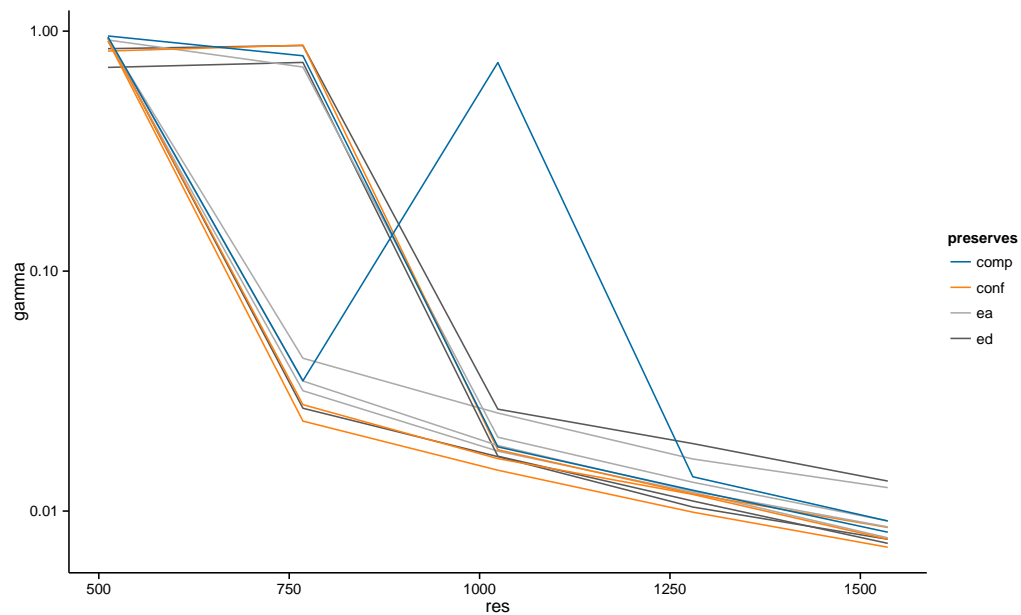


Figure 3:  $\gamma$  vs. Resolution Showing Projection Class

Table 3: Colouring for Figure 4

Characteristic	Colour Feature
Developable Surfaces	
Cone	Purple Background
Cylinder	Orange Background
Plane	Green Background
Projection Type	
Conformal	Blue Foreground
Equal Area	Green Foreground
Equidistant	Purple Foreground
Compromise	Red Foreground

Table 4: Projection Arrangements

	1	2	3	4
1	LCC	AEA	EqDC	Robin
2	Merc	Eck6	Moll	EQC
3	VanDG	Stere	LAEA	AEqD

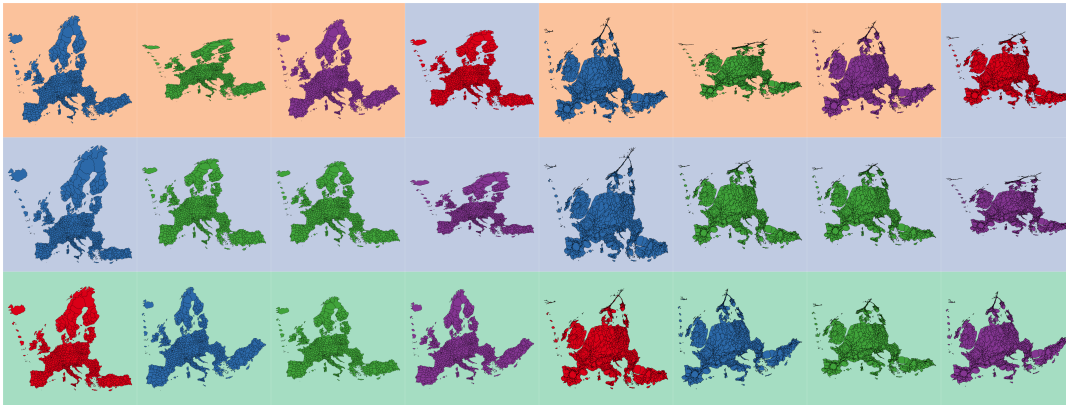


Figure 4: Original Projections (LHS) and Cartograms (RHS)

perhaps an unsurprising fact, but the cartograms seem to take on the squashed/elongated characteristics of the conventional projections used to produce them.

## 4 Conclusions

In relation to the questions asked earlier, it seems that the cartogram algorithm is relatively robust to the initial choice of map projection - despite expectations that equal area projections should outperform others. Of the equal area projections, Eckert Type VI performs best, but is outshone by several other projections. One possible explanation, at least at the Europe-wide scale, is that other factors, such as the inaccuracies due to rasterisation come into play, and their influence outways that of projection type. For work in the near future, world maps will be considered - and presented at the talk we propose here.

In terms of ‘best practice’, choice of projection seems less important than resolution of the raster approximation. Roughly speaking, once some very poor results for low resolutions have been encountered, deviation from a ‘perfect cartogram’ as measured by  $\gamma$  reduces exponentially with resolution. We would suggest checking  $\gamma$  as a way of deciding appropriate resolution, and choosing a squashed or elongated starting projection according to aesthetic preference, bearing in mind the aspect ratio of the resultant cartogram is influenced by this.

## 5 Acknowledgements

We gratefully acknowledge support from the ESPON Programme under the Multidimensional Database Design and Development (M4D) Project. Text and maps stemming from research projects under the ESPON Program presented here do not necessarily reflect the opinion of the ESPON Monitoring Committee.

## 6 Biography

Chris Brunsdon is Professor of Geocomputation and Director of Maynooth University National Centre for Geocomputation. His research interests involve spatial statistics, visualisation and geocomputation applied to a number of areas, including crime pattern analysis, and the analysis of environmental data.

Martin Charlton is a Senior Research Fellow and Deputy Director of Maynooth University National Centre for Geocomputation. His research interests involve geographical information systems, data analysis and geocomputation applied to a number of areas, including health data, and the analysis of housing data data.

Both Chris and Martin played rôles in the development of Geographically Weighted Regression, and are actively involved in developing and implementing tools for this and related techniques in the R programming language.

## References

- Bugayevskiy, L. and Snyder, J. (1995). *Map projections. A Reference Manual*. Taylor and Francis, London.
- Dorling, D. (1996). *Area cartograms: their use and creation*. Number 59 in CATMOG: Concepts and Techniques in Modern Geography.
- Gastner, M. T. and Newman, M. E. (2004). Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499–7504.
- Sagar, B. D. (2013). Cartograms via mathematical morphology. *Information Visualization*, 13(1):42–58.
- Tobler, W. R. (1973). A continuous transformation useful for districting. *Annals New York Academy of Sciences*, 219:215–220.

# Evolutionary Computing for Multi-Objective Spatial Optimisation

Daniel Caparros-Midwood, Stuart Barr and Richard Dawson

School of Civil Engineering and Geosciences, Newcastle University

## Summary

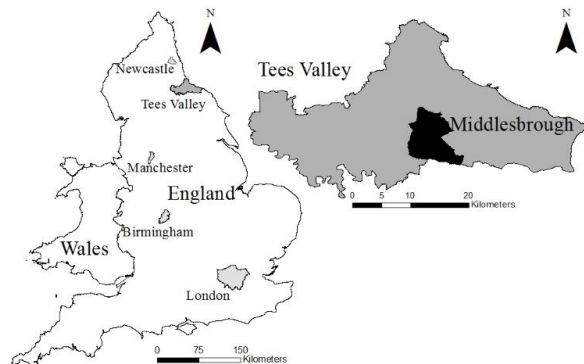
During the transition to more resilient and sustainable cities, planners require robust planning tools to ensure sustainability efforts do not conflict and negatively affect one another. In this paper spatial optimisation is used to provide best trade-off spatial plans between conflicting real world sustainability objectives during the spatial planning process. Using Pareto-optimal optimisation a series of spatial development strategies are derived that outperform all other possible development strategies in at least one objective. When applied to a case study for a north east local authority the resulting spatial Pareto-optimal strategies were found to significantly outperform the local authorities proposed development plan.

**KEYWORDS:** Sustainability objectives, spatial planning, optimisation.

## 1. Introduction

The processes of urbanisation and climate change are necessitating the transformation of cities towards sustainable cities that are robustly adapted to natural (and other) hazards, while simultaneously reducing energy and resource usage to mitigate further climatic change. However the policies required to achieve these frequently conflict with each other, negatively affecting sustainability as a whole. For example, urban intensification with the intention of lowering transport energy costs (Newman and Kenworthy, 1989; Williams, Burton et al., 2000) has been found to exacerbate urban heat islands and increase flood risk as well lead to poor health outcomes for residents (Hunt and Watkiss, 2011; Melia, Parkhurst et al., 2012; Holderness et al., 2013).

Therefore decision makers require robust planning tools to achieve the trade-offs necessary to ensure optimal sustainability (Dawson, 2011). This paper presents the use of a spatial optimisation framework as one method by which multiple positively and negatively correlated sustainability objectives can be evaluated in time and space to assist urban planning. A case study, applied to Middlesbrough Borough Council, a local authority area in the North East of England (Figure 1), demonstrates how a spatial Pareto-optimisation based on a Genetic Algorithm (GA) framework (Goldberg, 1989) can be employed to derive spatial development patterns that are sensitive to climate induced hazards such as heat and flood while accounting for current planning policies that seek to avoid fragmented urban growth and development on green space.

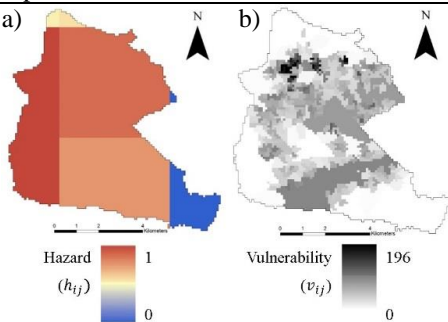
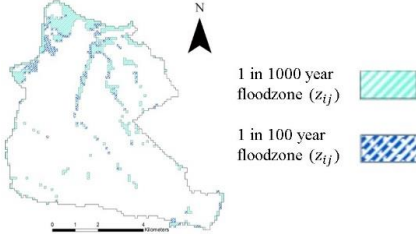
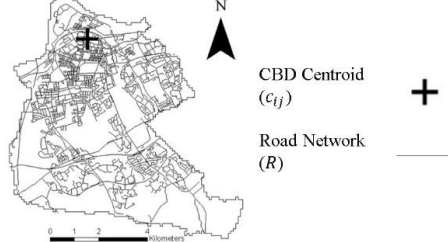
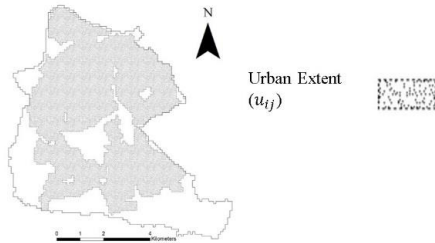
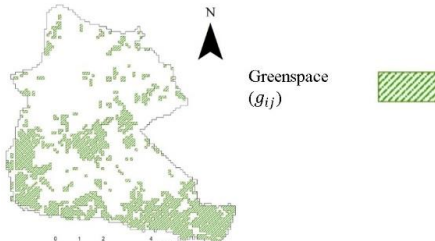


**Figure 1** The case study area of Middlesbrough within the Tees Valley.

## 2. Methodology

A subset of key, real world sustainability objectives were derived from a review of spatial planning and urban sustainability literature: namely; (1) minimizing risk from heat waves; (2) minimizing risk from flooding; (3) minimizing the distance of new development to the current CBD to minimize travel costs; (4) minimizing urban sprawl to prevent increased travel costs; and (5) preventing the development of green-space. Table 1 summarises their parameterisation and the spatial fields used in their calculation.

**Table 1** Parameterisation of Sustainability Objectives

Parameterisation	Inputs
<p><b>1. Minimise risk from heat waves: <math>f_{heat}</math></b>  Characterised by the increase in heat risk in the future relative to the baseline date based on the spatial assignment of new development sites. Heat risk is defined as the cross product of the probability of a heat hazard event and population vulnerability.</p> <p><b>Data Source:</b> a) Spatially disaggregated 2020 heat wave frequency projections (Jones et al., 2009); &amp; b) 2011 census (ONS, 2012) population figures at lower super output area (density per hectare).</p>	
<p><b>2. Minimise risk from flooding: <math>f_{flood}</math></b>  Characterized by a proportional risk assessment of development within 1 in 100 and 1 in 1000 year flood zones represented</p> <p><b>Data Source:</b> UK's Environmental Agency's (EA) Flood zone maps at a 100 meter resolution.</p>	
<p><b>3. Minimise the distance of new development to CBD to minimise travel costs: <math>f_{dist}</math></b>  Characterised by a shortest path between proposed development sites and designated CBD centroid.</p> <p><b>Data Source:</b> Roads and CBDs represented by Ordnance Survey Meridian 2 roads and digitised town centre centroid respectively.</p>	
<p><b>4. Minimise expansion of urban sprawl: <math>f_{sprawl}</math></b>  Calculated as a proportion of new development which falls outside the currently defined urban extent.</p> <p><b>Data Source:</b> Ordnance Survey Meridian Developed Land Use Area.</p>	
<p><b>5. Prevent development of greenspace:</b>  Spatial constraint prevents solutions locating development on areas designated as greenspace</p> <p><b>Data Source:</b> Rasterised Ordnance Survey MasterMap data with Natural theme, reduced to greenspace areas which exceed 2 ha as per Natural England guidelines.</p>	

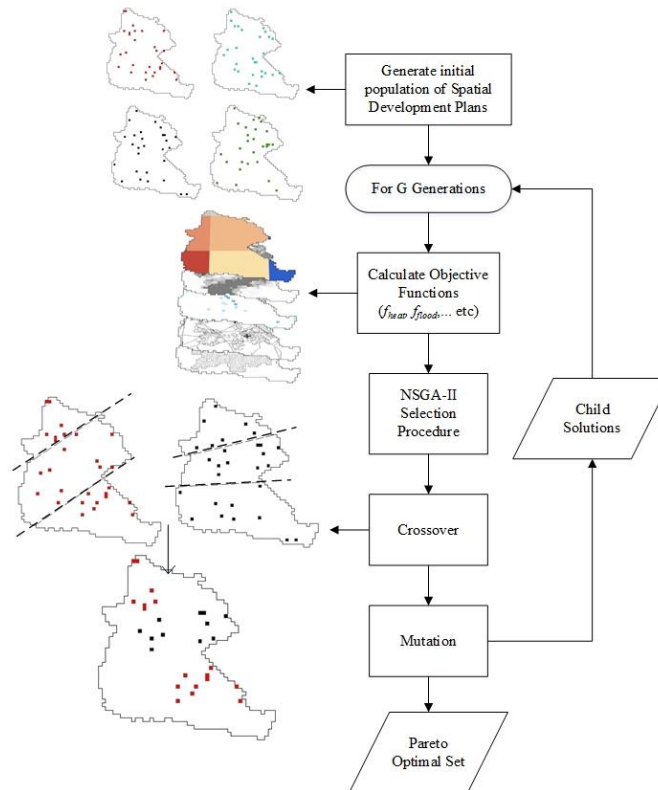
## 2.1 Spatial Optimisation Framework

The developed framework utilises a genetic algorithm which exploits the evolutionary operators of selection, crossover and mutation to converge on superior spatial configurations of development (Figure 2). The framework initialises with a series of random spatial configurations which are modified by the genetic algorithm operators for a set number of generations. The initial spatial plans are evaluated against the objective functions outlined before a selection operator chooses superior solutions to breed a new set of solutions. The selection operator is based on the NSGA-II (Deb et al., 2002) selection method which uses a unique crowding distance metric to ensure that a wide Pareto-front is maintained.

The resulting superior solution-set is then exposed to a crossover operator which combines features from two selected solutions using a two point crossover algorithm. Elements of each solution are exchanged around two randomly selected crossover points along the list of sites to generate two new solutions. This is done with the intention that sliced and merged solutions will lead to superior spatial development plans. Lastly a mutation operator is applied on a small probability which randomly alters the location of a site in order to maintain a diversity of sites in the solutions and prevents premature convergence.

After the prescribed number of GA generations is achieved, a set of Pareto-optimal solutions are returned. These are defined as spatial development plans that out perform all other spatial development strategies in at least one objective function. For a set of objective functions,  $f \in F$  a solution  $s^{(1)}$  is said to dominate solution  $s^{(2)}$  if:

1. The solution  $s^{(1)}$  is no worse than  $s^{(2)}$  in all objectives;  $f(s^{(1)}) \leq f(s^{(2)}) \forall f \in F$ ;
  2. The solution  $s^{(1)}$  is strictly better than  $s^{(2)}$  in at least one objective;  $f(s^{(1)}) < f(s^{(2)})$  for at least one  $f \in F$ .
- (Deb, 2001)

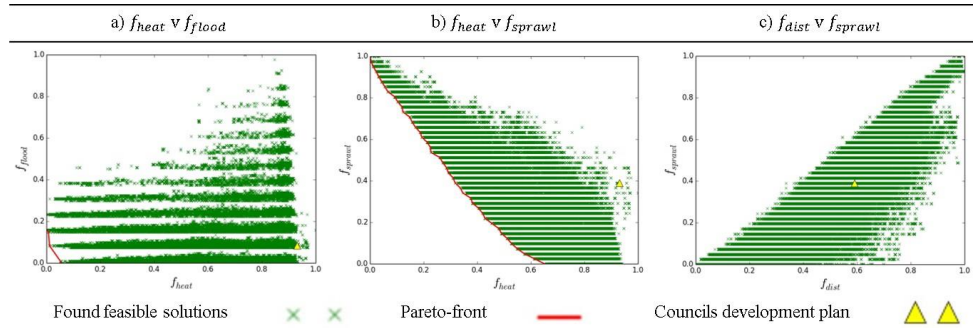


**Figure 2** Genetic Algorithm flowchart.

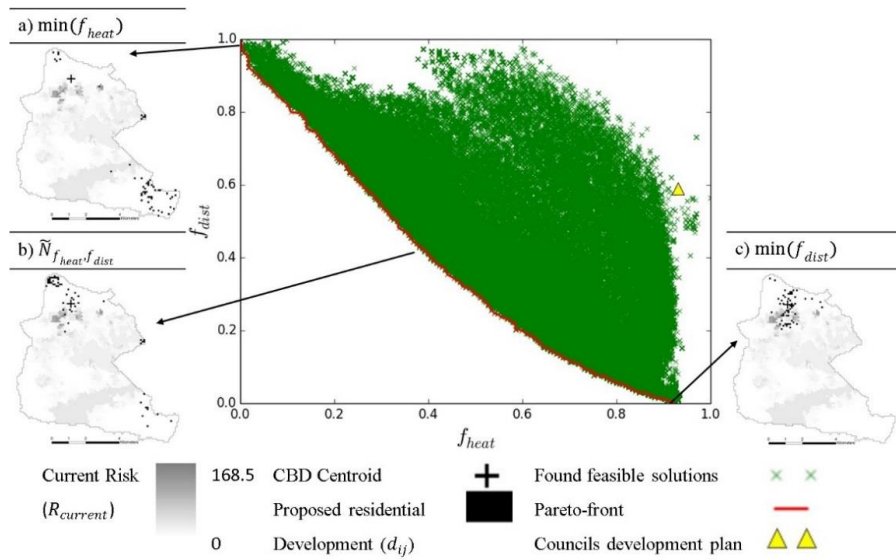
### 3. Results and Discussion

Figures 3 and 4 present the results of the analysis over the case study. The performance of the local authorities development plan (Middlesbrough Council, 2013) is highlighted for comparison, demonstrating that the spatial optimization framework significantly improves upon this in terms of the sustainability objectives investigated. Figure 3 presents the Pareto-fronts (best trade-offs) between sustainability objectives highlighting conflicts between  $f_{heat}$  and both  $f_{flood}$  (Figure 3a) and, to a much greater extent,  $f_{sprawl}$  (Figure 3b). The latter is a result of urbanized areas having higher vulnerability. Despite these conflicts, the spatial optimization is able to identify plans which are best trade-offs. Alternatively  $f_{dist}$  and  $f_{sprawl}$  are simultaneously optimized (Figure 3c) as locations close to the CBD correspond with being within the urban extent.

A major strength of this approach is the ability to co-present sustainability scores alongside the spatial configuration. Figure 4 highlights a series of Pareto-optimal optimised spatial configurations of future development which occur on the Pareto front between  $f_{heat}$  and  $f_{dist}$ . To fully minimise  $f_{heat}$  (Figure 4a) the optimal plan assigns development in the south east of the study at the expense of  $f_{dist}$  whilst to optimize  $f_{dist}$  (Figure 4c) development is assigned to areas surrounding the CBD at the expense of poor  $f_{heat}$  performance. Interestingly, the median solution in the Pareto front (Figure 4b) assigns development in the north of the study area, a trade-off location which avoids the most vulnerable areas whilst being located relatively close to the CBD. By combining mapped results with the Pareto front plot it becomes possible to understand spatially the outcome of selecting a particular optimal solution.



**Figure 3** Pareto front between pairs of objectives



**Figure 4** Pareto-front and resulting Pareto-optimal spatial plans.



#### 4. Conclusion

Spatial optimisation provides a powerful decision support tool to help planners to identify spatial development strategies that satisfy multiple sustainability objectives. The application of the spatial optimization framework demonstrates for the real-world case study the ability to recognize potential development patterns that are potentially more sustainable than the current development plan. Extraction of non-dominated Pareto-optimal spatial configurations provides planners with a clear quantitative and visual characterization of the potential conflicts present between sustainability objectives. Moreover, the use of the Pareto-optimal approach provides a rich set of diagnostic information on possible trade-offs, with the potential to constitute a spatial decision support tool. In-conjunction with further qualitative examination these results could directly inform final planning decisions.

#### 5. Acknowledgements

The research was funded by the UK Engineering and Physical Sciences Research Council, grant EP/H003630/1.

#### 6. Biography

**Daniel Caparros-Midwood** ([daniel.caparros-midwood@ncl.ac.uk](mailto:daniel.caparros-midwood@ncl.ac.uk)) is a PhD student in the School of Civil Engineering, Newcastle University. Daniel's research focuses on the use of optimised spatial planning to manage climate risks and achieve sustainable development. Previously a GIS graduate from Newcastle University.

**Stuart Barr** ([stuart.barr@newcastle.ac.uk](mailto:stuart.barr@newcastle.ac.uk)) is a senior lecturer in Geographic Information Science in the School of Civil Engineering and Geosciences, Newcastle University.

**Richard Dawson** ([richard.dawson@newcastle.ac.uk](mailto:richard.dawson@newcastle.ac.uk)) is the Chair of Earth Systems Engineering ([www.ncl.ac.uk/ceser](http://www.ncl.ac.uk/ceser)) at Newcastle University. Richard's research focuses on the analysis of risks associated with urban and infrastructure systems and the development of adaptation solutions that ensure our infrastructure and cities are resilient and sustainable in the face of intensifying global change.

#### References

- Deb K. (2001) *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons Ltd, Chichester.
- Deb K, Pratap A, Agarwal S and Meyarivan T (2002). A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182-197.
- Dawson R (2011). Potential pitfalls on the transition to more sustainable cities and how they might be avoided. *Carbon Management*, 2(2), 175–188.
- Goldberg, D. E. (1989). *Genetic Algorithms for Search, Optimization and Machine Learning*. Addison-Wesley, Reading.
- Holderness T, Barr S, Dawson R and Hall J (2013) An evaluation of thermal Earth observation for characterising urban heatwave event dynamics using the urban heat island intensity metric. *International Journal of Remote Sensing*, 34(3), 864–884.
- Hunt A and Watkiss P (2010). Climate change impacts and adaptation in cities: a review of the literature. *Climatic Change*, 104(1), 13-49.
- Jones P, Kilsby C, Harpham C, Glenis V and Burton A (2009) *UK Climate Projections Science Report*:

*Projections of Future Daily Climate for the UK from the Weather Generator*, University of Newcastle, UK.

Melia S, Parkhurst G and Barton H (2012). The Paradox of Intensification. *Journal of Transport Policy*, 18(1), 46–52.

Middlesbrough Council (2013). *Local Development Framework Housing Review Preferred Options: Sustainability Appraisal*. Middlesbrough, UK.

Newman, P. and Kenworthy J.R. (1989). *Cities and automobile dependence: a sourcebook*. Gower, Aldershot.

ONS (Office of National Statistics) (2012). *2011 Census: Population Estimates for the United Kingdom*. London.

Williams K, Burton E and Jenks M (2000). Achieving Sustainable Urban Form: Conclusions. In Williams K, Burton E. and Jenks, M. (ed.), *Achieving Sustainable Urban Form*. Routledge, London, 347-55.

# sDNA: how and why we reinvented Spatial Network Analysis for health, economics and active modes of transport

Crispin HV Cooper<sup>\*1</sup>, Alain J Chiaradia<sup>†1</sup>

<sup>1</sup>Sustainable Places Research Institute, Cardiff University

<sup>2</sup>School of Planning and Geography, Cardiff University

October 31, 2015

## Summary

We introduce sDNA, a GIS/CAD tool and methodology for analysis of spatial networks. The design decisions behind the tool are documented, in particular the choice of standardizing on the network link in order to match existing data standards and increase computational efficiency. We explore the effects of this decision on algorithm design, and present results that validate our decision to depart from a recent tradition and revive a much older one.

**KEYWORDS:** network analysis, space syntax, health, transport, pedestrian

## 1. Introduction

sDNA is both a methodology and a GIS/CAD plug-in for the analysis of spatial networks, compliant with widely accepted data standards. It has recently been used to provide environmental morphometrics for the 500,000-point UK BioBank database (UK Biobank 2013) – a major national health resource - in a project which won the 2014 RTPI award for excellence in spatial planning research (RTPI 2014). Its uses are not restricted to health, however; sDNA has also been used in mass transport investment option analysis (for Shanghai, 2014), environmental footprinting (Collins and Cooper 2014) and social cohesion studies (Cooper, Fone, and Chiaradia 2014) as well as numerous planning consultancy projects (for Trowbridge 2012, London Borough of Merton 2013, City of London 2014, Paris 2010-12). The plug-in is made available both freely and commercially.

In this short paper we discuss why we thought spatial network analysis needed reinventing (a risky undertaking given the software development time involved); how our design criteria helped to shape the new approach, and how the new approach has been validated. It thus serves of a record as to why the framework we offer is structured as it is.

## 2. Background

Spatial network analysis is an old discipline, dating back at least to Euler who in 1736 solved the problem of the Seven Bridges of Königsberg (Coupy 1851). This single piece of work necessitated the invention of both network codification and network generalisation and was the beginning of modern graph theory (Biggs, Lloyd, and Wilson 1986). Closeness - a mean shortest path measure on networks - and betweenness - a measure of flow derived from closeness, were defined in parallel in sociology (Bavelas 1948; Freeman 1977), communications networks (Shimbel 1953), transport network analysis (Garrison and Marble 1962; Ford and Fulkerson 1962; Kansky 1963), and geography (Haggett and Chorley 1969). Today these are applied in social network analysis research

---

\* CooperCH@cardiff.ac.uk

† ChiaradiaAJ@cardiff.ac.uk

(Freeman 2004), social media (Facebook) and Google's Page rank can be conceptualised as a re-invention of accessibility weighted by opportunities importance (Hansen 1959). The concept of mean shortest path was also used by Christaller (1933) and in Reilly's (1931) gravity model following Ravenstein (1885; 1889). Mean shortest path is in turn part of the early definition of accessibility, either weighed by the importance of destinations (Hansen 1959), or unweighted (Ingram 1971), which also underpin most of the last 50 years of transportation modelling.

More recently, spatial network analysis at a highly spatially desegregated (street link) level has been applied to a variety of problems in economic and transport modelling. Such analyses can be predicated on the existence of origin/destination data points (MIT 2011; ESRI 1999), or alternatively can dispense with collection of such data, instead studying only the structure of the network itself. Haggett and Chorley (1969), Kinsky (1963), early space syntax research (Hillier and Hanson 1984) and place syntax (Stahle, Marcus, and Karlstrom 2008) take this latter approach. Space Syntax used an axial line network codification; closeness and betweenness on axial lines gave good correlations with measured pedestrian and vehicle flows (Hillier et al. 1993; Penn et al. 1998). This was subsequently adapted to segmented axial line (Hillier and Iida 2005) and segmented link (Turner 2007).

### **3. Rationale**

Our aim in developing sDNA was to adapt such techniques to current cartography standards and datasets. On the network mapping side, most cartography uses the link-node standard: e.g. OS ITN (Ordnance Survey 2011a), the European road representation standard (ISO 2011), US Tiger lines, OpenStreetMap and indeed the data underlying any modern satnav system. It thus made sense to make link representation an intrinsic part of the algorithm in sDNA. There are additional reasons for doing so, however. Density of links - rather than of segments or network length - can give a good proxy for origins and destinations in the absence of such data (Chiaradia et al. 2012). For a given level of detail, the links in a network are uniquely defined (unlike axial lines - see Ratti 2004) and hence make sense as a mapping standard. Finally using whole links, rather than link segments, reduces the number of entities in an analysis thus increasing computational efficiency.

Basing sDNA on established data models has the further advantage of allowing us to leverage additional data such as OS Mastermap topography layer (Ordnance Survey 2011b), addressing layer (Address Layer2) as well as survey or census data. Currently this is done by joining related data to each network link, leaving network links as the fundamental unit of analysis - although in cases where greater precision is required, links may be split to accommodate data points. We take an approach that offers a spectrum of options between all- and no-data alternatives, both for origin and destination weights and for what is variously called cost or impedance: the quantity which is minimized in selection of shortest paths. (We prefer the simpler term 'distance', which may be defined differently depending on the analysis).

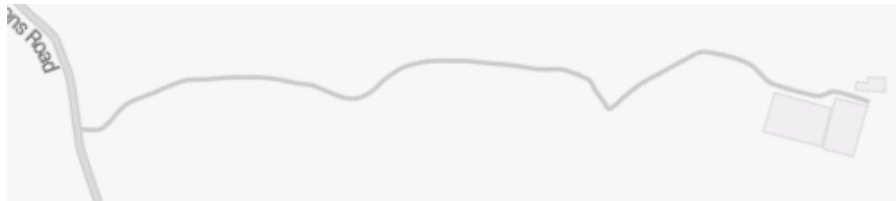
### **4. Methodology**

Standardizing on the network link required a new algorithmic approach, in particular with respect to determining the locality of analysis. As axial lines are a cognitive representation of space based on lines of sight, it is reasonable to treat them as discrete entities due to the cognitive cost of traversing between them. This is also true of angular segments, based on the cognitive and time costs incurred by turning corners (Caldwell 1961; Kirby 1966; Turner 2001). The same cannot be said of network links: while a user of a high street could probably tell you how many corners - if any - the street has, or how long it is, the number of links that constitute its length is likely of no direct relevance to them (illustrated in Figure 1 - although indirectly, the link density probably correlates to their reason for being there). Thus, while the approach taken by e.g. Hillier and Iida (2005) is to measure locality as a step radius, in terms of the number of axial lines or segments that are traversed, we choose instead to follow those who measure locality of analysis in Euclidean network terms (Kinsky 1963; Haggett and Chorley 1969; Sheffi 1985; Porta, Crucitti, and Latora 2006; Turner 2007), for example considering

all links within 1 km of each origin as measured along the network. We refer to such a locality as a network radius. But the similarity of approaches ends here, because links – unlike axial lines or angular segments – are not only perceptually non-discrete, but also cannot be treated as indivisible from this standpoint (Figure 2).



**Figure 1** Queen Street, Cardiff is composed of 8 network links although perceptually is close to being a single straight line. Base map © OpenStreetMap contributors



**Figure 2** A long rural lane with multiple corners consists of only one link, but perceptually consists of multiple segments. © OpenStreetMap contributors

As traversing an exact distance from each origin is likely to result in the inclusion of partial portions of links, we define two modes for handling this situation. In *discrete space*, we choose to include a link in the analysis if its centre falls within the network radius, and exclude it otherwise. In *continuous space*, our algorithm dynamically cuts each link at the point where the edge of the radius is reached, and in subsequent analysis the partial link receives a partial weighting (as shown in Equation 1) in any subsequent calculation of closeness, betweenness, network quantity or averaging of other network qualities such as diversion ratio (Chiaradia, Cooper, and Webster 2012):

$$P(L) = \frac{\text{length of } L \text{ falling within radius}}{\text{length of } L} \quad (1)$$

On large scales, the difference between the two is minimal as the portions of network ‘wrongly’ included or excluded in discrete space are small in comparison to the network which is wholly within the radius. On short scales however, particularly those relevant to pedestrian transport, continuous space analysis can produce a significantly more accurate result.

## 5. Results and Outcomes

Convenient though it may be to adapt spatial network analysis to match modern data representations, there was a risk was that the resulting models would not be representative of human behaviour. Fortunately, empirical testing proved their worth in validation against measured pedestrian and vehicle flows. Table 1 presents some correlations between sDNA measures and measured flows for the classic space syntax dataset in central London. This micro to macro behavioural performance is

what underpins sDNA's applicability and reliability in the studies we cited in the introduction. The theoretical implications will be discussed in a forthcoming paper.

**Table 1** Correlations between sDNA network variables and measured flows

Network variable		Num. measured points (n)	Mean Angular Distance (Closeness)	Angular Betweenness
$r^2$ with vehicle flow	Barnsbury	82	0.73****	0.74****
	Clerkenwell	42	0.82****	0.80****
	South Ken.	46	0.82****	0.76****
	Brompton	61	0.60****	0.58****
	(mean)		0.74	0.72
$r^2$ with ped. flow	Barnsbury	102	0.69****	0.60****
	Clerkenwell	51	0.78****	0.73****
	South Ken.	62	0.49****	0.58****
	Brompton	85	0.56****	0.47****
	(mean)		0.63	0.60

\*\*\*\* indicates significance  $P < 0.0001\%$

Recent adaptations to the software enable (1) analysis in full 3-d, (2) choice of theoretic hybrid distance metrics based on pedestrian route choice analysis (3) output of shortest paths and network buffers, (4) large scale network analysis (e.g. whole UK, over  $10^6$  links). As an example of a 3d hybrid metric, height changes can be valued differently according to whether they take place on stairs, escalator or elevator; and these factors in turn valued in proportion to Euclidean distance and angular change. Such features have been already been shown to be useful in modelling pedestrians in complex multi-level urban environments, and cyclist travel mode and route choice in relation to slope, with further applications in transport and health research as well as land use planning.

## 6. Biographies

Crispin Cooper is lead developer for the sDNA software, and also researches applications of network analysis in health, community cohesion and active modes of transport. Crispin has a Master's degree in Computer Science (University of Cambridge, 2002) and a PhD in City & Regional Planning (Cardiff University, 2010).

Alain Chiaradia is the concept lead for sDNA software design. He has 20 years' experience in software design and use of spatial design network analysis as honorary senior research fellow at UCL and as consultant to local authorities, developer, and urban designer in the UK and internationally. His research interest spans from economic geography to urban design economics.

## 7. Author contributions

AC developed the initial concept and conducted most of the literature review. CC worked on the rationale, software development and analysis. Both authors reviewed and edited the final version.

## References

- Bavelas, Alex. 1948. "A Mathematical Model for Small Group Structures." *Human Organisation* 7: 16–30.
- Biggs, N L, E Keith Lloyd, and Robin J Wilson. 1986. *Graph Theory 1736-1936*. Oxford: Oxford University Press.

- Caldwell, T. 1961. "On Finding Minimum Routes in a Network With Turn Penalties." *Communications of the ACM* 4: 107–8.
- Chiaradia, Alain, Crispin Cooper, and Chris Webster. 2012. *sDNA a Software for Spatial Design Network Analysis, Specifications*. UK: Cardiff University. <http://www.cf.ac.uk/sdna/wp-content/downloads/documentation/Detailed%20sDNA%20measure%20descriptions.pdf>.
- Chiaradia, Alain, Bill Hillier, Christian Schwander, and Martin Wedderburn. 2012. "Compositional and Urban Form Effects in Centres in Greater London." *Urban Design and Planning - Proceedings of the ICE* 165: 21–42.
- Christaller, W. 1933. *Die Zentralen Orte in Sddeutschland - The Central Places of Southern Germany*. Jena - Englewood Cliffs, NJ: Gustav Fischer Verlag - Prentice-Hall.
- Collins, Andrea, and Crispin H V Cooper. 2014. *The Environmental Impacts of Festivals: Reflections on the 2012 Hay Literature Festival, Wales*.
- Cooper, Crispin H. V., David L. Fone, and Alain Chiaradia. 2014. "Measuring the Impact of Spatial Network Layout on Community Social Cohesion: A Cross-Sectional Study." *International Journal of Health Geographics* 13 (1): 11. doi:10.1186/1476-072X-13-11.
- Coupy, E. 1851. "Solution D'un Problème Appartenant À La Géometrie de Situation, Par Euler." *Nouvelles Annales de Mathématiques Ire Série* 10: 106–19.
- ESRI. 1999. *ArcGIS Network Analyst*. <http://www.esri.com/software/arcgis/extensions/networkanalyst>.
- Ford, L R, and D R Fulkerson. 1962. *Flows in Networks*. 2nd ed. Princeton: Princeton University Press.
- Freeman, L C. 1977. "A Set of Measures of Centrality Based on Betweenness." *Sociometry* 40 (1): 35–41.
- . 2004. *The Development of Social Network Analysis*. North Charleston: BookSurge LLC.
- Garrison, W L, and D F Marble. 1962. *The Structure of Transportation Networks*.
- Haggett, Peter, and Richard J Chorley. 1969. *Network Analysis in Geography*. London, UK: Hodder & Stoughton Educational.
- Hansen, W G. 1959. "How Accessibility Shape Land Use." *Journal of American Institute of Planners* 25: 73–76.
- Hillier, Bill, and Julianne Hanson. 1984. *The Spatial Logic of Space*. Cambridge: Cambridge University Press.
- Hillier, Bill, and S Iida. 2005. "Network and Psychological Effects: A Theory of Urban Movement." In *Proceedings of the 5th International Space Syntax Symposium*. TU Delft: Techne Press.
- Hillier, Bill, Alan Penn, T Grajewski, and J Xu. 1993. "Natural Movement: Or, Configuration and Attraction in Urban Pedestrian Movement." *Environment and Planning B: Planning and Design* 20: 29–66.
- Ingram, D R. 1971. "The Concept of Accessibility: A Search for an Operational Form." *Regional Studies* 5: 101–7.

- ISO. 2011. *Intelligent Transport Systems - Geographic Data Files (GDF) - GDF5.0*. 14825.  
[http://www.iso.org/iso/home/store/catalogue\\_ics/catalogue\\_detail\\_ics.htm?csnumber=54610](http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=54610).
- Kansky, K J. 1963. *Structure of Transport Networks: Relationships between Network Geometry and Regional Characteristics*. Chicago: University of Chicago.
- Kirby, R F. 1966. "A Minimum Path Algorithm for a Road Network with Turn Penalties." In Sydney: 3rd Australian Road Research Board (ARRB) Conference.
- MIT. 2011. *Urban Network Analysis*. <http://cityform.mit.edu/projects/urban-network-analysis-toolbox>.
- Ordnance Survey. 2011a. *OS Mastermap ITN Layer Urban Paths Theme*.  
<http://www.ordnancesurvey.co.uk/docs/user-guides/ITN-paths-tech-userguide.pdf>.
- . 2011b. *OS Mastermap Topography Layer User Guide and Specification*.  
<http://www.ordnancesurvey.co.uk/docs/user-guides/os-mastermap-topography-layer-user-guide.pdf>.
- Penn, Alan, Bill Hillier, David Banister, and J Xu. 1998. "Configurational Modelling of Urban Movement Networks." *Environment and Planning B: Planning and Design* 25: 59 – 84.
- Porta, S, P Crucitti, and V Latora. 2006. "The Network Analysis of Urban Streets: A Primal Approach." *Environment and Planning B: Planning and Design* 33: 705–25.
- Ratti, Carlo. 2004. "Space Syntax: Some Inconsistencies." *Environment and Planning B: Planning and Design* 31 (4): 487–99. doi:10.1068/b3019.
- Ravenstein, George, Ernest. 1885. "The Laws of Migration." *Journal of the Statistical Society of London* 48: 167–235.
- . 1889. "The Laws of Migration." *Journal of the Royal Statistical Society* 52: 241–305.
- Reilly, W J. 1931. *The Law of Retail Gravitation*. New York: Knickerbocker Press.
- RTPI. 2014. *Excellence in Spatial Planning Research Awards*. <http://www.rtpi.org.uk/briefing-room/news-releases/2014/september/excellence-in-spatial-planning-research-awards-winners/>.
- Sheffi, Yosef. 1985. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Englewood Cliffs: Prentice Hall.
- Shimbel, A. 1953. "Structural Parameters of Communication Networks." *The Bulletin of Mathematical Biophysics* 15 (4): 501–7.
- Stahle, A, L Marcus, and A Karlstrom. 2008. "Place Syntax: Geographic Accessibility with Axial Lines in GIS." In Delft: Proceedings in 5th Space Syntax Symposium.
- Turner, Alasdair. 2001. "Angular Analysis." In *3rd International Space Syntax Symposium*. Atlanta.
- . 2007. "From Axial to Road-Centre Lines: A New Representation for Space Syntax and a New Model of Route Choice for Transport Network Analysis." *Environment and Planning B: Planning and Design* 34 (3): 539–55.
- UK Biobank. 2013. *BioBank*. <http://www.ukbiobank.ac.uk/>.



# HAG-GIS: A spatial framework for geocoding historical addresses

Daras K<sup>\*1</sup>, Feng Z<sup>†1</sup> and Dibben C<sup>‡2</sup>

<sup>1</sup> School of Geography & Geosciences, University of St Andrews

<sup>2</sup> School of Geosciences, University of Edinburgh

November 7, 2014

## Summary

The Digitising Scotland (DS) project aims to digitise the 24 million vital events record images (births, marriages and deaths) for all residents in Scotland since 1855 (ie transcribe them into machine encoded text). This will allow research access to information on individuals and their families for those who have ever lived in Scotland between 1855 to the present day. In this paper we present the methodology for geocoding these 24 million historical addresses in Scotland from 1855 to 1974 by introducing the Historical Address Geocoder – GIS (HAG-GIS) spatial framework and its matching algorithms implemented for the needs of the DS project. The matching processes link the historical addresses to the contemporary addresses by exact and fuzzy matching algorithms. Apart from geocoding the historical addresses, we also produce pseudo registration district boundaries using the pilot historical addresses from death event records in 1950 and 1951.

**KEYWORDS:** Digitising Scotland, HAG-GIS, geocoding, address matching

## 1. Introduction

In Scotland, the Registration Districts (RDs) were used from 1855 as result of the Registration of Births, Deaths and Marriages (Scotland) Act 1854 and registrars were appointed in each district to maintain the statutory registers (births, marriages and deaths). These records are very detailed and have changed very little in content over time, making them a highly valuable record of change over time and space. Apart from the great importance for research, the detailed nature of the record allows records for the same person to be linked and then link persons into families. Originally the Digital Imaging of the Genealogical Records of Scotland's people (DIGROS) project converted all statutory

---

\* kd54@st-andrews.ac.uk

† zf2@st-andrews.ac.uk

‡ Chris.Dibben@ed.ac.uk

vital events records into digital image format (PDF) with a supporting index, where only forenames, surname, year and RD code were indexed. The purpose of DS project is to build on the DIGROS index and extend it through digitisation for research purposes. It has four main objectives: 1) digitise the 24 million vital events record images (births, marriages and deaths) for all residents in Scotland since 1855 (ie transcribe them into machine encoded text), 2) standardise and code occupation descriptions to the Historical International Standard Classification of Occupations (HISCO), 3) standardise and code all deaths to a modified form of the International Classification of Disease 10 (ICD10) and 4) link address information to consistent geographies through time. This will allow research access to information on individuals and their families for those who have ever lived in Scotland between 1855 to the present day. The combination of temporal and spatial historical information will provide a rich demographic database system for the Scottish population of a similar potential depth and breadth as the Scandinavian and Low countries (Edvinsson, 2000; Mandemakers, 2000; Thorvaldsen, 2000).

In this paper we present the methodology for geocoding 24 million historical addresses in Scotland from 1855 to 1974 by introducing the HAG-GIS system and its matching algorithms implemented for the needs of the DS project. The overall matching process links the historical addresses to the contemporary addresses by exact and fuzzy matching algorithms. Apart from geocoding the historical addresses, we produce pseudo registration district boundaries using the pilot historical addresses from death event records in 1950 and 1951. This way we create a historical geographical dataset where new insights into the Scottish geographies of the past are available. A large number of projects (Fitch and Ruggles, 2003; Gregory and Healey, 2007; St-Hilaire et al., 2007) have used similar approaches to assist in developing historical GIS systems where the spatial boundaries are unknown but textual information is available in various public records.

## **2. HAG-GIS: System framework**

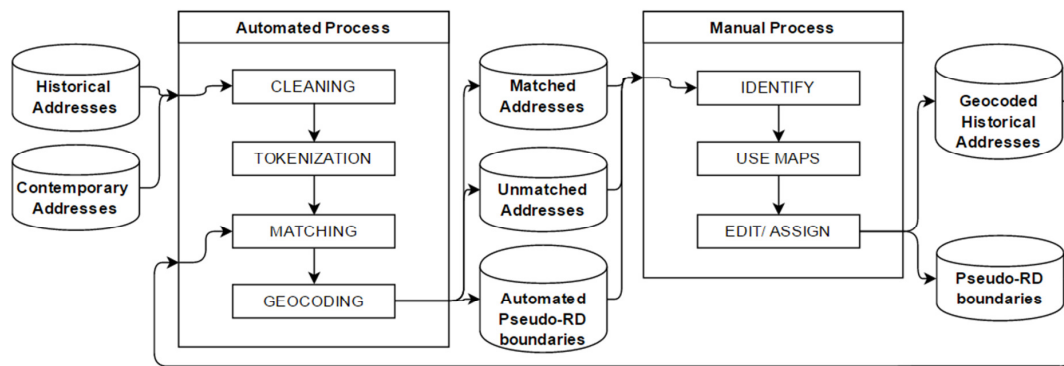
The HAG-GIS application has been designed to be used with historical data for Scotland, targeting special data characteristics and providing required tasks of data analysis and normalisation. Its implementation is based on the Python language using the SQLite database for storing the data providing portability to any operational system platform. The geocoding process introduced here is related to efficiently organising historical addresses by eliminating redundancy and ensuring data dependencies. The geocoding process has two distinct phases: a) the automated phase where the HAG-GIS system performs an exact and a fuzzy matching to link each record to a particular geographical reference point, and b) the manual phase where the clerks check/edit the unmatched addresses using the historical maps from Ordnance Survey (Figure 1). Importantly in the first phase, HAG-GIS exploits the two types of geographical information retained in the records to assign a geocode to the record: firstly the geographically high resolution but 'dirty' address information and

secondly the less high resolution but much ‘cleaner’ Registration District information. The programme cycles between these two types of information to fully geocode the dataset.

During the automated phase, the HAG-GIS system executes the following tasks using a SQLite database to store the data:

- a) the cleaning task where the system transforms all the address data into a normal form by lowercasing all the characters of address, stripping possible whitespaces, removing punctuations and removing duplicate addresses from the database,
- b) the tokenisation task where each address is split into word tokens and each token could be marked as possible unchanged token or linked to name aliases stored in the database,
- c) the address matching task where in order to match historical addresses to contemporary addresses we use exact and fuzzy matching algorithms such as the Measuring Agreement on Set-Valued Items (Passonneau, 2006) and Levenshtein edit-distance respectively,
- d) the geocoding task where the system assigns the appropriate reference point for each matched address and uses these reference points to build artificial polygons (Thiessen polygons). In addition we use the K-Dimensional Tree algorithm for nearest neighbours (Maneewongvatana & Mount, 2001) for creating series of point density surfaces for each RD code. Thus, the system can exclude reference points that are assigned to wrong RD codes or distant address locations.
- e) Finally the Thiessen polygons are assigned to the pseudo-RDs allowing us to enable a new matching phase through a more relaxed matching criterion and to direct a historical map linked clerical phase.

Even though the manual phase is not yet implemented, its basic framework components are presented for a complete introduction to the HAG-GIS system framework (figure 1). Thus, the pseudo-RDs will be used by the clerks to manually identify and geocode addresses to grid references in combination with the historical maps. The clerks will use the Ordnance Survey six-inch to the mile historical maps in Scotland for the period 1892-1905 for manually edit unmatched historical addresses. If the manual allocation of some addresses to the appropriate grid references is not feasible then those addresses are assigned to the center of the pseudo-RD. The quality of the geocoding process and the pseudo-RDs is secured by calculating the density of matched addresses for each pseudo-RD ( $\text{index} = \text{matched addresses} / \text{all addresses}$ ). This way, the manual process is concentrated on the pseudo-RDs with low index value.



**Figure 1** HAG-GIS system framework

### 3. Pilot study

In this pilot study we utilise 129,694 historical addresses derived from a sample digitisation of the death records for the 1950 and 1951 years in Scotland. Each historical address refers to an individual and provides information about the building number or name of property, the street name, the town name and the registration district (RD) code. In addition, the catalogue of contemporary addresses are provided by the NRS address register, which includes the building number or name of property, the street name, the town name, postcodes and national grid references.

Using the aforementioned historical and contemporary addresses and the HAG-GIS geocoding system, we performed an initial automated address matching process using a 80% match cut-off point in fuzzy matching so that all addresses whose matching scores are above a certain membership will be determined as a match. The initial address matching process was completed in under an hour using a single CPU processor. The HAG-GIS system was able to match 36,412 (42%) of the 87,495 cleaned historical addresses with 25,761 (29%) addresses being exact matched. In Table 1, we can see the initial matched addresses by matching thresholds.

**Table 1** The initial matched addresses in the pilot study by matching threshold

Num. of addresses	Match threshold			Un-matched addresses
	100%	$\geq 90\%$	$\geq 80\%$	
87495	25761 (29%)	33288 (38%)	36412 (42%)	51083 (58%)

After the end of initial matching process, the system created temporary Thiessen polygons using the 80% matched addresses assigning a weight for each polygon (Figure 2a & 2b). The weights are computed using the nearest neighbours' method (k-dimensional tree) for each RD number. We selected only the 80% matched addresses because the system can overcome the problem of partially

matched addresses with incorrect geo-reference and can provide a clean artificial tessellation for constructing the initial set of pseudo-RD boundaries. Having all the Thiessen polygons weighted for each RD, the system automatically assigned the RD number with the highest weight to the relevant Thiessen polygon. Then, the HAG-GIS system repeated the matching process for each unmatched address focusing on the contemporary addresses that fall within the pseudo-RD's bounding box of the unmatched address with the same RD code. This time we used a less strict threshold (50%) for configuring the system because the safe estimation of the initial pseudo-RD boundaries supported us with the geographical extent of each pseudo-RD and we could narrow down the search results considerably. In later phases of the work we will test the quality of the results by different thresholds in order to establish an optimal value. The HAG-GIS system was able to deliver a final match of 62,602 (72%) of the 87,495 cleaned historical addresses with 24,893 (28%) addresses being unmatched. In Table 2, we can see the final matched addresses by matching thresholds. The final pseudo-RDs have been improved extensively and they are imitating closely the Bartholomew Post Office Plan (1939-40) boundaries (figure 2c & 2d).

**Table 2** The final matched addresses in the pilot study by matching threshold

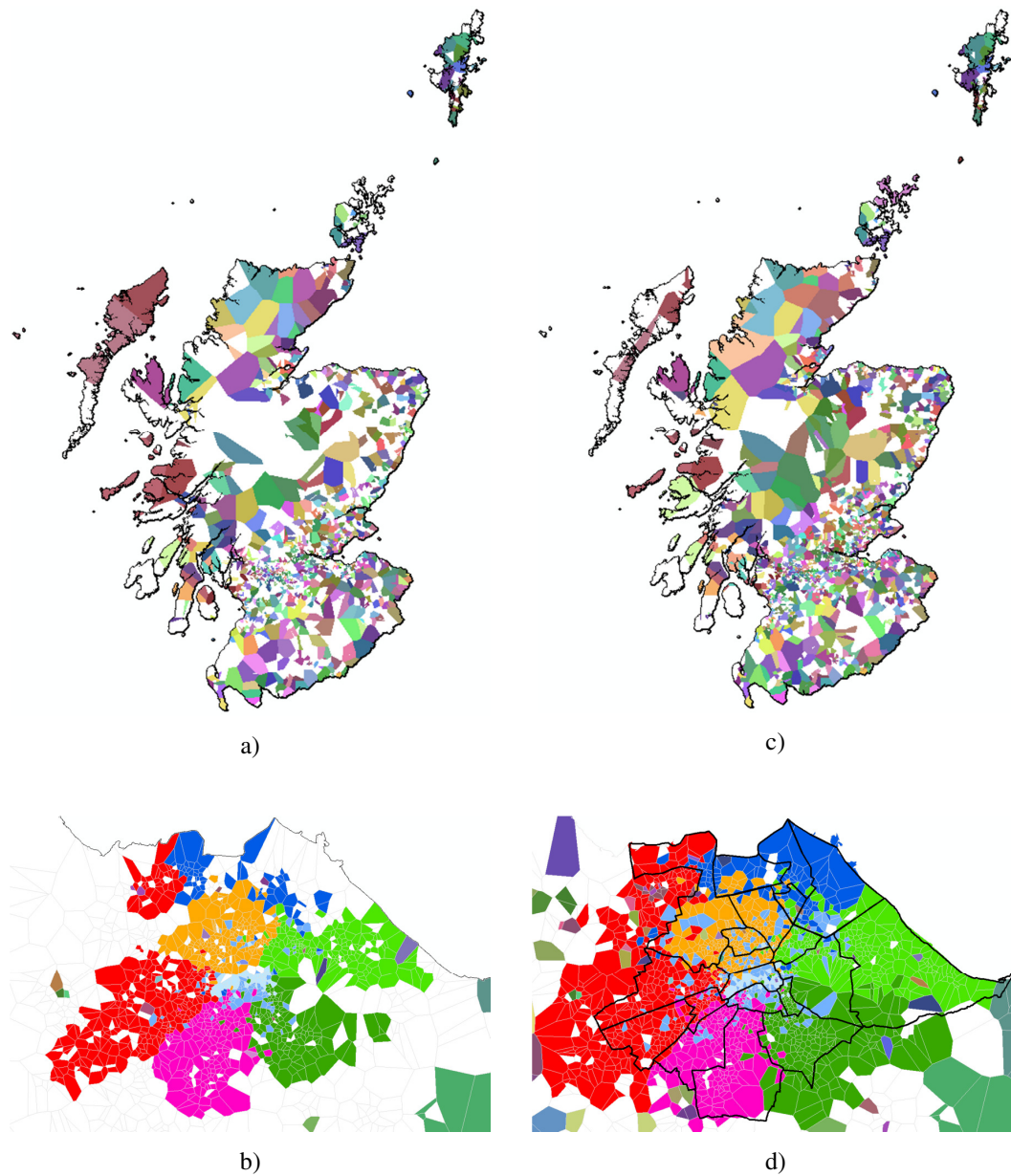
Total addresses		Match threshold						Un-matched addresses
		100%	≥90%	≥80%	≥70%	≥60%	≥50%	
Counts	87495	26067	33865	37671	41364	49869	62602	24893
%	100	30	39	43	47	57	72	28

#### 4. Conclusions

This paper has presented the HAG-GIS spatial framework for geocoding historical addresses in Scotland using the pilot data for two distinct years (1950 and 1951). The preliminary results suggest that the overall geocoding methodology performs well and the matching process is expected to further improve the final results when the pilot data will be replaced with the full data. However, there are possible improvements that can be facilitated to address issues related to the size of urban and rural pseudo-RDs, the precision of geocoder for long or short streets and better handling of various transcribing errors. Further work will involve the effects to the geocoding precision using other matching algorithms such as phonetic match algorithm etc. and the optimisation of essential phases during the automated and manual processes.

#### 5. Acknowledgements

The Digitising Scotland project is funded by ESRC. The support from National Records of Scotland is also gratefully acknowledged.



**Figure 2** Thiessen polygons created from the HAGGIS system: a) & b) pseudo-RDs of Scotland and Edinburgh at the initial phase, c) pseudo-RDs of Scotland at the final phase & d) Bartholomew Post Office Plan (1939-40) boundaries layered over the pseudo-RDs of Edinburgh at the final.

## References

- Edvinsson S (2000) The Demographic Data Base at Umeå University—a resource for historical studies. In: Hall PK, McCaa R, and Thorvaldsen G (eds), *Handbook of international historical microdata for population research*, Minneapolis: Minnesota Population Center, pp. 231–248.
- Fitch CA and Ruggles S (2003) Building the National Historical Geographic Information System. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 36(1), 41–51.
- Gregory IN and Healey RG (2007) Historical GIS: structuring, mapping and analysing geographies of the past. *Progress in Human Geography*, 31(5), 638–653.
- Mandemakers K (2000) Historical sample of the Netherlands. In: Hall PK, McCaa R, and Thorvaldsen G (eds), *Handbook of international historical microdata for population research*, Minneapolis: Minnesota Population Center, pp. 149–178.
- Maneewongvatana S and David M M (2001) *On the efficiency of nearest neighbor searching with data clustered in lower dimensions*, Springer Berlin Heidelberg, pp. 842–851.
- Passonneau R (2006) Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.
- St-Hilaire M, Moldofsky B, Richard L, et al. (2007) Geocoding and Mapping Historical Census Data: The Geographical Component of the Canadian Century Research Infrastructure. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 40(2), 76–91.
- Thorvaldsen G (2000) The Norwegian Historical Data Centre. In: Hall PK, McCaa R, and Thorvaldsen G (eds), *Handbook of international historical microdata for population research*, Minneapolis: Minnesota Population Center, pp. 179–206.

# Quantifying the deterrent effect of police patrol via GPS analysis

Toby Davies<sup>\*1,2</sup>, and Kate Bowers<sup>†1</sup>

<sup>1</sup>Department of Security & Crime Science, University College London, Gower Street, London

<sup>2</sup>Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London

January 8, 2014

## Summary

The efficacy of police patrolling as a means of crime deterrence remains a significant area of uncertainty within crime prevention study. It is, however, an issue of substantial practical importance, since the design of policing strategies depends crucially on knowledge of the form and intensity of intervention required to achieve a given effect. Here, we examine GPS traces of police vehicle movement in a major UK city in order to quantify precisely the patrol activity applied to each street segment. This is then compared against crime data to test, and estimate the magnitude of, the deterrent effect of patrolling.

**KEYWORDS:** police patrol, street networks, crime prevention, survival analysis.

## 1. Background

Patrol is one of the primary tactics employed by the police in the course of their efforts to prevent crime. In accordance with the principles developed by Sir Robert Peel, which remain influential in current models of policing, one of the functions of the police is to demonstrate the presence of a legal force that has the authority to punish transgressions of law. Patrol fulfils this remit by increasing the awareness among potential offenders of the risks associated with committing offences. That this should prevent crime can be rationalised by appealing to rational choice theory: if a prospective offender is considered to act with (bounded) rationality, any increase in risk will reduce the anticipated utility of an offence and thus reduce the likelihood that the individual will choose to offend. Patrolling therefore represents a well-grounded and efficient crime prevention activity, and has been adopted widely for this reason.

While it is common to the majority of policing models, however, the concept of patrolling is a broad one, and considerable variation can be observed in its use. To take one basic aspect, the volume of resource allocated can differ substantially: the relative emphasis placed on proactive patrolling and reactive response varies considerably across policing models. There are also differences, however, in the form of patrolling. For example, the role of vehicular patrol, as opposed to that carried out on foot, has increased over time, and the balance between the two continues to vary between settings. The nature of patrol activity is also dependent on overall strategy. The approach of ‘neighbourhood policing’, for example, suggests that officers should be integrated with the local community, fostering a familiar and co-operative relationship. This contrasts with a ‘zero tolerance’ approach, in which many resources are committed to individual areas and high-impact tactics are employed.

Given that such variation exists within police patrolling, the question naturally arises of which forms are most effective in deterring crime. This is an issue of clear practical importance: if patrolling is to be applied in an evidence-based manner, then it is essential to have some understanding of the level and type of patrolling necessary to achieve a given effect. This need is made even more urgent by

---

\* toby.davies@ucl.ac.uk

† kate.bowers@ucl.ac.uk



recent developments in the field, such as predictive policing. A crucial underlying assumption for systems of this type is that crimes occurring in predicted locations can be prevented by focussed police activity (of which patrolling is typically the principal form). The success of such strategies is therefore contingent on the deterrent effect of patrolling: if predicted crimes cannot be prevented, the accuracy of a predictive algorithm is immaterial. We argue that these issues are addressed only partially by existing research.

## **2. Previous Research**

A number of studies have been concerned with the evaluation of patrol-based interventions as a means of reducing crime. Many of these are focussed on the particular tactic of ‘hot spot policing’, in which small geographic areas with acute crime problems are patrolled intensively, and these are summarised in a systematic review by Braga *et al.* (2014). The meta-analysis presented demonstrated that focussed patrolling had been modestly but significantly successful at controlling crime problems. While this is a useful finding, however, it refers only to patrolling in one particular context: hot-spot policing is a specific intervention, applied over and above existing activities. The effect of routine day-to-day patrolling – which, we argue, is of greater overall significance for policing – must be considered independently of this.

Some indication of the effectiveness of day-to-day police activity can also be found by considering the relationship between overall police resource levels and crime. A number of studies examine this (and, indeed, it is the subject of a systematic review by Lee *et al.*, 2013); however, since the analysis is only carried out at macro level, there are a number of reasons why positive results in these studies cannot be taken as a reliable indication of the effect of patrolling. Officers working in better-resourced departments may have the freedom to undertake enhanced policing on a number of fronts: they might improve victim support, for example, or have faster response to real-time incidents. To presume that an increased capacity to patrol is the mechanism by which the increased crime reduction is found in such places is difficult to justify.

Inferential reliability is not the only reason to analyse patrol effects at the micro level; it is also necessary in order to understand properly the mechanism by which policing patrol might reduce crime. A number of possibilities exist for how the deterrent effect is manifested, some more viable than others. In hotspot policing studies, many authors assert that it is the visible presence of officers which is the fundamental driver of deterrence. This would stop offenders undertaking offences at the time of the police patrol only; however, it is also hypothesised that patrol activity also has a residual deterrence effect. In other words, it alters the behaviour of would-be offenders for a period after the patrol has gone. Koper (1995), for example, suggests that foot patrol has a positive deterrent effect on crime within an 11-15 minute timeframe, compared with a drive-by patrol approach. Again, this is only partial evidence, as the study in question is primarily observational.

## **3. The present study**

The scarcity of fine-grained quantitative research concerning patrolling is, in fact, unsurprising, when the practical requirements of such research are considered. The problem is primarily one of data: until recently, systematic recording of police locations at micro level simply did not take place, so that there was no means of systematically measuring the presence of police officers at particular locations. This is the information required to measure patrol ‘dosage’ – the key concept in assessing policing intensity in space and time – and its absence has traditionally precluded such analysis. Recent years, however, have seen the widespread proliferation among police officers of GPS-enabled devices, via which the required location data can be systematically recorded. The present research seeks to address a number of the questions raised in the previous section by analysing one such dataset.

### **3.1. Data and pre-processing**

The data in question relate to police vehicle movement for one area of a major UK city. In the

recording system used, vehicles are tracked at regular intervals, providing a log of movement as they are used by officers. In addition to locational information, a number of attributes are also recorded, including the speed of travel and ‘blue light’ status; these are both variables with which the effect of patrolling might be expected to vary. The objective of the research is to examine patrol intensity at the street segment level in terms of both its distribution and its effect on crime.

The first aspect of the research to be discussed will concern a number of technical issues which arise in the study of such data. GPS records of this type are, by nature, subject to significant uncertainty, both in their location and other attributes. Since the research is dependent on reliable association of patrol activity with specific street segments, however, this must be eliminated in order to ensure the validity of results. We will discuss the methods used for data cleaning and the process of associating records with the street network, both of which involve the use of bespoke geo-processing and routing algorithms.

### 3.2. Patrol distribution analysis

Having discussed the data processing, we will present analysis of the overall spatio-temporal distribution of police patrol activity. The comprehensive nature of the data means that each street segment can be profiled in terms of all visits by police vehicles: the time at which the segment was visited, the travel speed, and whether the vehicle stopped can all be examined. We will demonstrate that the overall distribution of activity displays particular interesting features: certain areas receive disproportionate levels of coverage, and the usage level of some streets conflicts with what would be expected on the basis of their network centrality (such as the network metric ‘betweenness’; see Figure 1). We will discuss what can be inferred about both individual officer behaviour and the influence of central command on the basis of these findings.

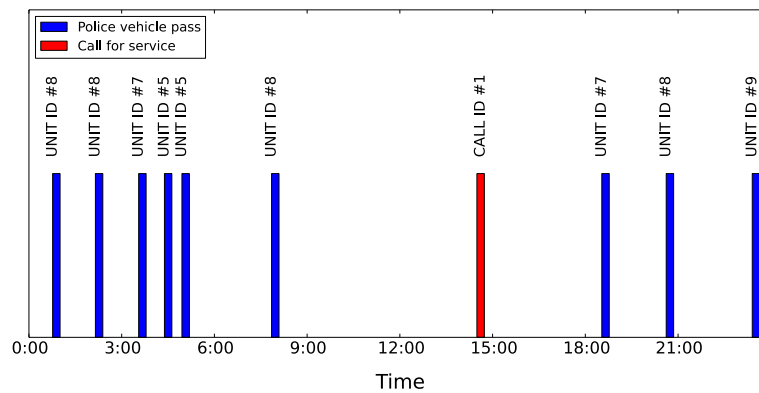


**Figure 1** Street segments coloured according to network betweenness, which provides a first-order estimate of the likely usage levels of each segment during travel through the network.

### 3.3. Testing and measurement of deterrence

Finally, we will present analysis of the relationship between patrolling and the occurrence of crime. Using data concerning crime events for the same area and time period as the vehicle movement data, each street segment can also be profiled in terms of the exact times at which incidents took place. The relationship between these two event types can therefore be analysed, with the street segment as the

unit of analysis. Figure 2 shows an example of one day's activity for a particular street segment: this exemplifies the non-uniformity typically observed in the temporal distributions. We will present analysis of the dependence between these two distributions, examining in particular the extent to which patrol visits tend to be followed by periods of no criminal activity. The findings of this analysis provide important evidence concerning the residual effect of patrolling. Finally, we will show how the situation can be framed in terms of survival analysis, with patrol activity playing the role of treatment and criminal events representing failures. Preliminary analysis of this type will be presented, and implications for practice subsequently discussed.



**Figure 2** Daily log of both patrol visits and calls for police service on an individual segment: examining the intervals between events allows the deterrent effect of patrol activity to be examined.

## References

Braga A A, Papachristos A V & Hureau D M (2014). The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis. *Justice Quarterly*, 31(4), 633-663.

Lee Y-J, Corsaro N & Eck J E (2013). Police Force Size and Crime: A Systematic Review of Research from 1968 to 2013. Presented at the *69th Conference of the American Society of Criminology (ASC)*, Atlanta, Georgia.

Koper C S, (1995). Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots. *Justice Quarterly*, 12(4), 649-672.

## Acknowledgements

This work is part of the project - Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks ([www.ucl.ac.uk/cpc](http://www.ucl.ac.uk/cpc)), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) is greatly appreciated.

## Biography

Toby Davies is a Research Associate working on the Crime, Policing and Citizenship (CPC) project at UCL. His background is in mathematics, and his work concerns the application of mathematical techniques in the analysis and modelling of crime. His research interest include networks and the analysis of spatio-temporal patterns.

Kate Bowers is a Professor in Crime Science at the UCL Department of Security and Crime Science. Kate has worked in the field of crime science for almost 20 years, with research interests focusing on the use of quantitative methods in crime analysis and crime prevention.

# Ephemeral Londoners: Modelling Lower Class Migration to Eighteenth Century London

Adam Dennett<sup>\*1</sup>, Adam Crymble<sup>†2</sup> Tim Hitchcock<sup>‡3</sup> and Louise Falcini<sup>§4</sup>

<sup>1</sup>Centre for Advanced Spatial Analysis, University College London

<sup>2</sup>Department of Humanities, University of Hertfordshire

<sup>3</sup>Department of History, University of Sussex

<sup>4</sup>School of Law, University of Reading

October 10<sup>th</sup>, 2014

## Summary

Between 1750 and 1801 the population of London grew from approximately 750,000 to 1.1 million people. Relocating to London in the eighteenth century only occasionally generated a paper trail, but a significant number of failed migrants were rounded up for ‘wandering and begging’ on the streets and sent back from whence they came to their parish of legal settlement. Records of these removals have been digitised and are used in this paper to model migration into London, to throw light onto the patterns of movement at this time.

**KEYWORDS:** Migration, Historical Data, Vagrancy, Spatial Interaction Modelling

Between 1750 and 1801 the population of London grew from approximately 750,000 to 1.1 million people. According to Wrigley (1967), to retain this rate of growth the metropolis needed to attract in excess of 8,000 more migrants per year than it lost through death or out-migration. But from where was London drawing these migrants?

Relocating to London in the eighteenth century only occasionally generated a paper trail. Eighteenth century scholars have no censuses, as do those of the mid-to-late nineteenth century. Those eighteenth century migrants who integrated successfully and managed to make a modest living have largely disappeared. Yet some have remained visible. A significant number of failed migrants were rounded up for ‘wandering and begging’ on the streets and sent back from whence they came to their parish of legal settlement (Taylor 1991). Under the authority of the 1744 Vagrancy Act, local magistrates had classified these people as ‘vagrants’ before forcibly removing them from the county under a scheme designed to reduce the burden of economically fragile migrants on local ratepayers Eccles, 2012). Between December 1777 and April 1786, the county’s vagrancy contractor, Henry Adams, submitted lists to the county eight times per year detailing the names as well as the final destinations of those he had transported (Hitchcock et al. Forthcoming). For Adams, the lists were his way of billing the county for completed work; for contemporary academics they represent a set of records that can be geo-referenced and modelled as a way of understanding from where a set of lower class failed migrants to London originated.

Many, but not all of these lists survive (42 out of a possible 65). This allows us to identify the place of settlement of 11,489 individuals removed from urban Middlesex to counties beyond, which according to a 1785 report by Adams to the Middlesex bench, amounts to roughly seventy-five per cent of the

---

<sup>\*</sup> a.dennett@ucl.ac.uk

<sup>†</sup> a.crymble@herts.ac.uk

<sup>‡</sup> T.Hitchcock@sussex.ac.uk

<sup>§</sup> l.falcini@reading.ac.uk

total number of vagrants transported. The gaps in the records are unevenly spread through the seasons, making it difficult to draw firm conclusions about annual trends in migration and expulsion – though these almost certainly do exist. Because of this happenstance of historical survival, we know relatively little about removals in May and August, compared to October.

The records also bias our knowledge of migration towards certain parts of the country. Henry Adams had been hired to shepherd vagrants north and west, but a different contractor, or possibly a different system entirely was used for those heading south and east. This leaves us nearly blind to migration from and removal to the counties of Norfolk, Suffolk, Essex, Kent, Sussex, Surrey, and Hampshire. Finally, the lists highlight two distinct types of vagrants that are not obvious to a casual observer: those waywards who had likely been arrested for disorderly behaviour and processed through the Houses of Correction (referred to hereafter as ‘vagrants’), and those who had volunteered to leave the area in exchange for free passage, particularly after 1783, and processed by the Lord Mayor (referred to hereafter as ‘volunteers’). This latter group included demobilized servicemen and seasonal labourers, and the origins of these voluntary leavers is significantly different from those arrested as disorderly. With this in mind, this research focuses each group separately, since the ‘vagrants’ typically arrived in London of their own volition, often by foot, and many of the ‘volunteers’ were dumped in the capital at the end of the American Revolutionary War, meaning a single explanation for the geospatial patterns we observe for both groups is unlikely to fit.

These two groups represent a very small subset of migrants to London during a relatively short period of time in the latter eighteenth century. A better understanding is possible by incorporating yet another group: those arrested in Middlesex between 1801 and 1805. These individuals and their parish of origin are recorded in the Middlesex Criminal Registers when they were checked into gaol and their demographic details written down (referred to hereafter as ‘criminals’). Though there is little evidence that vagrants, volunteers, and criminals were one and the same, they did all tend to come from those economically unprivileged groups who occupied the lower rungs of the socio-economic hierarchy. By looking at all three groups: vagrants, volunteers, and criminals, between 1776 and 1805, it becomes possible to determine to what extent migration patterns into London were following predictable patterns, and to what extent each group and each region of the British archipelago had a unique relationship with London.

A number of scholars have attempted to answer these questions through both indirect theories of human migration, and directly through targeted studies. The theories are old and well tested, laid down by Ravenstein (1885-9) and Zipf (1946). Ravenstein’s seminal work on migration theory in Britain is still the starting point for discussing migration. It was based on studies of the 1871 and 1881 censuses of England and Wales, and outlined a number of laws for human migration, including the propensity of migration to be step-wise, of women to travel shorter distances than men, and of long-distance migrants to end their journey in a great centre of commerce or industry. Half a century later, Zipf’s ‘P1 P2D Hypothesis’ mathematically modelled migration, and concluded people travel only as far as required to find an acceptable economic opportunity, reinforcing the importance of short-haul migration.

In this paper we draw heavily on the work of Zipf and many who have followed since employing gravity/spatial interaction models to help understand the patterns and processes of flows of economically underprivileged individuals to London in the late eighteenth and early nineteenth centuries. To our knowledge this is the first time these techniques have been used to cast light on historical migration flows of this kind and as such we hope that as increasing volumes of historical data are digitised, georeferenced and archived in fields such as History and the Digital Humanities, Geographic Information Science can continue to offer new analytic insights.

## **Biography**

Adam Dennett is a Lecturer in Urban Analytics in the Centre for Advanced Spatial Analysis, University College London. Adam is a Geographer, with interests broadly in the areas of population, quantitative methods, GIS and spatial analysis. Prior to joining UCL, Adam completed his PhD in the School of

Geography at the University of Leeds.

Adam Crymble is a Lecturer of Digital History at the University of Hertfordshire. He did his PhD in history and digital humanities at Kings College London. He earned his MA in Public History, his BA in History, and his Certificate in Writing and Rhetoric, from the University of Western Ontario

Tim Hitchcock is Professor of Digital History at the University of Sussex. He completed his DPhil on the evolution of eighteenth-century parochial workhouses at St Antony's College, Oxford in 1985, and began his academic career at the then Polytechnic of North London. He has published ten books on the history of poverty, sexuality and street life, primarily focussed on eighteenth-century London.

Louise Falcini is a Sessional Lecturer at the University of Reading. Louise's research examines the evolution of regulation concerning civic and personal cleanliness in eighteenth-century London. She has worked extensively on the judicial and poor law archives of Middlesex.

## References

- Eccles, A. (2012) *Vagrancy in Law and Practise under the Old Poor Law*, ch. 2 & 8. Ashgate, London
- Hitchcock, T. Crymble, A. and Falcini, L. (Forthcoming) 'Loose, Idle and Disorderly: Vagrant Removal in Late Eighteenth-Century Middlesex', *Social History*
- Ravenstein, E.G. (1885) "The Laws of Migration." in the *Journal of the Royal Statistical Society*. 48:167-227
- Ravenstein, E.G. (1889) "The Laws of Migration" in the *Journal of the Royal Statistical Society*. 52:241-301
- Snell, K.D.M. (1991) 'Pauper Settlement and the Right to Poor Relief in England and Wales', *Continuity and Change* 6, no. 3: 382.
- Taylor, J.S. (1976) 'The Impact of Pauper Settlement 1691-1834.' *Past & Present* no. 73: 42-74. doi:10.2307/650425
- Wrigley, E.A. (1967) 'A Simple Model of London's Importance in the Changing English Society and Economy 1650-1750', *Past and Present* 37, no. 1: 46.
- Zipf, G.K. (1946) 'The P1 P2D Hypothesis: on the Intercity Movement of Persons', *American Sociological Review*, Vol. 11, No. 6, 677-686.

# Identifying perpetuation in processes driving fish movement

Matt Duckham<sup>\*1</sup>, Antony Galton<sup>†2</sup> and Alan Both<sup>‡1</sup>

<sup>1</sup>University of Melbourne, Australia

<sup>2</sup>University of Exeter, UK

January 9, 2015

## Summary

This extended abstract explores ongoing work that is developing new models and algorithms capable of identifying the environmental drivers of human and animal movement. Specifically, the paper presents an algorithm able to identify the perpetuating conditions for movement: those ranges of environmental variables that are necessary for movement to occur.

Our algorithm is tested on fish movement data from a large, long-term ecology study in Australia, combined with environmental data about water temperatures, levels, and salinity. The results demonstrate the types of rules that can generated from real movement patterns using our algorithm.

**KEYWORDS:** causation, processes, perpetuation, context-aware movement analysis, environmental monitoring, data mining.

## 1 Introduction

Context-aware movement analysis (CAMA) aims to relate movement to the underlying geographic context in which that movement is embedded (Laube, 2014). Understanding how geographic space drives movement patterns is arguably of much greater importance in most applications than analyzing the geometry of the movement patterns (a perspective often ignored by traditional movement analysis approaches, cf. Laube et al., 2005; Buchin et al., 2011). In this extended abstract we explore ongoing work developing techniques capable of identifying the environmental drivers of movement. Specifically, the work reported concerns the identification of perpetuating conditions for movement: those ranges of environmental conditions that are necessary for movement to occur, even though these conditions may not directly be the causes of movement. The approach, tested on fish movement data from a large long-term ecology study in Australia, demonstrates the types of rules generated from real movement patterns.

---

<sup>\*</sup>mduckham@unimelb.edu.au

<sup>†</sup>apgalton@ex.ac.uk

<sup>‡</sup>alan.both@gmail.com

## 2 Perpetuation

Our algorithm for identifying perpetuation is a component of a larger attempt to identify other types of causation too, including previous work on identification of causal events in fish movements (Bleisch et al., 2013, 2014). Our general model of causation has as its primary focus the relationships between events, some of which are *causes* while others are *effects*. The most general form of causal rules we are working towards handling is:

$$[\text{Causes} \mid \text{Conditions}] \Rightarrow \text{effect after Delay},$$

where

- **Causes**  $\subset \mathcal{C}$ ,  $\mathcal{C}$  is a set of causes (environmental events),
- **Conditions** is a set of *conditions*, where each condition is a triple  $c = (p_c, v_c^-, v_c^+) \in \mathcal{P} \times \mathbb{R} \times \mathbb{R}$ , and  $\mathcal{P}$  is the set of all environmental processes
- **effect**  $\in \mathcal{M}$ ,  $\mathcal{M}$  is the set of effects (movement events).
- **Delay** is a *delay interval*  $[d^-, d^+]$ , where  $d^-, d^+$  are integers such that  $0 \leq d^- \leq d^+$ .

In a condition,  $v_c^-$  and  $v_c^+$  are the limits of a range within which the value of  $p_c$  must fall in order for the condition to hold. The environmental process  $p_c$  is drawn from the set of environmental processes  $\mathcal{P}$  (which might include processes such as temperature, turbidity, water level, and so forth in the case of fish movements). Note that following Galton (2012), only events can strictly *cause* other events. When one ongoing process is responsible for the continuing operation of another, we refer to this as *perpetuation* rather than cause. Within our general scheme, perpetuation effects can be simulated by dropping the **Causes** term and using the **Conditions** to encode the perpetuating processes. The **effect** is then an event acting as a proxy for the perpetuated process (as might arise, for example, from discrete observations of what is in reality a continuous process). This results in rules of the form:

$$\text{Conditions} \Rightarrow \text{effect after Delay}$$

which are the target of the investigations reported here.

### 2.1 Problem statement

Now assume a data set recording the occurrences of movement events  $\mathcal{M}$  at every timestep over some time period  $T = [0, n]$  along with the continuously varying values of relevant environmental processes  $\mathcal{P}$  across the geographic space in which movement takes place. Our problem is to identify a compact set of rules of the form **Conditions**  $\Rightarrow$  **effect after Delay** that accurately describe this data.



## 2.2 Algorithm

Although our full algorithm for dealing with causing events and their interactions with perpetuating processes is rather complex, the component for dealing with perpetuation alone can be relatively simply sketched. Informally, we can:

1. label each time step  $t \in T$  as “good” (if the movement effect occurs) or “bad” (if the movement effect does not occur);
2. sort the labeled time steps according to the values of a chosen environmental process  $p_c$ ;
3. identify consecutive runs of time steps within the sorted sequence; and
4. take the union of the sets of intervals delimiting the start and end points of this each;

In cases where there is only one rule driven by only one perpetuating environmental process, this procedure performs well. Extending the approach to multiple rules (explaining different movement effects), and multiple environmental process (with different conditions), and time delays (resulting from inevitable temporal and spatial granularity effects upon the detection of movement and monitoring environmental processes), makes our full algorithm somewhat more complex. Further, although this algorithm is explicitly temporal, it is not explicitly spatial (as indeed might be expected for a general treatment of causality). Our results also consider the spatial coincidence of perpetuating processes and effects, on the assumption that causal relationships operate in only over immediate temporal and spatial proximity. However, a fuller discussion of the algorithm and these issues is beyond the scope of this extended abstract.

## 3 Preliminary results and outlook

Our data set involved the set of movement events of more than 1000 tagged fish in the Murray River, south eastern Australia, monitored over five years using a network of 18 logging towers that partition the river system into 24 zones (Koehn et al., 2008). Figure 1 depicts the 24 zones and their downstream adjacencies.

The effects were taken to be the set of fish movements. In our analysis we distinguished upstream and downstream movements, as well as movements between different zones. For example, Figure 2 summarizes the number of downstream movement events between adjacent zones over the five year time period. The Figure shows significantly more movement in later years (most likely a larger-scale effect itself of the decade-long drought in southern Australia which ended in 2010).

Data about water temperature, water level, and salinity from monitoring stations along the river was used as the environmental processes. Space in this extended abstract does not permit an exploration of the full set of results. However, Table 1 gives examples of the best rules found for water level in selected zones, but typical of other zones and processes. The final column in Table 1 shows the  $F_1$  score for these rules. The  $F_1$  score is calculated as the harmonic mean of the precision (positive predictive value) and sensitivity (true positive rate) and provides a useful measure of rule

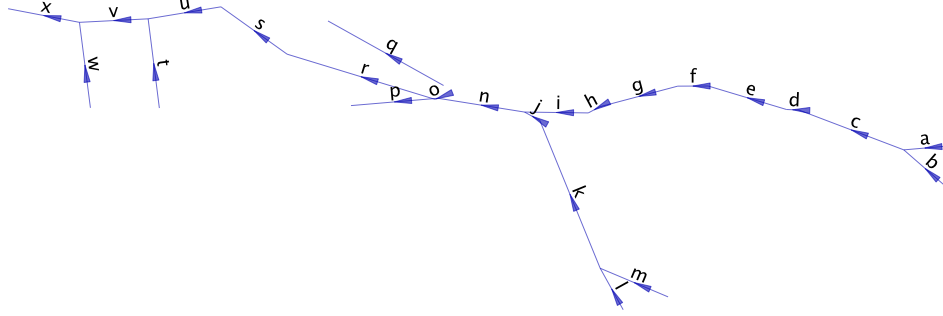


Figure 1: Schematic of zones of monitored river system in the Murray River, Australia, highlighting downstream adjacencies (Koehn et al., 2008). The total length of monitored river is approximately 200km.

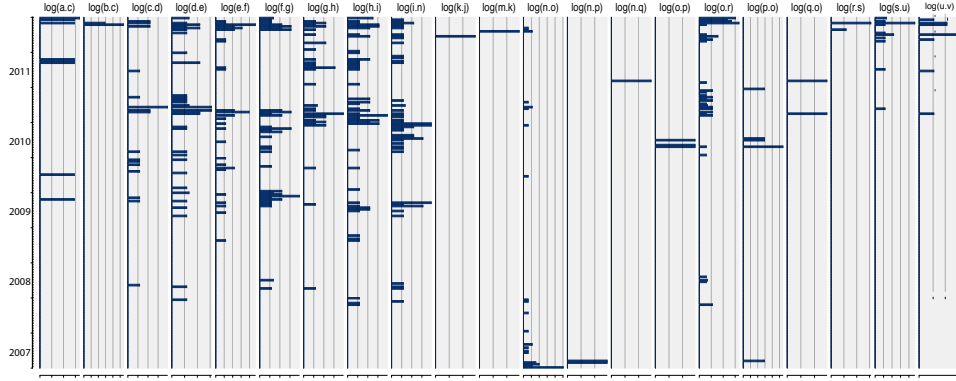


Figure 2: Log of number of fish downstream movements between zones (identified along the columns) over elapsed time.

accuracy ( $F_1$  score of 1.0 indicates no false positives or false negatives, a score of 0.0 indicates only false positives and false negatives). These rules vary between moderately poor performance (e.g., the best rule for zone  $n$ ,  $F_1=0.26$ ) up to remarkably good performance. For example, for zone  $r$  the rule  $1.60 \leq wl \leq 6.75 \Rightarrow \text{after}$  alone can be used to account relatively reliably for movements from  $r$  to  $o$  within 0 to 5 time steps ( $F_1=0.62$ ).

Broadly, these results are encouraging in the context of the limitations that 1. these rules only concern the single most effective rule in a zone (additional rules found might further increase the accuracy); 2. the rules only concern water level (other environmental variables might be necessary to explain many fish movements); 3. the rules do not yet account for uncertainty in the data or causal rules; and 4. the rules only concern conditions of perpetuating processes, and do not yet incorporate events (such as moon phase, flood events, or the end of droughts). Current work is investigating each of these limitations.

Zone	Best rule found	F <sub>1</sub> score
<i>i</i>	$126.41 \leq wl \leq 131.53 \Rightarrow$ moving from <i>i</i> downstream to <i>h</i> after $[0, 5]$	0.45
<i>j</i>	$126.89 \leq wl \leq 126.92 \Rightarrow$ moving from <i>j</i> downstream to <i>k</i> after $[4, 5]$	0.26
<i>n</i>	$124.67 \leq wl \leq 124.75 \Rightarrow$ moving from <i>n</i> downstream to <i>i</i> after $[0, 5]$	0.26
<i>o</i>	$1.60 \leq wl \leq 6.75 \Rightarrow$ moving from <i>o</i> downstream to <i>n</i> after $[0, 5]$	0.46
<i>r</i>	$3.02 \leq wl \leq 6.75 \Rightarrow$ moving from <i>r</i> downstream to <i>o</i> after $[0, 5]$	0.62
<i>s</i>	$2.24 \leq wl \leq 6.40 \Rightarrow$ moving from <i>s</i> downstream to <i>r</i> after $[0, 5]$	0.42
<i>u</i>	$2.33 \leq wl \leq 6.40 \Rightarrow$ moving from <i>u</i> downstream to <i>s</i> after $[0, 5]$	0.58
<i>v</i>	$2.78 \leq wl \leq 6.40 \Rightarrow$ moving from <i>v</i> downstream to <i>u</i> after $[0, 5]$	0.48

Table 1: The best rules discovered for selected upstream movement effects, typical of the wider results.

## Acknowledgments

AG’s research is supported by the UK EPSRC grant EP/M012921/1 “Collectives and Causality”. MD’s research is supported by funding from the Australian Research Council (ARC) under the Discovery Projects Scheme, project DP120100072 “From environmental monitoring to management: Extracting knowledge about environmental events from sensor data.” AB’s research is supported by funding from the ARC under the Discovery Projects Scheme, project DP120103758 “Artificial intelligence meets wireless sensor networks: Filling the gaps between sensors using spatial reasoning.” The authors are also grateful for helpful support and input from the Jarod Lyon and Adrian Kitchingman at the Arthur Rylah Institute (ARI), Melbourne, Australia.

## Biography

Antony Galton is Reader in Computer Science at the University of Exeter. His research focuses on spatial and temporal knowledge representation, with applications to artificial intelligence and GI science, including areas such as collective phenomena, and processes and causation in general.

Alan Both is a PhD student at the Department of Infrastructure Engineering, University of Melbourne, Australia. His PhD topic is “Decentralized computation of qualitative spatial relationships in mobile geosensor networks.”

Matt Duckham is Professor at the Department of Infrastructure Engineering, University of Melbourne, Australia. His research focuses on distributed and robust computation with uncertain spatial and spatiotemporal information.

## References

Bleisch, S., Duckham, M., Galton, A., Laube, P., and Lyon, J. (2014). Mining candidate causal relationships in movement patterns. *International Journal of Geographical Information Science*,

28(2):363–382.

- Bleisch, S., Duckham, M., Lyon, J., and Laube, P. (2013). Identifying candidate causal relationships in fish movement. In *Proc. GISRUK 2013*, Liverpool, UK.
- Buchin, M., Driemel, A., van Kreveld, M., and Sacristán, V. (2011). Segmenting trajectories: A framework and algorithms using spatiotemporal criteria. *Journal of Spatial Information Science*, (3):33–63.
- Galton, A. (2012). Formal ontology in information systems. In Donnelly and Guizzardi, editors, *States, Processes and Events, and the Ontology of Causal Relations*, pages 279–292. IOS Press, Amsterdam.
- Koehn, J., Nicol, S., McKenzie, J., Lieschke, J., Lyon, J., and Pomorin, K. (2008). Spatial ecology of an endangered native Australian Percichthyid fish, the trout cod *Maccullochella macquariensis*. *Endangered Species Research*, 4:219–225.
- Laube, P. (2014). *Computational Movement Analysis*. Springer.
- Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

# Inequality in access to education, and inequality in access to information about allocation of school places

Duke-Williams O, Shepherd E, Eveleigh A

<sup>1</sup>Department of Information Studies, UCL

January 9, 2015

## Summary

The process of allocation of school places generates administrative data that can be used to explore ideas about access to education, and to allow parents / guardians to make a more informed choice when applying for school places for children. A case study explores changing patterns of pseudo catchment areas in the London Borough of Waltham Forest, and illustrates difficulties of assembling some of the relevant data. Similar analysis is carried out for other local authorities, and it is shown that the amount of data available and the ease with which it can be retrieved varies considerably between authorities.

**KEYWORDS:** school allocation, administrative data, active commuting, spatial literacy

## 1. Introduction

The paper explores patterns of equality in access to education (as evidenced by allocation of school places) at two levels. Firstly, some patterns of school allocations and pseudo 'catchment areas' for both primary and secondary schools are explored at a local level for a case study area. This can be discussed in the light of ideas about spatial literacy, and the degree to which visualised (mapped) indicators may be preferable to solely tabular data. Secondly, at a more general level, different areas within England are compared in terms of the amount of information that is made available to parents / guardians of pupils on whose behalf applications for school places are made. Differences exist between local authorities / local educational authorities in terms of the rules used to allocate school places, most pertinently in the case of addressing over-subscription for some schools. On that basis, families in different areas are likely to have different information requirements in order to make an informed decision about the likelihood of a successful application for any particular school. However, it is shown that even where information requirements are broadly similar, the availability of information is variable between authorities.

School place allocation is an administrative operation carried out by local authorities that sets out to achieve a particular requirement, but in turn generates sets of administrative data that can inform research into access to education.

## 2. Data sets

Two groups of data are used: firstly, national (England) level schools performance data are used as a possible proxy for school preference, and secondly, sets of data published by local authorities are used to illustrate results of school applications.

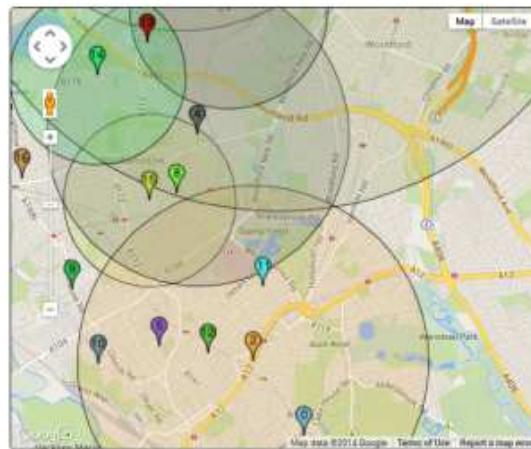
Whilst the use of performance table data as a measure of school quality may be the subject of critique (Goldstein et al, 1996), that is not the focus of this paper. School place allocation is a politically contested area, and one for which a number of alternate optimisations might be suggested; for example: maximising application success (most applicants getting a preferred choice), minimising aggregate distance, with a view to reducing vehicle usage, or allocating so as to maximise actual or potential for active commuting. Patterns of allocation have, for example, been considered in terms of

social segregation (Taylor and Gorard, 2001) and active commuting (Cooper et al 2005).

### 3. Case study area

Historical results are shown for a local authority with which the corresponding author is familiar. The London Borough of Waltham Forest has a relatively standard set of rules applied in order to allocate school places: preference is typically given to pupils with special educational needs, then to siblings of current pupils, and remaining places are then allocated on the basis of the distance between home and school, with preference given to those living nearer to the school. Where a school is over-subscribed, then the furthest relevant distance under which a pupil application was accepted is subsequently published.

Whilst these distance cut-offs are presented as tabular data in published reports, it is also possible to map the data, using circles of a fixed radius: Figure 1 illustrates such a map.



**Figure 1:** School distance cut-offs in Waltham Forest

Whilst there are many caveats about such a map – it misrepresents the actual distribution of students – anecdotal evidence suggests that it is easier to interpret than a simple table of data listing distance values for each school. The map also allows an assessment to be made of the degree to which parents / guardians have a choice of schools: those living in the areas overlapped by multiple circles have a wider apparent choice than those living elsewhere, perhaps only a short distance away. Various forms of overlay can also be used to indicate the extent to which cut-off distances have varied over time; again visualisation of this may allow the data to be easier to interpret.

From a data visualisation perspective, alternative mapping strategies will be discussed.

#### 3.1 Data discovery

A map such as that illustrated in Figure 1 is relatively easy to construct, and for map-users to explore, but it relies on availability of data on which to base the map. A description of the process of locating relevant components of the full data set used to draw the map will be given for this case study area. These data are all implicitly open data: they are published by a local authority, with no restrictions indicated regarding re-use. In order to map data for a single year, data must be manually transcribed from a PDF document, and several different lists of a fixed set of schools must be used to assemble a single canonical set.

In order to extend this into a time series, former publications must be located; in the case of machine readable versions this required the use of Google searches revealing relevant URLs, rather than any bespoke index or archive at the local authority website. This is time consuming, and requires some

level of research skill in identifying and downloading the relevant documents. Not all persons applying for school places will be able to do this, and families are thus make decisions with different amount of information at their disposal.

#### **4. Comparison of information availability between authorities**

In order to compare change over time, two further local authorities were selected: one in Essex, and one in West Yorkshire. An attempt was made to retrieve the data required to construct similar time-series maps. It was discovered that there were considerable differences in the range of data made available to prospective applicants, as well as in the ease with which data could be retrieved. Specifically, whilst 'distance' was used as a criterion for selection, it was not always possible to determine from the published documentation what the cut-off distance had been in previous allocation rounds. Where this is the case, applicants are required to make a decision without all of the relevant information available to them.

Again, the data discovery process depended on web search strategies that in themselves are not particularly difficult, but would not be possible for all potential applicants.

*[Author note: full examples of these will be given, and additional authorities added]*

#### **5. Discussion**

Ideally, applications for school places should be made with applicants having suitable information available to them. Where distance is a key element in the success of an application, it is useful for applicants to know whether any application they make is likely or unlikely to be successful: an application for a school for which they are very unlikely to be eligible on distance grounds might be avoided, along with the emotional investment in such an application by both the parent and child. It is important to couch this in terms of spatial and statistical literacy, and for applicants to understand that distance cut-offs vary from year to year, with a number of factors influencing them, not all of which are easily predictable. At present, information made available to applicants is variable from place to place, and can be demonstrated to be only partially complete. It is argued that data of this sort would benefit from collation at national level, although it is recognised that there are obstacles to this, most notably that different authorities operate different sets of allocation systems, (and that there is within-authority variation), and therefore different data items are relevant in different cases.

#### **6. Acknowledgements**

This work has been supported by the ESRC via grant ES/L007517/1, 'Administrative Data Research Centre, England (ADRCE)'.

#### **7. Biography**

Oliver Duke-Williams is a lecturer in Digital Information Studies, with research interests in the dissemination and analysis of demographic data, specialising in UK Census data.

Elizabeth Shepherd has a chair in archives and records management, and research interests in the development of the UK archive profession, and links between records management and information policy compliance.

Alexandra Eveleigh is a Research Associate for the Administrative Data Centre for England (ADRC-E), working on information governance issues in respect of government administrative data.

#### **References**

Cooper, A. R., Page, A. S., Foster, L. J., & Qahwaji, D. (2003). Commuting to school: are children

- who walk more physically active?. American journal of preventive medicine, 25(4), 273-276.
- Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society. Series A (Statistics in Society), 385-443.
- Taylor, C., & Gorard, S. (2001). The role of residence in school segregation: placing the impact of parental choice in perspective. Environment and Planning A, 33(10), 1829-1852.



# Assessing spatial distribution and variability of destinations in inner-city Sydney from travel diary and smartphone location data

Richard B. Ellison<sup>\*1</sup>, Adrian B. Ellison<sup>†1</sup> and Stephen P. Greaves<sup>‡1</sup>

<sup>1</sup>Institute of Transport and Logistics Studies, The University of Sydney

January 9, 2015

## Summary

Relatively high densities and low car ownership levels in inner Sydney are associated with much lower levels of car use than other parts of Sydney’s Metropolitan Area but it is unknown how this affects the distribution nor the variability in destinations. Following processing of a dataset derived from a seven week travel diary and smartphone app, spatial density analysis is conducted on the destinations by variables including mode, purpose and day of the week. The results show substantial differences in choice of destinations depending on what mode is used and the purpose of the trip.

**KEYWORDS:** Spatial density, transport, destinations, trip purpose, smartphone tracking.

## 1 Introduction

The daily travel of residents in Sydney, Australia has generally been characterised by being largely car-based given the metropolitan area’s relatively low population density (by global standards) and high car ownership levels (Greaves et al., 2014). However, the increasing affluence and density of inner-city suburbs coupled with an increasing push towards public transport and active travel by the City of Sydney council has resulted in a somewhat different mode share for travel in these areas (Bureau of Transport Statistics, 2013). Although this difference in market share has been well documented, it is not clear how this is related to the spatial distribution of trip destinations nor how this varies between individuals depending on what mode they use. With this in mind, this paper uses data on one week of travel of over 600 inner Sydney residents collected using a combination of an online travel diary and a location tracking smartphone app to analyse the spatial variability and distribution of trips with the aim of identifying the relationships between mode choice and destination choice.

---

<sup>\*</sup>richard.ellison@sydney.edu.au

<sup>†</sup>adrian.ellison@sydney.edu.au

<sup>‡</sup>stephen.greaves@sydney.edu.au

## 2 Background and Context

Sydney is a city with a relatively large area given its population and this means that many people travel considerable distances by car. Although the average daily travel distance is approximately 32km, this varies substantially with inner city areas seeing average daily travel distances of approximately 17km (10 miles) and some outer suburbs having average daily travel distances of over 60km (37 miles) (Bureau of Transport Statistics, 2013). This wide discrepancy in the distance of trips is also evident in the choice of modes with trips by car ranging from 28 percent of trips to nearly 90 percent of trips. A large component of this variation between suburbs is associated with different levels of population density and car ownership levels throughout Sydney. In Figure 1 population density and vehicle ownership have been plotted on a map using a two dimensional colour gradient. Bright blue indicates areas with high population density (over and above 50,000/square kilometre) and low car ownership levels (cars per household). Bright red indicates areas with low population density and high car ownership levels. Dark colours indicate low values of both population density and car ownership levels and purple indicates medium values of both variables. As is clear from Figure 1, the higher concentration of areas with high population density and low car ownership is in and around Sydney’s Central Business District (CBD).

The concentration of high density areas near the CBD coupled with an increasing focus on active travel by the City of Sydney council (albeit with reluctant support from the State government) means that the choice of destinations of residents and the modes used to access those destinations may also be changing (Pucher et al., 2010). Although these broad changes are slowly becoming apparent in aggregate statistics produced from census data and Sydney’s continuous (one-day) household travel survey, it is not clear how destinations and modes vary within individuals as well as between them.

## 3 Data and methodology

A recent (and ongoing) multi-wave study designed to determine the effects of new bicycle infrastructure being built in Sydney on bicycle use in the inner city has resulted in the collection of both seven day travel diary data and corresponding location tracking data from smartphones for over 600 inner Sydney residents (Rissel et al., 2013). This dataset provides the opportunity to gain further insight into the travel patterns of residents in inner city areas and allows for the analysis of week long and repeated (over several years) travel data. The web-based travel diary used by respondents included questions common to many travel diary (including trip departure and arrival times, mode and purpose) as well as some intended to provide more detailed information on short and incidental trips, primarily short walking trips to and from public transport or local shops that are often forgotten by people completing travel diaries. In addition to this, some questions were asked to elicit further information about any reported bicycle trips that were largely focused on their use of separated bicycle infrastructure. The smartphone app was designed to complement (rather than replace) the travel diary and as such was designed simply to passively collect location data, primarily through WiFi location, approximately every five seconds and provide participants

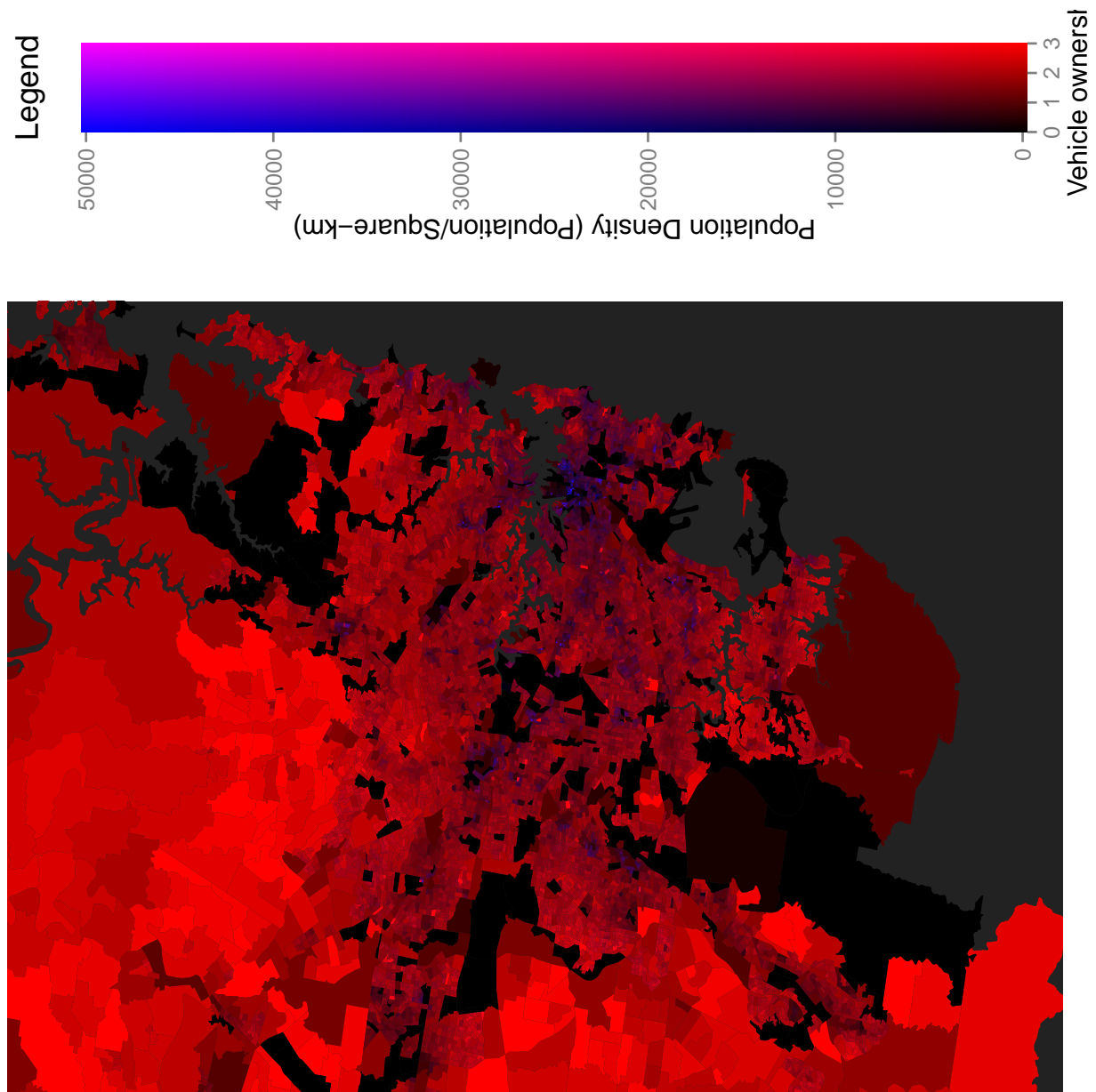


Figure 1: Population Density and Vehicle Ownership in Sydney Metropolitan Area

with the ability to view where they had travelled on a map (Greaves et al., 2014). Although the smartphone app was optional it was used by a significant proportion of participants with a total of approximately 54 million observations (so far).

The geocoded travel diary and smartphone datasets were combined into a single dataset recording all the trip destinations, purposes and mode as entered into the travel diary as well as intermediate stops calculated using the smartphone data. Additional destinations that were recorded by the smartphone tracking app but that respondents failed to record in their travel diary were also included. This combined dataset also included the calculation of other inferred trip variables including if the participant had travelled to work that day, the number of unique destinations in each trip tour<sup>1</sup> and the “main” mode for the trip. Using the combined dataset, a spatial density analysis was performed using several combinations of the trip variables included in the dataset. The spatial density analysis was used to assess the location and density of concentrations of destinations for a variety of different combinations of the trip variables. This was also conducted for different temporal classifications (e.g., weekdays and weekends, and morning, afternoon and evening) as a method of determining if the choice of destinations are associated with non-discretionary travel.

## 4 Results and Discussion

The results of the spatial density analysis on the combinations of trip purpose and travel mode show clear differences between the destinations of trips using each mode and for each purpose. However, there are also some similarities between modes/purposes with different modes having contours of similar shapes but different sizes suggesting a strong influence not only of the available destinations but also of the location of transport corridors and public transport services. Although conclusions from combinations of variables are perhaps most interesting, it is of use to look at mode and purpose separately before looking at the combination of the two variables.

For purpose alone, there are high concentrations of trip destinations for some purposes associated with specific land-use in Sydney (see Figure 2). This is most strongly evident (as can reasonably be expected) in destinations where the purpose is to attend university with destinations concentrated around the three main universities located in the area. Similarly, commuting trips are strongly concentrated in the city’s CBD as well as the University of Sydney<sup>2</sup>.

The spatial distribution of several of the other trip purposes are also reasonably concentrated with most trips having destinations either close to home or in the CBD. Shopping trips appear to be particularly concentrated in a relatively narrow band stretching from the CBD to slightly further West than the suburbs included in this study. In contrast, trip purposes associated with recreational activities including sport, visiting friends and family, and religious activities cover a reasonably wide area.

The analysis of destinations by (main) mode suggest that there is also a difference in the density and distribution depending on the mode. Although this is to be expected to some extent given

---

<sup>1</sup>A series of trips starting and ending at home.

<sup>2</sup>One of the largest employers in inner Sydney is the University of Sydney.

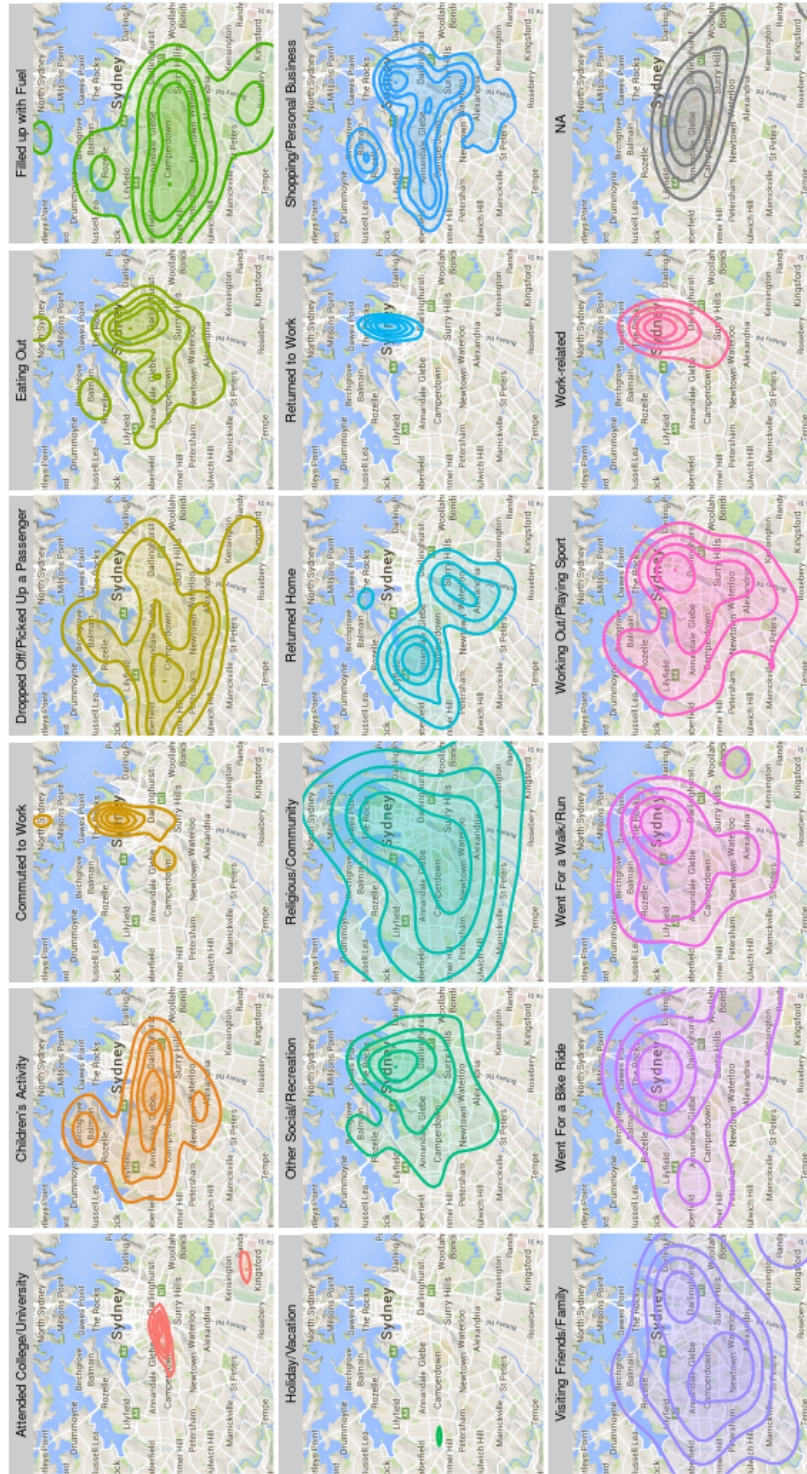


Figure 2: Spatial density plot of trip destinations by trip purpose

constraints limiting the use of some modes of transport to specific corridors (train in particular), the differences are not as large as may have been expected. Furthermore, car trips are not significantly less concentrated than those of other modes. This is likely in part the result of the high availability of public transport in inner Sydney but may also be related to the large number of services available in the area that residents can reach by both cars and other modes.

When mode and purpose are assessed together, the differences between the modes become more apparent. This is particularly true of commuting trips (i.e., work/office destinations) in which the relatively small proportion of participants commuting by car travel to a wider geographic area but is also evident (to a smaller degree) with other purposes. One somewhat surprising result is that despite the flexibility of buses compared to trains, for many purposes destinations of bus trips are just as concentrated in (often similar) patterns to the train. This is despite Sydney's buses covering areas that are not very close to railway stations.

Analysis of some of the other trip variables also showed some differences in the distribution of trips. Although weekend destinations were rather less concentrated than weekday destinations, this varied substantially by purpose. Shopping and eating out were still relatively highly concentrated in similar areas during both weekdays and weekends with visiting friends and family being substantially less concentrated.

## **5 Conclusions**

The analysis of the spatial distribution and variability of destinations by inner city residents of Sydney showed that the choice of destination is very much related to both the mode and purpose but also other trip characteristics (such as the day of the week). Furthermore, the high population density and low vehicle ownership of inner Sydney compared to the rest of the metropolitan areas has a clear relationship to the choice of destinations and the mode used to get there. In contrast to the travel of residents in outer areas of Sydney, destinations of inner Sydney residents is highly concentrated and characterised by repeated visits to several nearby areas for a variety of purposes.

## **6 Biographies**

Richard B. Ellison is a Research Fellow at ITLS. His current research interests include modelling of freight transport and its environmental effects. He is also involved in several projects on cycling as well as broader research on the interaction between transport infrastructure investments and other wider benefits.

Adrian B. Ellison is a Research Fellow at the Institute of Transport and Logistics Studies (ITLS), The University of Sydney. Adrian's main research interests are in road safety, active travel and the use of GPS and smartphones to collect spatially aware data.

Stephen P. Greaves is a Professor of Transport Management at ITLS. Stephen's current research is focused around the health/environmental/safety impacts of transport, active travel including cycling, and innovative travel data collection methods using the latest technologies.

## References

- Bureau of Transport Statistics (2013). 2011/2012 Sydney HTS: Summary transport statistics by Local Government Area. Technical report, Transport for NSW.
- Greaves, S. P., Ellison, A. B., Ellison, R. B., Rance, D., Standen, C., Rissel, C., and Crane, M. (2014). A Web-Based Diary and Companion Smartphone app for Travel/Activity Surveys. In *10th International Conference on Transport Survey Methods*, Leura, Australia.
- Pucher, J., Dill, J., and Handy, S. (2010). Infrastructure, programs, and policies to increase bicycling: an international review. *Preventive medicine*, 50 Suppl 1:S106–25.
- Rissel, C., Greaves, S., Wen, L. M., Capon, A., Crane, M., and Standen, C. (2013). Evaluating the transport, health and economic impacts of new urban cycling infrastructure in Sydney, Australia - protocol paper. *BMC public health*, 13:963.

# Spatiotemporal Identification of Trip Stops from Smartphone Data

Adrian B. Ellison<sup>\*1</sup>, Richard B. Ellison<sup>†1</sup>,  
Asif Ahmed<sup>‡1</sup>, and Stephen P. Greaves<sup>§1</sup>

<sup>1</sup>Institute of Transport and Logistics Studies, The University of Sydney, Australia

09 January, 2015

## Summary

As part of a three-year study on cycling infrastructure, a smartphone app was used to passively collect location information resulting in 54 million observations. These data are then used to identify trip stops using a new method that employs a moving average position. In total 12,849 stops are identified with a median time of one hour and a spatial distribution consistent with the travel diary data collected as part of the same study.

**KEYWORDS:** Transport, Stop Detection, Algorithm, Smartphones, Cycling

## 1. Introduction

Collecting travel data using smartphones is gaining increasing attention due to the ability to (largely) unobtrusively collect data over time and space. These methods facilitate the collection of ever larger datasets providing geographic data at a detailed and highly disaggregate level across time and space. This being the case, these datasets have potential to be used for analyses of travel behaviour that incorporates both the spatial and temporal variation in behaviour. However, as a direct consequence of this these datasets tend to be very large and, therefore, become too labour intensive to prepare and analyse through existing labour-intensive methods designed largely for traditional travel surveys. In many cases, researchers resort to aggregating the data such that many of the spatial characteristics are lost.

This paper describes a method of identifying trip stops using location data collected from smartphones as part of a study on the impact of cycling infrastructure (Rissel et al. 2013) in Sydney, Australia. In total approximately 54 million observations were recorded over two data collection periods in 2013 and 2014. The aim here is two-fold. The first of these is to develop a method that accurately and consistently identifies the locations and durations of stops between and during travel from smartphone data. That is to say, stops that occur at a destination (or activity), intermediate stops and stops while waiting for a train, bus or other transport mode. This enables continuous analysis of travel data on a large scale. The second objective is to use the latitude and longitude of these detected stops as a basis for identifying the spatiotemporal characteristics of these stops by combining the smartphone data with existing geographic data.

### 1.1. Context

---

<sup>\*</sup> adrian.ellison@sydney.edu.au

<sup>†</sup> richard.ellison@sydney.edu.au

<sup>‡</sup> asif.ahmed@sydney.edu.au

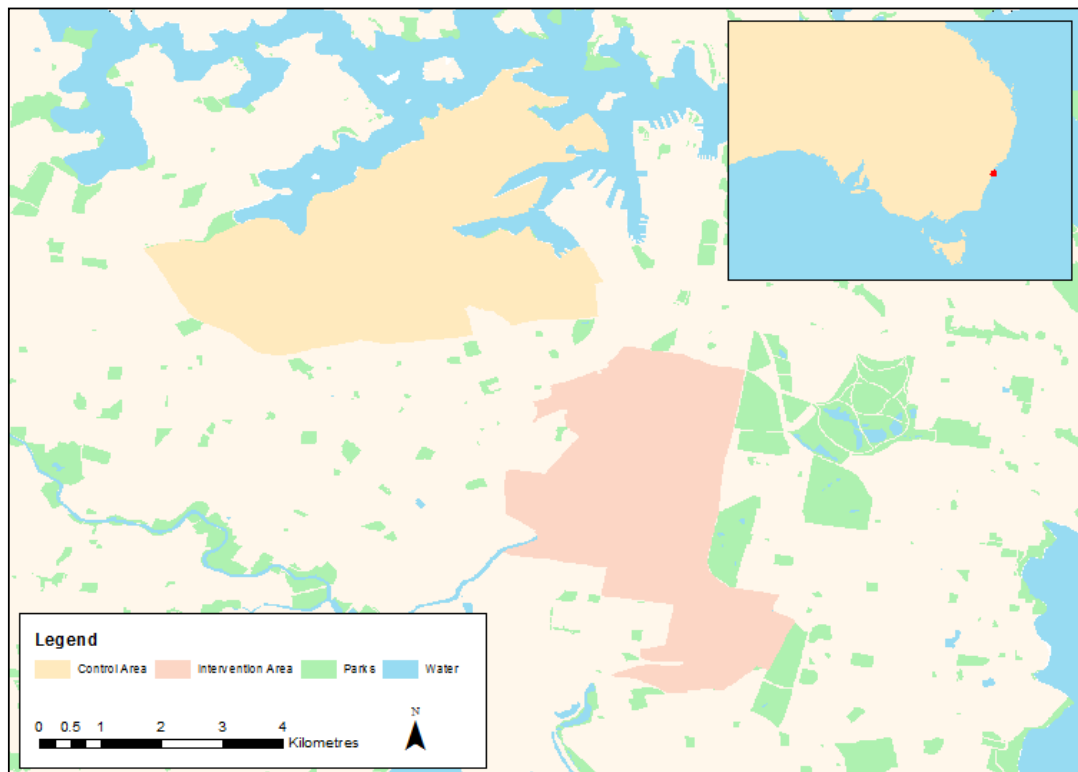
<sup>§</sup> stephen.greaves@sydney.edu.au



While transport research contains a long history of using personal GPS devices to measure travel (for example, Murakami and Wagner 1999; Stopher, FitzGerald, and Zhang 2008) alone or in combination with a travel diary, not all of the methods employed are transferrable or scalable to smartphone data collection. For instance, many studies using GPS employ methods that are dependent on the availability of reliable Doppler speed (Stopher, FitzGerald, and Xu 2007; Doherty, Papinski, and Lee-Gosselin 2006); a measure that is not available in data collected using smartphones without the aid of the (battery-draining) built-in GPS. In any case, even with speed data available many studies rely on manual map-editing to identify any false-positives and false-negatives of which there are many. This also potentially introduces the issue of inconsistency of the analyst or analysts involved. Although this may be practical when there are a finite number of participants carrying these devices for a small number of days (typically one to three) with the potential increase in scale this becomes infeasible. The time lag necessary to accomplish this also reduces the ability for researchers to ask participants for additional information due to the limitations of memory recall.

## 2. Data Collection

The data were collected as part of a three-year study on the impact on cycling of new dedicated bicycle paths. The study combined a travel and health questionnaire, an online seven-day travel diary and a smartphone app to passively record travel at five-second intervals in three data collection periods in the (Australian) spring of 2013 (baseline), 2014 and 2015. Participants were recruited from two inner-city areas shown in Figure 1, an intervention area in which the bicycle infrastructure was being built and the control area in which there was no change in the provision of bicycle infrastructure (Greaves et al. 2014). To simplify management, access and analysis, all data collected during the study is stored in a single relational database that can be queried as necessary on an *ad-hoc* basis or using algorithms such as the one described in this paper.



**Figure 1** Location of Control and Intervention Areas

The entire smartphone dataset contains almost 54 million observations collected from 469 study participants (many of whom over two years) and an additional 68 unidentified participants who likely downloaded the (free) smartphone app despite not being involved in the study. This represents 12,384 person-days of data albeit with an unequal distribution of days by user.

To verify the accuracy of the stop detection method an alternative source of data is needed. In this case, the seven-day travel diary is the best available source of comparison data. As such, this paper excludes smartphone data collected from those without a valid and complete travel diary leaving 401 unique participants with 38 million observations during 9,552 person-days. For the same reason, data collected outside the seven-day diary period was excluded from this particular analysis.

### **3. Stop Identification Method**

The stop identification method was designed to identify the times and locations in which users stopped from their smartphone data alone. Crucially, stops in this case are not intended to delineate trips and therefore stops to change mode or to pick up something or someone are considered to be valid stops. Similarly, a day spent entirely at home would generate one 'stop'. Simultaneously, the algorithm was designed to move away from the rule-of-thumb time duration methods common in GPS trip identification by focusing on distance.

For each eligible participant, the algorithm starts by retrieving the observations in chronological order from one participant and one day at a time. Subsequently, the algorithm loops through each observation with a recorded accuracy of less than 1,000 metres simultaneously using the latitude and longitude to maintain a rolling average position and a total time within a 150 metre radius of the average position. If an observation is within 150 metres of the average position then the average position and the amount of time within the 150 metre radius is recalculated including the observation. If this is not the case (i.e. the observations is located greater than 150 metres away from the rolling average position) then the time between the current and previous observation is removed from the time spent inside the 150 metre radius.

Once the total time spent within the 150 metre rolling average position exceeds five minutes, then a stop is deemed to have occurred at that location. At that point the algorithm continues but reverses the situation in which time spent outside the 150 metre rolling average position is added to a time spent moving and time spent inside the 150 metre rolling average position is subtracted from the same variable until zero is reached. If the time spent outside the rolling average position reaches 50 seconds then the previous stop is considered to have ended.

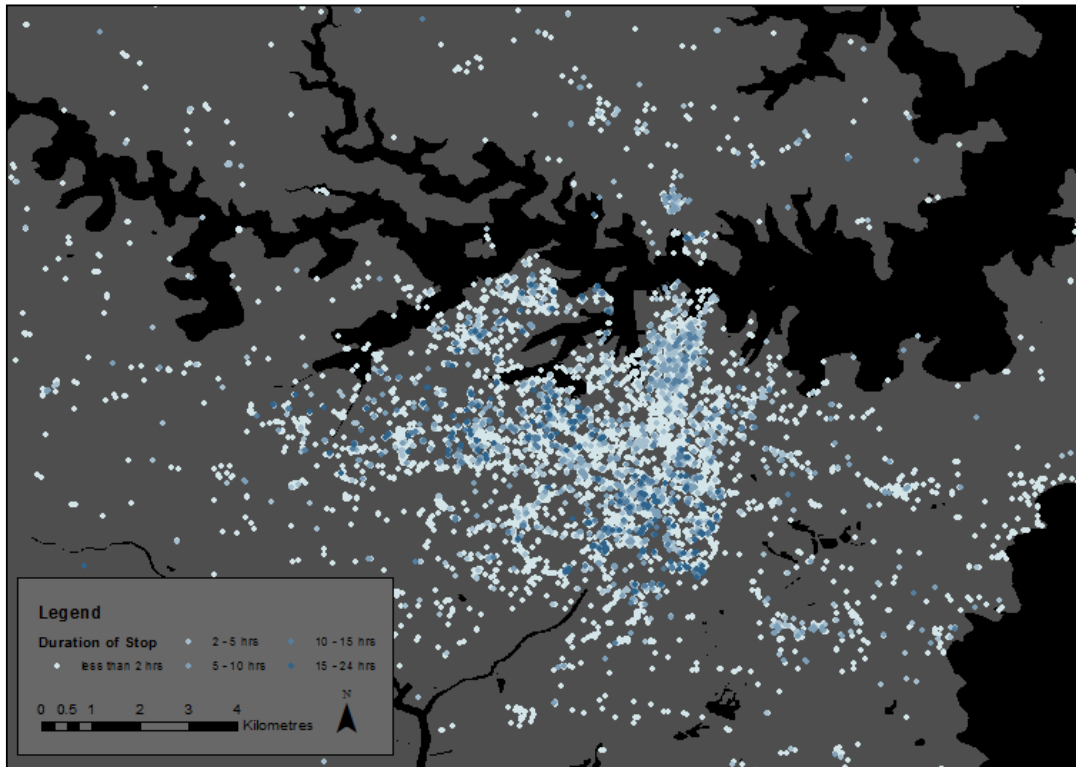
A final step is to loop through each of the detected stops and check that consecutive stops are at least 300 metres apart. This is done to as it was found that when indoors, moving from one side of the building to another could create spurious stops in the data due to (on average) less accurate positions.

This process ensures that the algorithm is not as susceptible to spurious data which would otherwise suggest there is movement where there is none. This is particularly problematic when somebody is located inside as this tends to increase the proportion of spurious locations observed in the dataset. The methodology is also computationally efficient allowing it to be used while data is being collected in addition to as a post-processing tool.

### **4. Identified Stops**

Running the algorithm resulted in 12,849 stops (shown in Figure 2) being detected from 382 participants. This compares to 16,660 trip 'legs' reported in the travel diary by the same participants. The average number of stops per participants was 34, the median was 31 and the highest was 99. Unsurprisingly given the study area, most stops occurred within the control area, the intervention area or the Central Business District (CBD).

In terms of time, the minimum stop duration was 75 seconds as measured from the first observation within a stop to the last observation within the stop. The maximum stop time was 24 hours due to the requirement for at least one stop to be identified each day. As such, where only one stop was detected this was equivalent to staying in the same location for one calendar day. The average and median stop durations were three hours and one hour respectively illustrating the extent to which (in absolute terms) most stops are short. The second highest concentration of stop durations was observed at eight hours of duration corresponding to the time between midnight (i.e. the start of a new day) and 08:00 in the morning and departures to work.



**Figure 2:** Location and Duration of Stops

## 5. Conclusions

This paper describes a new method of identifying stops dynamically from large location datasets collected using smartphones. It examines the way in which clusters of observations within 150 metres of a rolling average position can be used to indicate that a stop has occurred in that location. This can then be compared to other sources of data or combined with geographic data to allow for spatiotemporal analyses of travel (or lack thereof). Most importantly, it permits the analysis of very large datasets in a manageable and scalable manner increasing the potential to extract useful information.

## 6. Biography

Adrian B. Ellison is a Research Fellow at the Institute of Transport and Logistics Studies (ITLS), The University of Sydney. Adrian's main research interests are in road safety, active travel and the use of GPS and smartphones to collect spatially aware data.

Richard B. Ellison is a Research Fellow at ITLS. His current research interests include modelling of freight transport and its environmental effects. He is also involved in several projects on cycling as well as broader research on the interaction between transport infrastructure investments and other wider benefits.

Asif Ahmed is a PhD student at ITLS. His thesis title is Analysis of travel time expenditure and budgets from multi-day multi-year GPS data. This study aims to undertake a completely new exploration of the concept of stable travel-time budgets using multi-day and multi-year data collected by personalised GPS devices.

Stephen P. Greaves is a Professor of Transport Management at ITLS. Stephen's current research is focused around the health/environmental/safety impacts of transport, active travel including cycling, and innovative travel data collection methods using the latest technologies.

## References

- Doherty, Sean T, Dominik Papinski, and Martin Lee-Gosselin. 2006. "An Internet-Based Prompted Recall Diary with Automated GPS Activity-Trip Detection: System Design." In *Annual Meeting of the Transportation Research Board*. Washington D.C.
- Greaves, Stephen P, Adrian B. Ellison, Richard B Ellison, Dean Rance, Chris Standen, Chris Rissel, and Melanie Crane. 2014. "A Web-Based Diary and Companion Smartphone App for Travel/Activity Surveys." In *International Conference on Transport Survey Methods*. Leura, Australia.
- Murakami, E, and D P Wagner. 1999. "Can Using Global Positioning System ( GPS ) Improve Trip Reporting?" *Transportation Research Part C* 7: 149–65.
- Rissel, Chris, Stephen P Greaves, Li Ming Wen, Anthony Capon, Melanie Crane, and Chris Standen. 2013. "Evaluating the Transport, Health and Economic Impacts of New Urban Cycling Infrastructure in Sydney, Australia - Protocol Paper." *BMC Public Health* 13 (January). BMC Public Health: 963–70. doi:10.1186/1471-2458-13-963.
- Stopher, Peter, Camden FitzGerald, and Min Xu. 2007. "Assessing the Accuracy of the Sydney Household Travel Survey with GPS." *Transportation* 34 (6): 723–41. doi:10.1007/s11116-007-9126-8.
- Stopher, Peter, Camden FitzGerald, and Jun Zhang. 2008. "Search for a Global Positioning System Device to Measure Person Travel." *Transportation Research Part C: Emerging Technologies* 16 (3): 350–69. doi:10.1016/j.trc.2007.10.002.

# Exploring new ways of digital engagement: a study on how mobile mapping and applications can contribute to disaster preparedness

Gretchen Fagg, Enrica Verrucci, and Patrick Rickles

Department of Civil, Environmental and Geomatic Engineering University College London

5 June, 2015

## Summary

Disaster can happen at any time, and no community can consider itself completely safe from its direct or indirect impacts. Digital technologies, such as Geographic Information Systems (GIS), are becoming globally pervasive (World Bank, 2014), with smartphones hosting excellent mobile mapping, data collection and information-providing platforms. A report was compiled to investigate web and mobile applications that provide preparedness information and stimulate community empowerment, some using maps as a medium to convey the information. This body of work discusses the purpose, results and implications of this analysis for further work to be undertaken to address the identified research gap.

**KEYWORDS:** GIS, Hazards, Disaster Preparedness, Citizen Science, Web and Mobile Applications

## 1. Introduction

Disaster can happen at any time, and no community can consider itself completely safe from its direct or indirect impacts. Effective preparedness in communities has emerged in the literature as a crucial asset for limiting losses and for ensuring rapid and sustainable recovery (Paton & Johnston, 2001; Paton, 2000). Nonetheless, a number of factors, spanning from emotional and socio-cultural foundations to the lack of information on how to prepare, were found to be very influential on individuals' sense of agency in preparedness (Joffe et al., 2013; Paton et al., 2008; Morrissey and Reser, 2003; Duval & Mulilism, 1999).

Digital technologies, particularly Geographic Information Systems (GIS), are becoming increasingly pervasive both in developed and developing countries (World Bank, 2014), with smartphones hosting excellent mobile mapping, data collection and information-providing platforms. Scientists are debating the potential that the deployment of this new generation of social and web technologies could have in helping the general public to be better informed and actively involved in preparedness (Troy et al., 2008). Self-efficacy, community awareness, sense of agency, and resilience are the recurrent themes of this enquiry.

On the one hand, it is unquestionable that novel web and social technologies provide manifold channels of information about the occurrence, the intensity, and the area of impact of damaging events. The effectiveness of such technologies for preparedness is yet to be proven. Nonetheless, as some studies do support the idea that new web technologies and mobile applications facilitate learning (Corbeil & Valdes-Corbeil, 2007), a growing number of institutions are developing applications to communicate information on disaster preparedness through various web and mobile platforms.

This paper discusses the preliminary work of the Challenging RISK team – an interdisciplinary group of researchers seeking to understand and improve how people prepare for fires and earthquakes using multidisciplinary techniques, including geospatial

technologies and citizen science activities. Part of the project focuses on understanding how GIS and mobile technology can be used to break new ground in disaster communications, by providing actionable information to promote community engagement in disaster preparedness.

## **2. Web and Mobile Applications for Communicating Disaster Preparedness via GIS**

A preliminary investigation was undertaken to examine potential strengths and weaknesses of currently active online and mobile-distributed applications that provide preparedness information and stimulate community empowerment. The investigation aimed at gauging information on the main distributors and on the intended users of preparedness information through web and mobile applications. These main distributors varied between governmental and non-governmental organisations, as well as by jurisdiction (e.g. local, regional, or national). Interactivity and modality of outreach (e.g. newsletter, RSS feed, text alert, mobile maps) were also examined.

The investigation identified and analysed the contents of 97 active websites and 159 web and mobile applications, 82% of which are hosted on websites or Android and Apple mobile platforms. All the resources were intended for the general public and did not target a specific portion of the population. It was also found that there is a strong predominance of earthquake-focused resources (90%), with very few addressing fire (7%) and even fewer dedicated to both hazards (3%). Most of the resources and applications (88%) focused on passively conveying hazard information to users. Though passive information may be helpful, it begs the question of whether this means of preparedness messaging actually translates into successful uptake of information by the users.

Members of the general public tend to frame risk in a personal way (Dransch et al., 2010), and the inclusion of mapping components in web or mobile applications may support the user in framing the hazard risk from a personalised and location-specific perspective. Of the various applications analysed, a number of them used maps to convey locational information, including where earthquakes have occurred within the last 24 hours, where neighbourhood shelters may be, and the ability to contact emergency services and convey accurate location-based information. The displayed information can be extremely important for making life-saving decisions and maps provide an excellent medium for doing so.

Preliminary investigations are promising, but it is believed that shortcomings will need to be addressed in order to improve the effectiveness and likelihood of uptake of mobile mapping tools for community preparedness. Individuals are more capable of contributing and improving their community when they are better prepared and willing to participate in it (Mäkinen, 2006). For web and mobile applications to encourage participation, digital divide and interactivity issues will have to be considered to facilitate holistic, two-way communication between people and organisations to enact change on a local level.

## **3. Conclusions and Further Work**

This work investigated the various digital platforms that were available for conveying important, life-saving information on preparedness, including those that utilise maps in innovative ways. From the study, it can be seen from the various sources that there are a variety of applications across platforms that focus largely on earthquakes, as opposed to fire, and that opportunity for user contribution is lacking. Accessibility, interactivity, and two-way communication should be considered for future tools to facilitate greater participation and a more holistic dialogue between people and the institutions delivering the applications. As part of the ongoing project work, the researchers will build off of the lessons learned from this report to strengthen the case for geospatial technologies to act as a platform to empower people, communities and organisations to positively impact people's preparedness for

disasters, through workshops to inform communities; we will pilot this work in Seattle, Washington (USA), eventually broadening to other countries in order to better understand social-cultural challenges in deployment and uptake of methods to positively impact earthquake and household fire preparedness.

#### **4. Acknowledgements**

The authors of this paper would like to thank EPSRC and the Challenging RISK project for providing the opportunity to investigate this interesting area of research.

#### **5. Biography**

Gretchen Fagg is an MPhil/PhD Candidate in the Extreme Citizen Science (ExCiteS) research group working on the Challenging RISK project, working on community engagement and motivation. She has an extensive background in disaster and emergency management and seeks to apply those skills through implementation of digital technologies to empower individuals.

Patrick Rickles is a Research Associate in the Extreme Citizen Science (ExCiteS) research group working on the Challenging RISK project, where he coordinates interdisciplinary efforts and develops required GIS technologies. His personal research is on how to effectively teach and increase uptake of GIS by researchers on interdisciplinary projects.

Enrica Verrucci is a GIS researcher in Disaster Risk Reduction and Response. She combines a strong technical skill set in Environmental Engineering, Remote Sensing, and GIS with a deep interest for social studies. She currently holds the position of Research Associate in the EPICentre research group at University College London.

#### **References**

- Corbeil, J.R., & Valdes-Corbeil, M.E. (2007). Are you ready for mobile learning? *Educause Quarterly* 30(2), 51.
- Duval, T.S. and Mulilism J.P. (1999). A person-relative-to-event (PrE) approach to negative threat appeals and earthquake preparedness: a field study. *Journal Applied Psychology*, 11, 495-516.
- Dransch, D., Rotzoll, H. and Poser, K. (2010). The contribution of maps to the challenges of risk communication to the public. *International Journal of Digital Earth*, 3(3), pp.292-311.
- Joffe, H., Rossetto, T., Solberg, C., & O'Connor, C. (2013). Social Representations of Earthquakes: A Study of People Living in Three Highly Seismic Areas. *Earthquake Spectra*, 29(2), 367-397.
- Mäkinen, M. (2006). Digital Empowerment as a Process for Enhancing Citizens' Participation. *elea*, 3(3), p.381.
- Morrissey, S.A., Reser, J.P. (2003). Evaluating the effectiveness of psychological preparedness advice in community cyclones preparedness materials. *Australian Journal Emergency Management*, 18(2), 46-61.
- Paton, D. (2000). Emergency Planning: Integrating community development, community resilience and hazard mitigation. *Journal of the American Society of Professional Emergency Managers*, 7, 109-118.

Paton,D., Johnston,D. (2001). Disasters and communities: vulnerability, resilience and preparedness. *Disaster Prevention and Management: An International Journal*, 10(4), 270 – 277

Paton, D., Smith, L, Daly, M, Johnston, D. (2008). Risk perception and volcanic hazard mitigation: Individual and social perspectives. *Journal of Volcanology and Geothermal Research*, 172,179-188.

Troy, D A., Carson, A., Vanderbeek, J., Hutton, A. (2008). Enhancing community-based disaster preparedness with information technology. *Disasters*, 32(1), 149 -165.

World Bank (2014). World Development Indicators. Accessed on Oct 15th 2014 at: <http://data.worldbank.org/indicator/IT.NET.USER.P2/countries?display=map>



# A Comparison of Three Modelling Approaches for the Prediction of House Prices

Yingyu Feng and Kelvyn Jones

School of Geographical Sciences,  
University of Bristol

## Abstract

Multi-Level Models and Artificial Neural Networks are employed to model house prices in this study. The results are also compared with the widely deployed standard Hedonic Price Model. Historical house sale records in the Greater Bristol area during the period of 2001-2013 are used and the results indicate that MLM offers good predictive accuracy with high explanatory power, especially if neighbourhood effects are explored at multiple spatial scales.

**KEYWORDS:** House Prices, Multilevel Modelling, Artificial Neural Networks, Predictive Accuracy

## 1 Introduction

The principal objective of this paper is to present two advanced quantitative approaches, Multi-Level Models (MLM) and Artificial Neural Networks (ANN) in the house price predictions. They are also compared against the baseline, widely-deployed, the standard Hedonic Price Model (HPM) in terms of goodness-of-fit, predictive accuracy and explanatory power. There is no published work to date comparing ANN with MLM and the use of a much larger dataset than previous publications is another important contribution of this study.

The rest of the paper is organized as follows. Section 2 presents the study area, data, scenarios and performance measures for competing models. The results are presented in section 3 with the conclusions in section 4. The conceptual specifications of the three approaches and a review of the previous studies using those approaches are included in the full paper.

## 2 Data and scenario comparison

### 2.1 Setting and Data

This study selects Greater Bristol as the study area. The house prices and property attributes (displayed in Table 1) are from the Land Registry of England and Wales. The neighbourhood characteristics (displayed in Table 2) are abstracted from the 2001 census data and the Neighbourhood Statistics website. The location of each sale is geocoded based on the unit postcode of the property. 2001 Output Areas (OAs) are selected as the lowest level of neighbourhoods, nested in 2001 lower layer super output areas (LSOAs) and then middle layer super output areas (MSOAs). The full dataset is 65,302 house sales, out of which 61,161 sales in 2001-2012 are used for model calibration and the rest 4141 sales in 2013 are used for prediction.

**Table 1 Definition and explanation of house price data and attributes**

Variables	Definition and explanation
Price	Sale price stated on the Transfer deed in thousands (£'000)
Yrmth	The year and the month when the sale was completed as stated on the Transfer deed, expressed in numerical form
Type	“Det” for Detached house “Semi” for Semi-Detached house “Terr” for Terraced houses “Flat” for Flats/Maisonettes
New	"New" for a newly built property "Old" for an established residential building
Duration	Types of legal interests in land: “Free” for freehold, where the legal interest in land is held by the owner of the land “Lease” for leasehold, where the interest in land or property is held by the tenant who lets the property from the landlord
East	The Ordnance Survey postcode grid reference: Easting
Nth	The Ordnance Survey postcode grid reference: Northing
Dist	Euclidean distance to the city centre, Cabot Circus in Bristol

**Table 2 Definition of neighbourhood variables**

Variables	Definition and Explanation	Level
IMD	2004 Index of Multiple Deprivation score (IMD)	LSOA
IMDbar	2004 deprivation score on “Barriers to Housing and Services”, measuring barriers to housing such as affordability and geographical barriers to key local services.	LSOA
IMDenv	2004 Deprivation score in the living environment, comprising the ‘indoors’ living environment which measures the quality of housing and the ‘outdoors’ living environment for air quality and road traffic accidents.	LSOA
IMDcrime	2004 Deprivation Crime Domain Score, which measures the rate of recorded crime for four key dimensions of crime: burglary, theft, criminal damage and violence.	LSOA
Green	Green space area percentage of total land use area	OA
Det_area	Proportion of detached house	OA
Terr_area	Proportion of terrace house	OA
Flat_area	Proportion of flats	OA
Room	Average number of rooms per household, used as proxy of average size of properties	OA
Noheat	Proportion of houses that have no central heating	OA
Unemploy	Proportion of people aged 16-74 who are not in employment, including retired, students aged over 16 years old and other people	OA
Lnincome	Natural log of Experian income at MSOA level in 2004	MSOA
SocRent	Proportion of social rented from council or others	OA
PriRent	Proportion of private rented from council or others	OA
Occupancy	The Occupancy Rating provides a measure of under-occupancy and over-crowding. It relates the actual number of rooms to the number of rooms ‘required’ by the members of the household	OA
Young	Proportion of people aged 17 or under	OA
Old	Proportion of people aged 65 or older	OA
Black	Proportion of black ethnic	OA
Noedu	Proportion of people have no academic or professional qualifications	OA
Degree	Proportion of people have at least First degree or Higher degree	OA

## 2.2 Scenarios for comparison

Three scenarios are designed with different representations of space and place. Scenario 1 use property characteristics only, and three models are specified for each modelling approach, HPM1, MLM1 and ANN1. The dependent variable is the logarithm of house prices expressed in thousands of pounds and the predictor variables include house characteristics with full interactions and the time of the sale as a third order polynomial. In scenario 2, the absolute location of the property location (represented by the Grid references) and the relative location (measured as the distance to the city centre) are included as additional explanatory variables. The models are named HPM2, MLM2 and ANN2. In scenario 3, the Grid references and distance to city centre are replaced by 20 measured neighbourhood characteristics. Another three models are specified in this scenarios, HPM3, MLM3 and ANN3. We have examined the predictors for multicollinearity by calculating variance inflation factors (VIFs) (Belsley et al., 1980) and all VIFs of the predictors are under 10, indicating that there is not major multicollinearity problem.

In addition, neighbourhood delineations are reflected in MLM for all three scenarios, specified as the nested multiple neighbourhood structure (OAs, LSOAs, and MSOAs). Neither ANN nor HPM is capable of including neighbourhood delineation in the model due to the large number of categorical variables required for practical specification

## 2.3 Performance measures

In order to reach a balanced view of a model's performance, we have used  $R^2$  as the goodness-of-fit measure, Mean Absolute Error (MAE) as the accuracy measure. The explanatory power of property and neighbourhood variables in accounting for house price are also evaluated for MLM and ANN.

## 3 Empirical results and discussion

Due to limited space, detailed model results are not included here but available on request. The comparison of goodness-of-fit are presented in Table 3 and the accuracy comparisons are based on the 4141 hold-out samples in 2013 and summarised in Table 4. All performance measures show that MLM is superior to ANN and HPM in each scenario, indicating that the specification of neighbourhood is helpful in house price predictions, even when the locations or neighbourhood characteristics have been included in the model. Once the appropriate hierarchical structure of housing market has been defined in MLM, location and neighbourhood characteristics will only further explain the price variation between neighbourhoods, but will not further improve the predictive accuracy.

**Table 3 Comparisons of Goodness-of-fit**

	Scenario 1			Scenario 2			Scenario 3		
	HPM1	MLM1	ANN1	HPM2	MLM2	ANN2	HPM3	MLM3	ANN3
<b>R<sup>2</sup> (in-sample)</b>	0.39	<b>0.75</b>	0.39	0.43	<b>0.75</b>	0.47	0.68	<b>0.75</b>	0.69
<b>R<sup>2</sup> (hold-out)</b>	0.23	<b>0.75</b>	0.23	0.31	<b>0.75</b>	0.38	0.65	<b>0.74</b>	0.67

**Table 4 Comparisons of prediction accuracies for hold-out samples**

Hold-out sample:	Scenario 1			Scenario 2			Scenario 3		
	HPM1	MLM1	ANN1	HPM2	MLM2	ANN2	HPM3	MLM3	ANN3
MAE(log Price)	0.319	<b>0.178</b>	0.318	0.303	<b>0.178</b>	0.286	0.210	<b>0.178</b>	0.216
MAPE(log Price)	5.89%	<b>3.29%</b>	5.85%	5.59%	<b>3.29%</b>	5.26%	3.89%	<b>3.30%</b>	4.00%

In terms of model explanatory power, we have presented the relative importance of each predictor graphically. Figure 1 shows the predicted house price (after exponentiation) on a common scale. The predictions are for all sixteen types of property, holding everything else at their mean value (Figure 1a). The time effects (Figure 1b) are for the typical property in a typical neighbourhood. The effects of the neighbourhood characteristics (Figure 1 c-f) show the relationship between house price and each variable, holding all other variables constant at their typical value. In each figure, 95% confidence intervals of the in-sample model are also plotted. It can be seen that the average size of the property in OA, property characteristics and the time when it was sold are very important predictors.

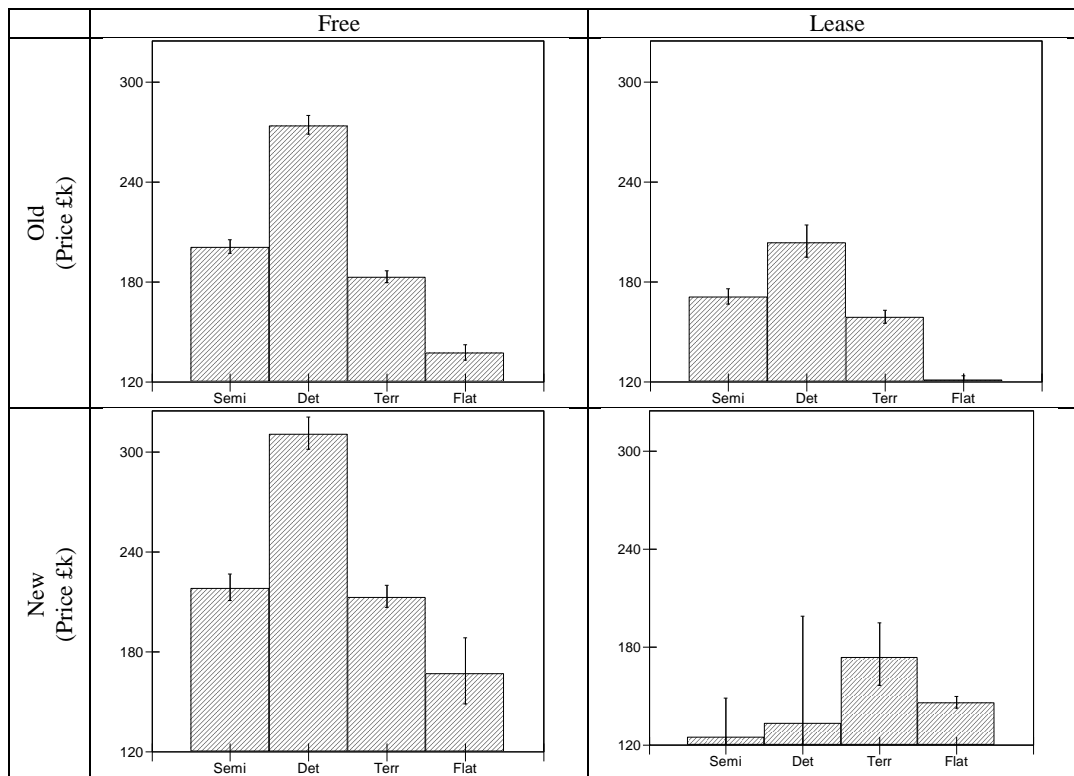


Figure 1 (a) Effect size of predictors

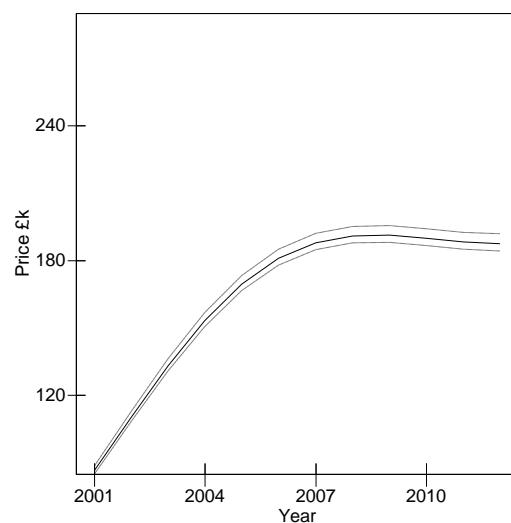


Figure 1 (b)

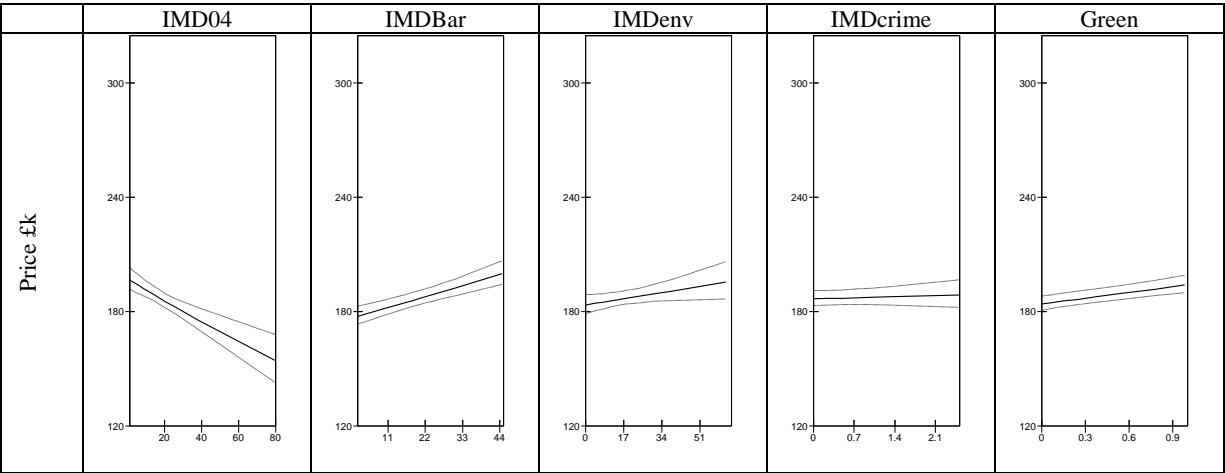


Figure 1 (c)

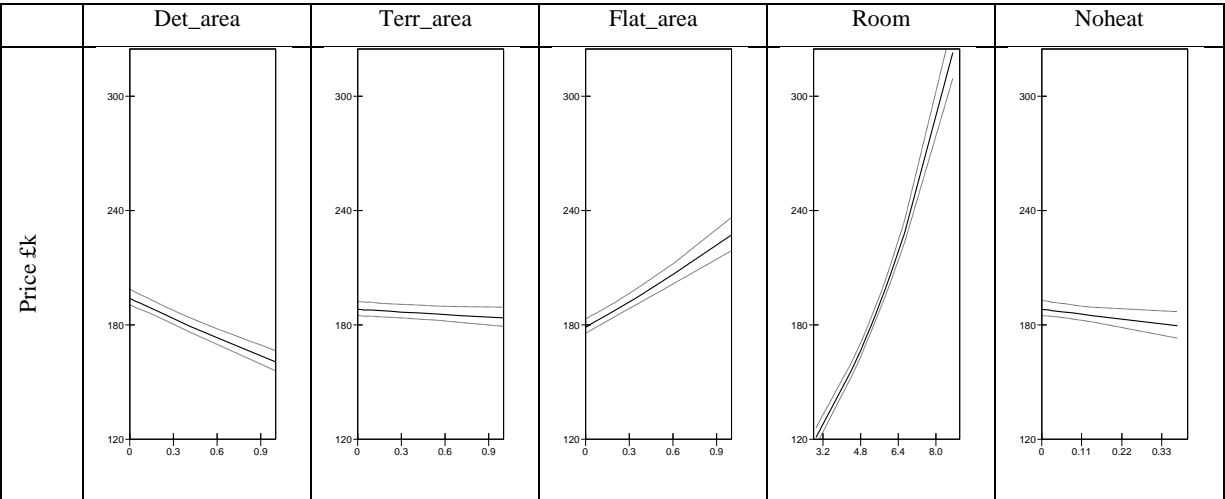


Figure 1 (d)

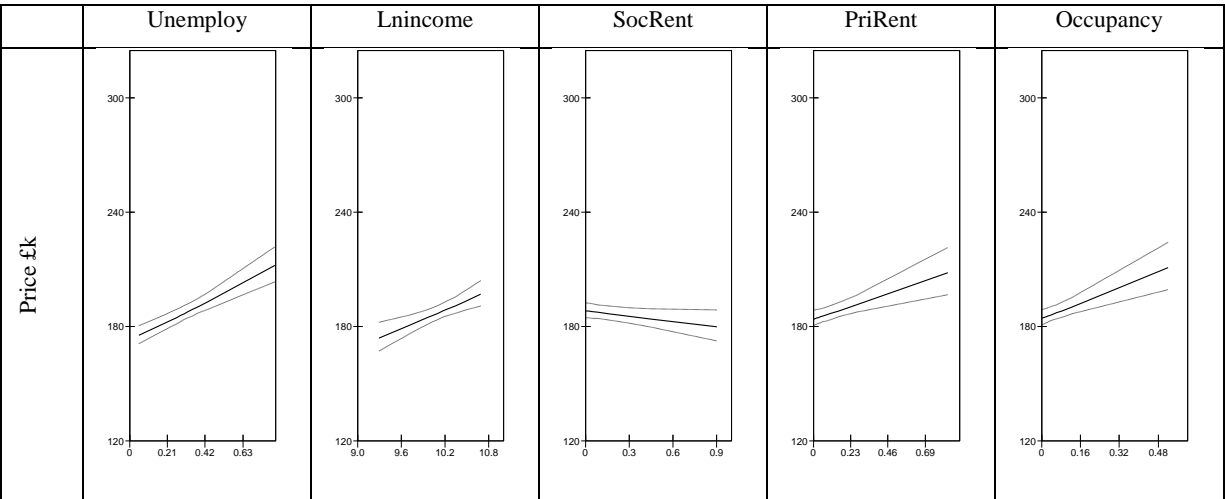


Figure 1 (e)

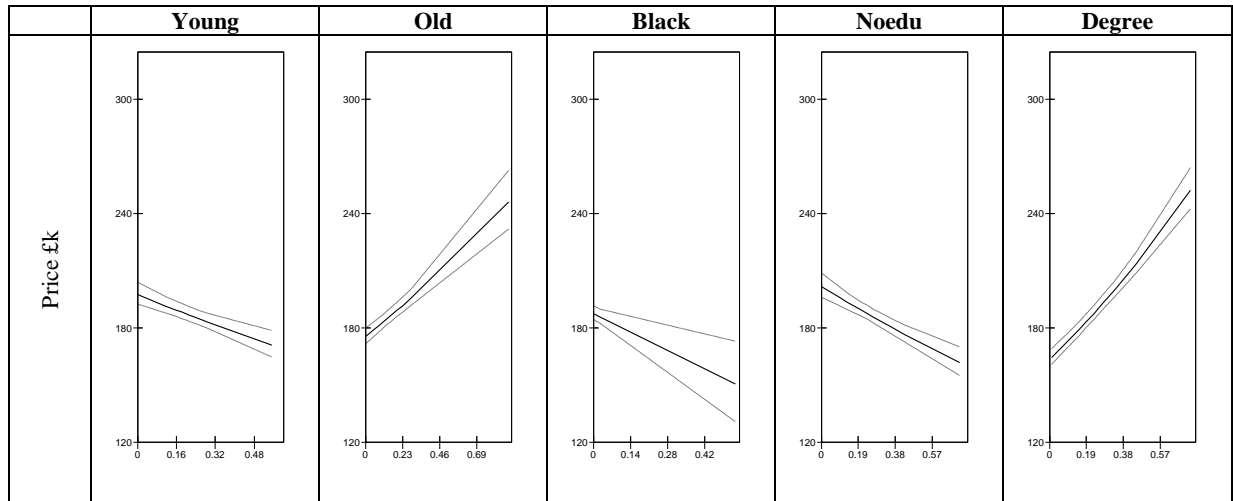


Figure 1 (f)

#### 4 Conclusions

This paper illustrates the use of MLM and ANN approach to modelling housing prices and compares them with the widely accepted HPM approach in terms of goodness-of-fit, predictive accuracy and explanatory power. Neither ANN nor HPM is capable of including neighbourhood delineation in the model due to the large number of categorical variables required for specification, while MLM is able to specify by simply defining them as macro-level units. The results indicate that MLM offers good predictive accuracy with high explanatory power, especially if neighbourhood effects are explored at multiple spatial scales.

#### References

- Belsley, David. A., Edwin. Kuh, and Roy. E. Welsch. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons
- Caudill, M. (1988). Neural Network Primer: Part III, AI Expert, pp53-59.
- Goldstein, H.(1999). *Multilevel Statistical Model*. Arnold.
- Lancaster, K.J. (1966), A New Approach to Consumer Theory, *Journal of Political Economy*, 74, pp. 132-157.
- Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*, Vol. 82, pp. 34-55.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning Representations By Back-Propagating Errors. *Nature*, 323(6088), pp.533–536.
- Snijders, Tom A B and Bosker, Roel J. (1999). *Multi- level Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage

#### Acknowledgements

This work was supported by the Economic and Social Research Council [grant number EDUC.SC3325]. The house price paid data covers the transactions received at Land Registry in the period 1<sup>st</sup> Jan 2001 to 31<sup>st</sup> Dec 2013. © Crown copyright 2013

**Biography**

Yingyu Feng is a PhD candidate at the University of Bristol. Her research interests include spatial modelling, multilevel analysis, GIS technology, neural networks and their applications in spatial analysis.

Kelvyn Jones is Professor of Quantitative Human Geography at the University of Bristol. He is an Academician of the Social Sciences and featured in the top 20 most cited human geographers of the last half century as of 2009. In 2013 he was awarded the Murchison Award of the Royal Geographical Society for 'publications on quantitative geography'.

# Assessing the need for infrastructure adaptation by simulating impacts of extreme weather events on urban transport infrastructure

Alistair Ford<sup>1\*</sup>, Maria Pregnolato<sup>1</sup>, Katie Jenkins<sup>2</sup>, Stuart Barr<sup>1</sup>, and Richard Dawson<sup>1</sup>

<sup>1</sup>School of Civil engineering and Geosciences, Newcastle University, Newcastle upon Tyne

<sup>2</sup>Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK.

November 6, 2014

## Summary

Cities face risks from climate change, placing increased pressure on infrastructure extremes. A methodology to assess the impacts of extreme weather events on urban networks has been developed, using a catastrophe modelling approach to risk assessment by overlaying spatial data, applying hazard thresholds, and testing potential adaptations. Utilising future climate projections, downscaled using stochastic weather generators, future urban temperature and flooding extremes are simulated. These are coupled with spatial urban transport network models and, applying thresholds, disruption to the networks can be simulated. Results for heat and surface water flooding events, and the impacts on the travelling public, are demonstrated.

**KEYWORDS:** climate change, infrastructure, network, flooding, heat, transport, impact

## 1. Introduction

The IPCC 5<sup>th</sup> Assessment Working Group 2 (IPCC, 2014) report on climate impacts highlights the risks faced in urban areas by future climate change, but also that the complex nature of urban areas and their interconnected systems means they cannot be considered in absolute terms but as “system of systems” (Lhomme et al., 2011). In particular, infrastructure in cities will be placed under more pressure in the future due to the changes in climate extremes (e.g. rainfall and temperature) and the concurrent increase in demand from population growth and urbanisation (Hallegatte and Corfee-Morlot, 2011). With the frequency of extreme weather events expected to increase, causing severe damage to buildings and infrastructures (Dawson, 2007), addressing robustness of the urban environment under multiple hazards is pivotal. The Tyndall Centre’s Urban Integrated Assessment Framework (UIAF) was developed (Hall et al, 2010) to allow the assessment of the urban impacts of climate change coincident with other changes which may be seen in cities.

The work presented in this paper highlights a rapid assessment methodology using the UIAF for understanding potential future impacts on the users of urban transport networks from extreme weather events. This begins with climate downscaling using the UKCP Weather Generator and, in the case of extreme rainfall, simulation of surface water flooding using the CityCat model. The spatial footprints of resulting climate hazards are then overlaid on the urban transport networks and thresholds applied to understand where impacts will be felt. These impacts can then be assessed in terms of increased travel time for the users of the transport infrastructure and the total cost of disruption calculated. This methodology is demonstrated in this paper for both extreme heat and extreme rainfall events, on public transport and road networks in the UK.

---

\* [a.c.ford@ncl.ac.uk](mailto:a.c.ford@ncl.ac.uk)

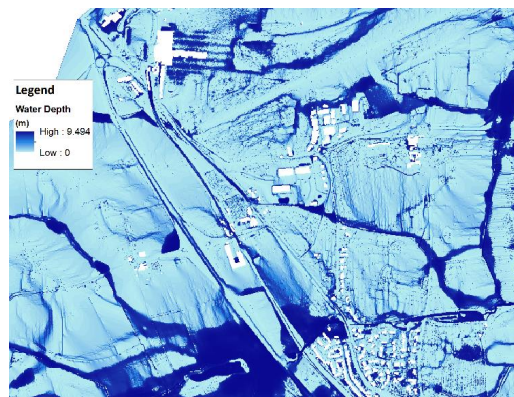


## 2. Method

### 2.1. Hazard modelling

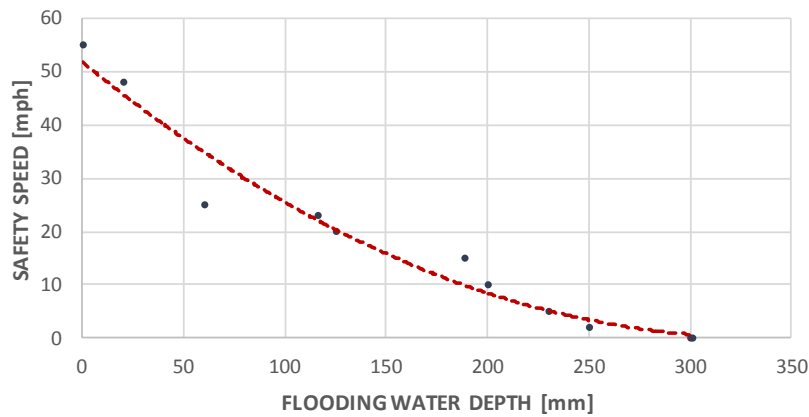
The initial step of the risk-based approach is to understand the hazards to which the system may be exposed. An Urban Weather Generator (UWG) has been produced to supplement the UKCP09 outputs and provide hourly time series of weather variables, such as rainfall or maximum temperature, for future climate scenarios at 5km resolution. The UWG uses a stochastic rainfall model coupled to change factors using probabilistic projections from UKCP09 (Jones et al., 2009). Recent advances in the UWG give an improved reproduction of extreme temperatures, spatial correlations in weather (Kilsby et al., 2011), and urban heat island effects (McCarthy et al., 2012).

The outputs from the UWG are used to assess the spatial and temporal variation of hazards in the urban area. A thresholding approach is applied with impacts assessed when the climate inputs exceed a certain level of severity. For extreme heat events, temperature thresholds are defined above which it is expected disruption will begin to be felt on transport networks. For extreme rainfall events, a further intermediate step is needed to translate heavy rain into flood extents using City Catchment Analysis Tool (CityCAT) developed by Newcastle University (Glenis et al., 2013), the thresholding being applied to the resultant water depths, an example output shown in Figure 1.



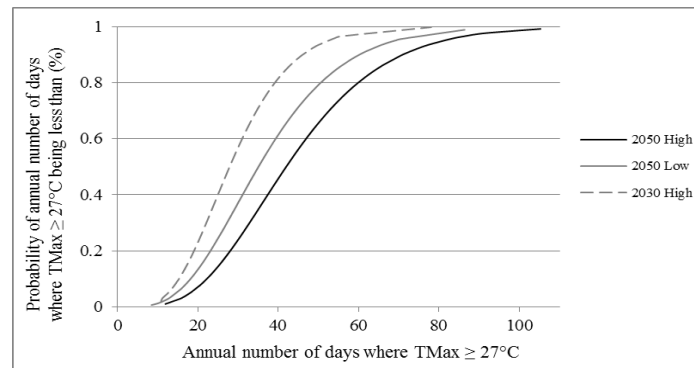
**Figure 1** An example output from CityCat showing the depth of water (hazard map) across an urban area during an extreme rainfall event at 1m resolution (2 hours of duration, 200 ys return period).

To calculate the disruptive effect of flooding on transport networks, a function that relates water depth to safe driving car speed has been developed from combining data from experimental reports (Morris et al., 2011), safety literature (Great Britain Department for Transport, 1999), analysis of videos of cars driving through floodwater and expert judgement.



**Figure 2** Representation of the safety driving speed as a function of the flooding water depth.

Heat impacts are considered on the railway network. Dobney et al., (2009, 2010) showed that disruption to railways in London and the South East can occur when temperature exceeds 27°C, based on analysis historic rail buckling events in the Network Rail Alteration database and the corresponding observed temperature. The frequency with which 27°C is exceeded in UWG outputs for a given climate scenario is assessed. Figure 3 shows the probability of the annual number of days where the maximum temperature in one or more grid cell in the study area exceeds the 27°C threshold for a range of time-periods and climate scenarios.



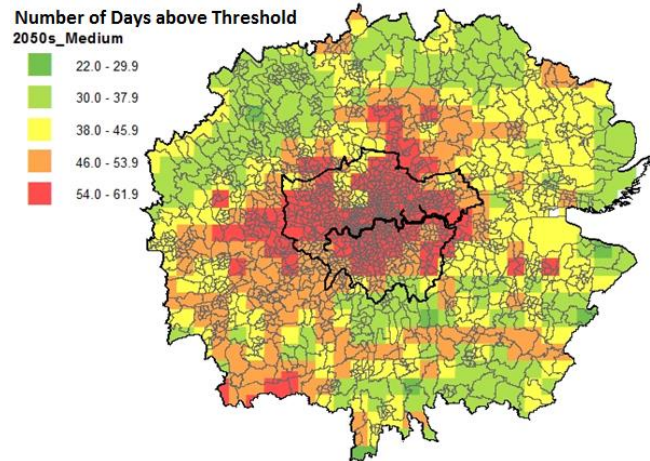
**Figure 3** Probability of the annual number of days where TMax exceeds 27°C for one or more grid cells in the study area, for the 2030s and 2050s under low and high emission scenarios

To analyse the impact of these events on the transport network, further temperature thresholds are related to speed restrictions imposed on railway lines (see Table 2). Single days in the UWG outputs are identified at least one grid cell exceeds one of these thresholds and then the maximum daily temperatures for each of these events can be mapped spatially across the study area on the 5km grid. Figure 4 shows the number of times each grid cell in the London study area exceeds the 27°C temperature threshold.

**Table 2** Temperature thresholds where speed restrictions are imposed.

Threshold	Speed restriction
<27°C	None
<b>Poor</b> Rail Track $\geq 27^{\circ}\text{C} < 28^{\circ}\text{C}$	30mph
<b>Poor</b> Rail Track $\geq 28^{\circ}\text{C}$	20mph

<b>Moderate</b> Rail Track $\geq 33^{\circ}\text{C}$ $<35^{\circ}\text{C}$	60mph
<b>Moderate</b> Rail Track $\geq 35^{\circ}\text{C}$	20mph
<b>Good</b> Rail Track $\geq 36^{\circ}\text{C}$	90mph
<b>Good</b> Rail Track $\geq 42.6^{\circ}\text{C}$	60mph



**Figure 4** Number of days in a given 100-year simulation from the UWG for Greater London where the maximum daily temperature exceeds the  $27^{\circ}\text{C}$  threshold.

## 2.2. Exposure modelling

As the study is focused disruption to commuter journeys, a simple model of network trips was developed using ArcGIS. The model a Frank-Wolf-style trip assignment algorithm (Dafermos and Sparrow, 1968) to load journey-to-work (JTW) observations from the 2001 UK census onto network representations. Networks were constructed from publicly-available data sources (e.g. Ordnance Survey ITN and Meridian data, UK NAPTAN data for public transport stops) supplemented with speed and capacity information. These extra elements allow the calculation of shortest routes in terms of time between origin and destination locations for the JTW observations. Using this model, the appraisal of both the relative importance of network segments (in relation to their levels of use) and the number of users considered in the disrupted network could be achieved.

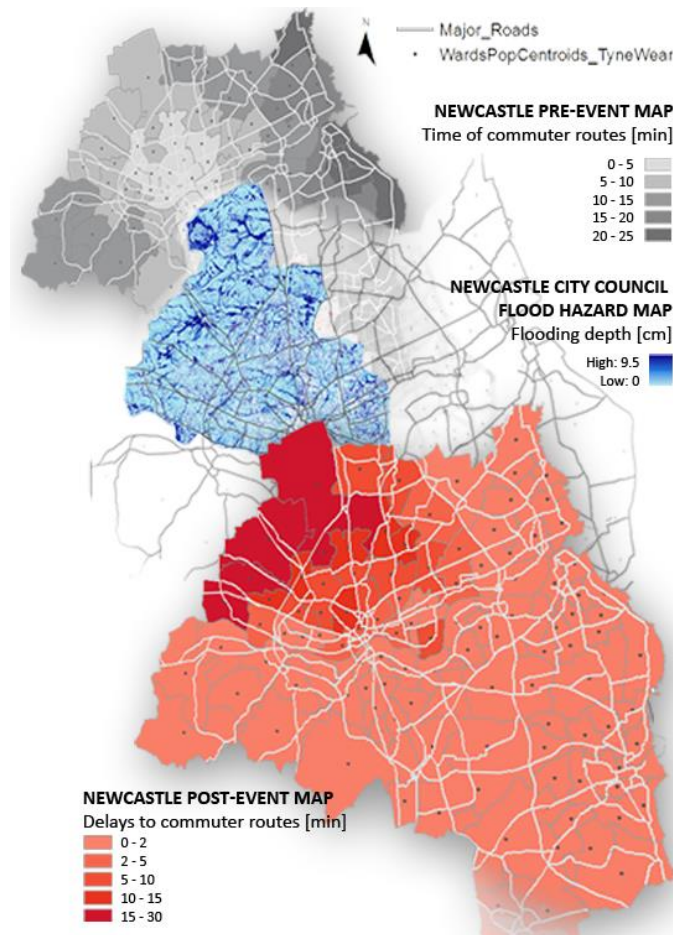
## 2.3. Risk Modelling

Spatial footprints of hazards, either from heat or flooding, are overlaid on transportation networks in GIS to calculate of disruption. Thus, the spatial footprint of the hazard (i.e. a 5km grid cell in which the defined temperature threshold has been exceeded, or a 1m grid cell in water depth has exceeded the a given value from Figure 2) is overlaid on the transport network, the travel speeds on the disrupted network segments adjusted, and new travel times between sets of origins and destinations calculated. By comparing the perturbed travel times with travel times before disruption, the impact on commuting journeys can be estimated, and since the number of journeys using that route is known from the JTW table, the total impact in terms of Person-Minutes can be computed.

## 3. Results

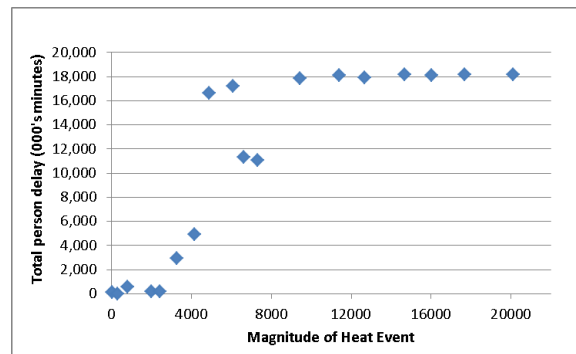
For flooding, analysis has been conducted on the Newcastle road network based on the comparison between the pre-event and post-event travel time maps (see Figure 5). In the pre-event map, the free-flow car speeds are assumed, whilst in the post-event map travel speeds are adjusted according to the

water depth-speed curve and new travel times generated. By comparing pre- and post-event maps, delays caused by flooding to commuter routes can be estimated in terms of delay minutes and economic metrics, based on the generalised cost of travel. Adaptation options can then be assessed, through adjustments of land use and building characteristics in CityCat, and comparison between scenarios determines the cost-effectiveness of the solution considered.



**Figure 5** The results maps from the analysis in Tyne and Wear

For heat disruption, 18 daily events in Greater London were produced by sampling across a range of weather generator simulations. For each of these events, a map of daily maximum temperature was produced and overlaid on the railway network as described above, and the impact on the railway network calculated in terms of speed restrictions and thus increased travel times. Figure 6 shows the relationship between the magnitude of each sampled event and the total person delay in minutes which results from the disruption to the network. In this example, it is assumed that all track in the simulation is of Poor quality (see Table 2). In order to represent simple adaptation to future temperature changes, similar simulations were also run with assumptions that all track was Moderate or Good.



**Figure 6:** Total delays for each of the 18 events of different magnitudes

#### 4. Conclusion

This paper has presented a methodology to investigate the impacts of extreme weather events on urban environment, in particular infrastructure networks, through a combination of climate simulations and spatial representations. By overlaying spatial data on hazard thresholds and transport networks, disruptions to commuting journeys are evaluated. This approach can be applied to the present conditions as well as future uncertain scenarios, allowing the examination of the impacts alongside socio-economic and climate changes.

#### 5. Acknowledgments

This research was funded by the Engineering & Physical Sciences Research Council ARCADIA: ARCADIA: Adaptation and Resilience in Cities: Analysis and Decision making using Integrated Assessment project (Grant No. EP/G061254/1). Richard Dawson is funded by an EPSRC fellowship (EP/H003630/1). Maria Pregolato is supported by an EPSRC funded Doctoral Training Account studentship. All maps contain Ordnance Survey data © Crown copyright and database right 2014.

#### 6. Reference

- Ayyub B M (2014). Natural Hazards in a Changing Climate: Impacts, Adaptation and Risk Management. *Vulnerability, Uncertainty and Risk*, 1-12
- Dafermos S and Sparrow F (1968). The Traffic Assignment Problem for a General Network. *Journal of Research of the National Bureau of Standards*, 73B, 91-118
- Dawson R J (2007). Re-engineering cities: a framework for adaptation to global change. *Philosophical Transactions of the Royal Society*, 365(1861), 3085-3098
- Dobney, K., Baker, C. J., Quinn, A. D. & Chapman, L. (2009). Quantifying the effects of high summer temperatures due to climate change on buckling and rail related delays in south-east United Kingdom. *Meteorological Applications*, 16, 245-251.
- Dobney, K., Baker, C. J., Chapman, L. & Quinn, A. D. (2010). The future cost to the United Kingdom's railway network of heat-related delays and buckles caused by the predicted increase in high summer temperatures owing to climate change. Proceedings of the Institution of Mechanical Engineers Part F, *Journal of Rail and Rapid Transit*, 224, 25-34.
- Doll C, Trinks C, Sedlacek N, Pelikan V, Comes T and Schultmann F (2014). Adapting rail and road networks to weather extremes: case studies for southern Germany and Austria. *Natural Hazards*, 72(1), 63-85

- Glenis V, McGough A, Kutija V, Kilsby C and Woodman S (2013). Flood modelling for cities using Cloud computing. *Journal of Cloud Computing*, 2(1), 1-14
- Great Britain Department for Transport (1999). Surface Drainage of Wide Carriageways. *Design manual for roads and bridges*, Part 1 TA 80/99, HMSO
- Hall JW, Dawson RJ, Barr SL, Batty M, Bristow AL, Carney S, Dagoumas A, Ford A, Tight MR, Walsh CL, Watters H, Zanni AM, (2010). City-scale integrated assessment of climate impacts, adaptation and mitigation. In: Bose, R.K, ed. *Energy Efficient Cities: Assessment Tools and Benchmarking Practices*. Washington, DC, USA: World Bank, pp.43-64.
- Hallegatte S (2009). Strategies to adapt to an uncertain climate change. *Global Environmental Change*, 19(2), 240-247
- Hallegatte, S. & Corfee-Morlot, J. (2011). Understanding climate change impacts, vulnerability and adaptation at city scale: an introduction. *Climatic Change*, 104, 1-12.
- IPCC WGII (2014). Climate Change 2014: Impacts, Adaptation, and Vulnerability. [Online]. Available at: <http://ipcc-wg2.gov/AR5/>
- Jenkins, K., Glenis, V., Ford, A., Hall, J.(2012) "A Probabilistic Risk-Based Approach to Addressing Impacts of Climate Change on Cities: The Tyndall Centre's Urban Integrated Assessment Framework". *UGEC Viewpoints, Connecting Past and Present Lessons in Urbanization and the Environment*, 8.
- Jones P D, Kilsby C G, Harpham C, Glenis V and Burton A (2009). UK Climate Projections science report: Projections of future daily climate for the UK from the Weather Generator. *UK Climate Projections science report*, University of Newcastle, UK
- Kilsby, C., Jones, P., Harpham, C., Glenis, V. & Burton, A. (2011). Spatial Urban weather Generator for Future Climates. ARCADIA Task 3 Report. Available at: <http://www.arcc-cn.org.uk/wp-content/pdfs/ARCADIA-7.pdf>.
- Kirshen, P., Caputo, L., Vogel, R., Mathisen, P., Rosner, A. and Renaud T (2014). Adapting Urban Infrastructure to Climate Change: A Drainage Case Study. *Journal of Water Resources Planning and Management*, 04014064, 1-11
- Lhomme S, Serre D, Diab Y and Laganier R (2011). A methodology to produce interdependent networks disturbance scenarios. *Vulnerability, Uncertainties and Risk*, 724-731
- Love G, Soares A and Püempel H (2010). Climate Change, Climate Variability and Transportation. *Procedia Environmental Sciences* 1(0), 130-145
- McCarthy, M. P., Harpham, C., Goodess, C. M. & Jones, P. D. (2012). Simulating climate change in UK cities using a regional climate model, HadRM3. *International Journal of Climatology*, 32, 1875-1888
- Morris B, Boddington K, Notley S and Rees T (2011). External factors affecting motorway capacity. *TRL Report*

## 7. Biography

Alistair Ford is a research associate in the Centre for Earth Systems Engineering Research and School of Civil Engineering and Geosciences at Newcastle University. His work involves developing models

of climate impacts on urban areas, assessing future climate change and socio-economic change concurrently. He was worked on the Tyndall Centre Cities programme, the EPSRC ARCADIA project, and is currently working on the EC Framework 7 RAMSES project.

After a Master Degree in Building-Engineering Architecture at the University of Pavia (Italy) and at Tongji University of Shanghai (China), Maria Pregnolato started her PhD at Newcastle University (UK). Her research consists of advancing multi-hazard modelling and decision-support study to shape the management of risk and associated uncertainties of urban systems.

Professor Richard Dawson is Director of the Centre for Earth Systems Engineering Research (CESER) at Newcastle University.

Dr Stuart Barr is Senior Lecturer in Geographic Information Science at Newcastle University.

# Calculating the Overbuilding Potential of Municipal Buildings in London

Joanna Foster<sup>1</sup>, Claire Ellul<sup>1</sup>, Philippa Wood<sup>2</sup>

<sup>1</sup>Department of Civil, Environmental & Geomatic Engineering, University College London (UCL), Gower Street, LONDON WC1E 6BT, UK

Tel. +44 (0) 20 7679 4118 Fax +44 (0) 20 7380 0453

<sup>2</sup>WSP Group, WSP House, 70 Chancery Lane, London, WC2A 1AF70

Tel. 0044 (0) 020 7314 4642, Philippa.Wood@WSPGroup.com, [www.wspgroup.com](http://www.wspgroup.com)

KEYWORDS: housing, solution housing crisis, overbuilding, municipal buildings, decision support tool

## 1. Introduction

The world population is set to increase significantly, with the urban population expected to rise by 72% in the next 40 years from 3.6 billion in 2011 to 6.3 billion by 2050 (Nations, 2012). It is therefore necessary to take into consideration the continuing demands that the growing population is going to place on the city's infrastructure, and make preparations now for the future.

### 1.1 UK and England

This study has been carried out with WSP | Parsons Brinckerhoff, one of the world's leading professional services firms in the built and natural environment. According to a report by KPMG and Shelter, a total of 250,000 new homes are needed annually in England to meet rising demand; in 2013 only 110,000 were built (KPMG and Shelter, 2014). The housing shortage is not limited to a single area, however due to the migration to London and the South East, the demand for housing here has increased. This shortfall in house building has had a major impact on house prices. Excluding London and the South East, UK house prices increased by 6.4% in the 12 months to May 2014. This is compared to a rise of 20.1% in London, 9.6% in the South East and 8.6% in the East during the same period (ONS, 2014a). These figures of house price inflation are much higher than the European average of 1.1%.

It is clear that there are major challenges in providing affordable housing, transportation infrastructure and employment for London's growing population (WSP, 2013). Recent projections by the Office of National Statistics (ONS) highlight that the population of London is to grow by 13% in the next ten years by 2022, with the latest estimates indicating



that by 2031 the population may exceed 10 million (ONS, 2014b, NLA, 2014). However, there is currently a substantial discrepancy between the increasing size of the population in London and the volume of housing available. It comes as little surprise then that one of the key messages taken from the Mayor's *2020 Vision* is tackling 'perhaps the gravest crisis the city faces – the shortage of housing' (Johnson, 2013). According to the 2011 Census, an estimated 40,000 new homes are needed every year to support the growing population, a figure that is failing to be generated (Hill, 2013). However, this figure is continually inflating, with new figures suggesting that more like 50,000 new homes are needed (Savills, 2014). Furthermore, even with the projected average of just under 35,000 homes are planned to be built every year for the next 5 years, there would still be a shortfall of 15,000 residential units.

## 2. Methodology

This study focuses on the concept of overbuilding, and calculating the potential solution this may provide to the chronic housing shortage currently faced throughout the UK. This concept is based on the multi-use developments that have been built across the world, including the Beekman Building, New York and The Plimsoll Building, London. This looks at taking existing public sector buildings and redeveloping them with public sector facilities on the lower floors and residential units above. This would involve public and private sectors working collaboratively together to provide the residential units that are needed and to improve existing public sector infrastructure.



Figure 1 WSP's Artist's Impression of Overbuilding NHS Hospital (WSP, 2014a)

## **2.1 Planning Permissions**

The Department of Communities and Local Government currently is in charge of implementing ‘the planning guidance within which the plan for Greater London is developed’ (Fainstein, 2008). The Mayor of London is then responsible for development under the guidance of The London Plan. The London Plan is the ‘overall strategic plan for London, and sets out a fully integrated economic, environmental, transport and social infrastructure framework for the development of London until 2031. The plan has the power to determine the shape of the London skyline, by constricting tall buildings to areas that are deemed good opportunities and that has the required transportation infrastructure needed to support the higher development density (GLA, 2012).

### **2.1.1 Height Restrictions**

London has multiple guidelines in place that are specific to tall buildings, including the English Heritage and CABI’s Guidance on Tall Buildings. The purpose of this document is to enable informed decisions to be made in evaluating planning applications for tall buildings, and is endorsed by the government (Heritage and CABI, 2007).

Building height restrictions within London are determined on a borough basis. Whilst some areas welcome the development, others are deemed more sensitive to them, and other areas are considered to be inappropriate. In addition to this, one of the main concerns within London and the height of new developments are adhering to the protected vistas.

### **2.1.2 Protected Vistas**

Outlined in the revised 2012 London View Management Framework (LVMF), there are currently 27 protected views of London’s landmarks from different vantage points across the city. In addition to this, 13 viewing corridors are protected and require additional consultation when new plans and designs are being considered within these areas, Figure 2. Ultimately, this framework protects the views of both St Paul’s Cathedral and the Palace of Westminster as seen from London’s larger parks.

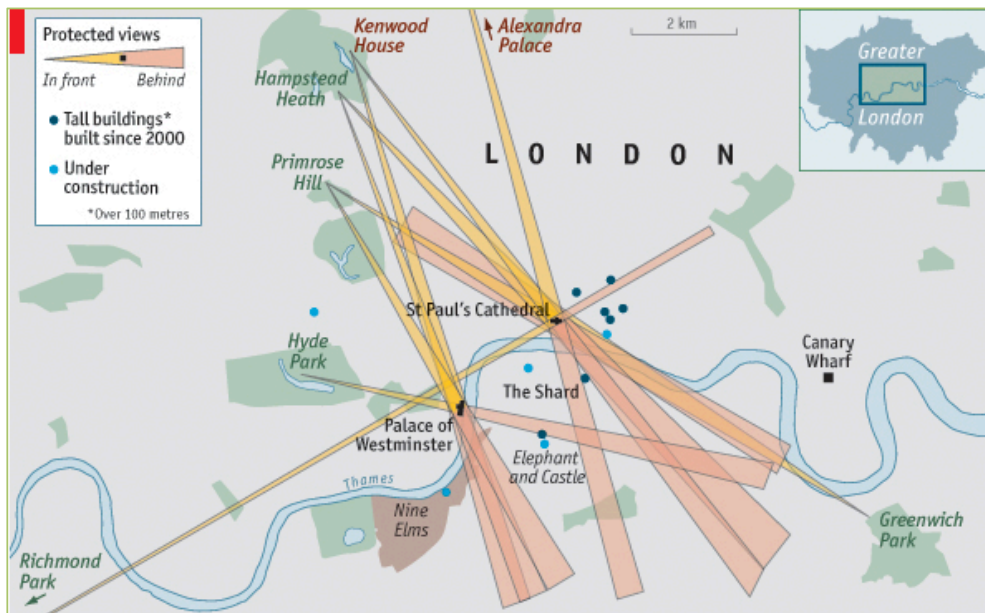


Figure 2 – London's Protected Views (The Economist, 2014)

## 2.2 Data

Property data was obtained in the form of point data from a number of different sources including 'data.gov.uk', 'data.london.gov.uk' as well as local authority websites. An analysis of the availability of land and property data for London highlighted that there was a varied level of data available regarding public sector property for each of the boroughs. It was clear that of the 33 boroughs, Lambeth provided a large volume of suitable data, and as such this borough formed the case study area for the rest of the study. This data availability was summarised within Table 1, which indicates the data available directly from Local Authorities. Data regarding GP practices, leisure centres, hospitals, and schools were available nationwide and therefore for all boroughs, as such these have not been included within the matrix. This was further mapped by WSP, Figure 3.

In addition to the point of interest data, a base map was needed of building footprints. It was decided that OS MasterMap (OSMM) would be used, which provides a good level of accuracy of building footprint areas needed to complete the area calculations. Once the data had been obtained from the relevant sources, a spatial join (intersect) was completed within ArcMap. This joined the point of interest point data to the polygon building data of OSMM.

Table 1 – Data Matrix showing data available from Local Authorities (Accurate as of 20<sup>th</sup> August 2014)

Borough	Libraries	Police Stations	Fire Stations	Council Buildings	Places of Worship	Recycling Banks	Conservation Areas	Protected Line of Sight
Barking and Dagenham								
Barnet								
Bexley								
Brent								
Bromley								
Camden								
City of London								
Croydon								
Ealing								
Enfield								
Greenwich								
Hackney								
Hammersmith and Fulham								
Haringey								
Harrow								
Havering								
Hillingdon								
Hounslow								
Islington								
Kensington and Chelsea								
Kingston								
Lambeth								
Lewisham								
Merton								
Newham								
Redbridge								
Richmond								
Southwark								
Sutton								
Tower Hamlets								
Waltham Forest								
Wandsworth								
Westminster								

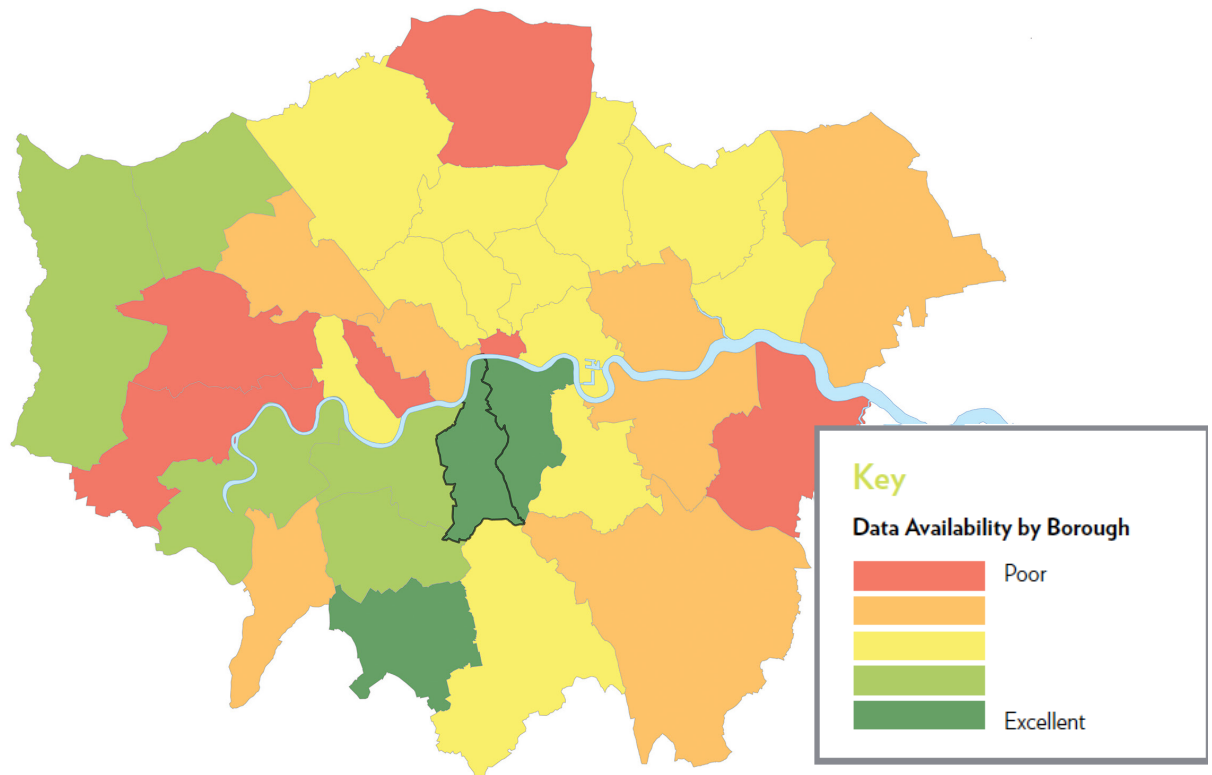


Figure 3 – Data availability mapped across London boroughs (WSP, 2014b)

### 2.3 Decision Support Tool

Web mapping applications have gained in popularity in recent years. Many more sites are using maps within their system as a way of visually representing necessary data and information to their users. For this study it was decided that a decision support tool would be created in the form of an interactive web map application. The integration of both the advantages of a web mapping application and the functionality of a GI system would allow for a more dynamic tool to support the decision making process. Due to the integration of different teams and disciplines within WSP for this project, meant that there were differing views on the importance and development potential of the various data layers. This meant that a flexible system was required in order for them to explore the data as necessary.

### 2.3.1 System Architecture

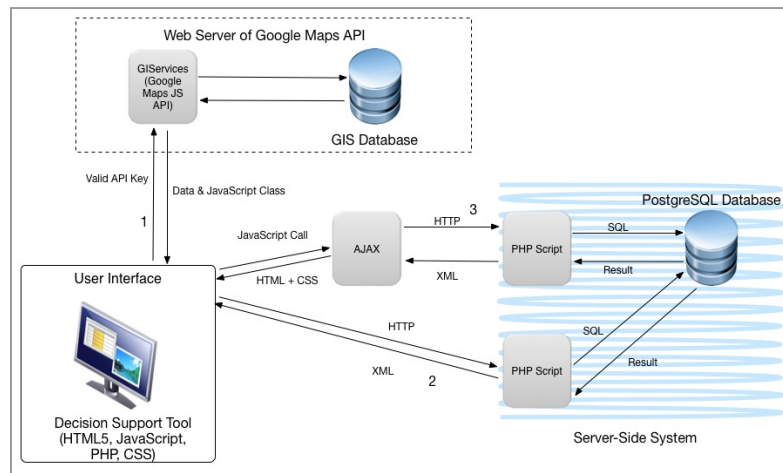


Figure 4 – Decision support tool system architecture

There are three main functions within this tool:

Load the Google Maps JavaScript v3 API (Process 1)

Load layers when toggle button is pressed (Process 2)

Compute the spatial query (Process 3)

The resulting decision support tool was created using a combination of different programming languages including: HTML5, Java Script, PHP, CSS and SQL.

## 3. Results and Analysis

### 3.1 Area Calculations

Table 2 is a summary of the area calculations that were obtained from the point of interest and OSMM building spatial join. There are currently 332 buildings in Lambeth that have been identified as being owned by Central Government, GLA or Local Authority, 78 of which have a listed status associated with them. This total of 332 does not take into account all of the buildings classified as a hospital, with only a single unit being counted for each of these sites. Table 3 represents the area calculation figures in terms of residential units, as estimated by WSP (WSP, 2014b). These are computed by estimating that a single residential unit would be 100 m<sup>2</sup>.

Table 2 - Summary of Area Calculations for Lambeth

	Building Footprint Single Storey Area (m <sup>2</sup> ) (minus listed buildings)	6 Extra Storeys Area (m <sup>2</sup> ) (minus listed buildings)	12 Extra Storeys Area (m <sup>2</sup> ) (minus listed buildings)
Sub-Total	264,134	1,584,804	3,169,608
Total Minus Duplicated Buildings	237,089	1,557,759	3,142,563

Table 3 – Residential Unit Estimates for Lambeth based on Area Calculations and a single residential unit measuring 100m<sup>2</sup> (WSP, 2014b)

Residential Unit WSP Estimates	6 Extra Storeys	12 Extra Storeys
Sub-Total	15,848	31,696
Total Minus Duplicated Buildings	15,578	31,426

Based on Lambeth, estimates of London's full municipal land overbuilding potential for all categories of municipal land have been calculated and mapped by WSP. This has been derived from scaling the Lambeth total potential to each borough based on the proportion of total population of each borough. This uses the assumption that supply and demand for municipal amenities are proportional to population.

Using single storey additions to all municipal buildings would provide 639 hectares of developable space.

- Six storey additions to all buildings would provide 4200 hectares
- 12 storey additions to all buildings would provide 8475 hectares

WSP have therefore calculated a mixed height solution (half six and half 12 storeys) would provide 6337 hectares. On this land, based on 100m<sup>2</sup> a unit, you could build 633,700 residential units.

Contains Ordnance Survey data © Crown copyright and database right 2014.

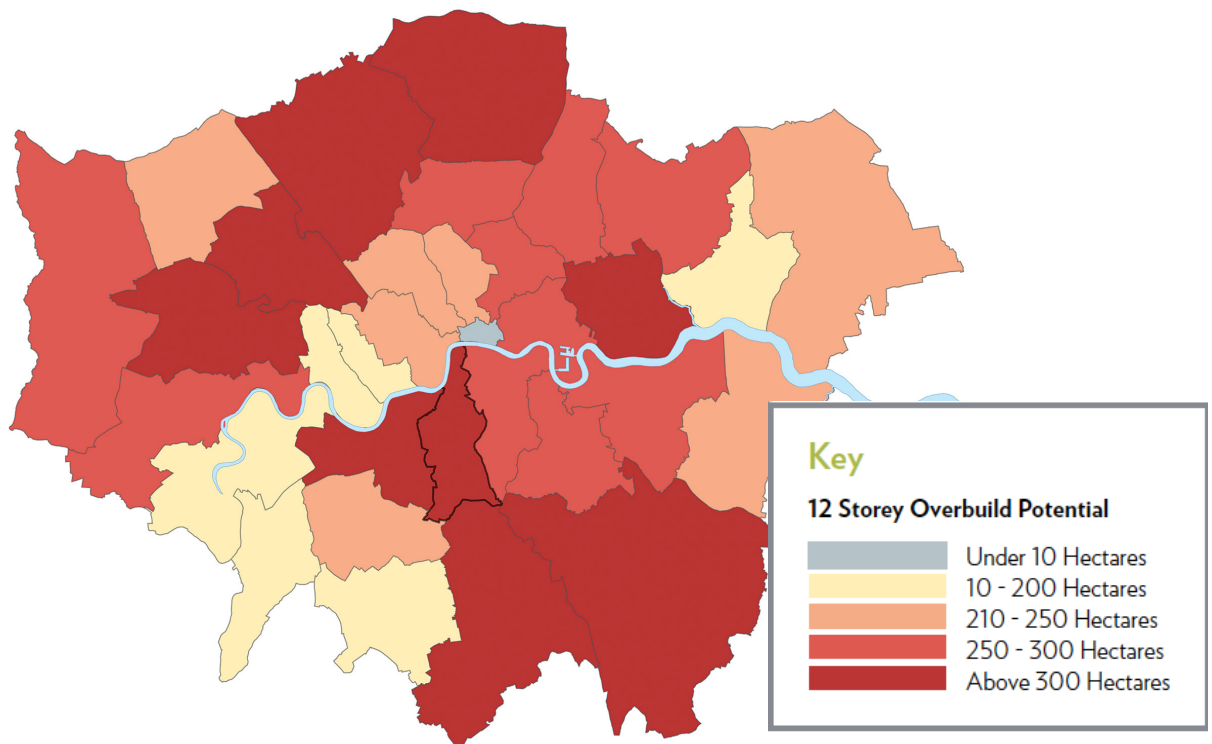


Figure 5 – Estimated 12 Storey Overbuilding Potential of London (WSP, 2014b)



### 3.2 Decision Support Tool

The web mapping decision support tool that has been presented is within the prototype stage of its development. This means that the basic functionality of the tool has been created, however there are a number of additional elements and functions that could be included.

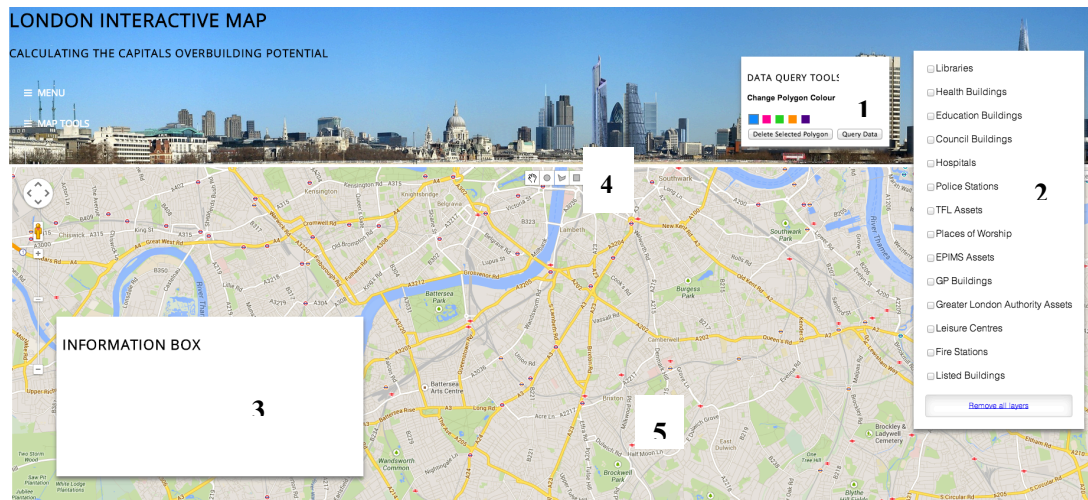



Figure 6 – Screen Shot of the Final Decision Support Tool

1. These are the Data Query Tools. From here the users is able to change the polygon colour and delete a polygon when selected. An additional button has been placed here, 'Query Data' which is currently not operational but would conduct the query when clicked rather than when the polygon shape was drawn.
2. This is the map layers toolbar. Layers can be added to the map by clicking on its associated button
3. This is the information box that is updated when the building in polygon query is started
4. There are the drawing tools. The  button is the necessary drawing tool to construct the polygon.
5. This is the map created from the Google JavaScript APIv3

### 4. Discussion

From these results, it can be seen that this study has achieved its aim of investigating the overbuilding potential for Lambeth. The most significant challenge in this project has been the access to open data. Although Government had requested complete data transparency as outlined in the Local Government Transparency Code 2014 (Department for Communities and Local Government, 2014), gaining accurate information regarding the location and size of land and buildings owned by local authorities across London has been extremely difficult.

Although Lambeth's data is more complete than most other sources, it could be improved further by providing metadata relating to the method of data collection. Detailed metadata to support the data that is published is crucial for the public to be able to understand the quality

and completeness of the data.

Making specific residential unit calculations at this stage is very difficult, as it is necessary to gain accurate building height data, as well as specific building floor usage for each of the points of interest. In addition to this, making calculations in terms of residential units could be misleading, as architects and designers would determine total property size at case specific stages within future site development. What this project has done is to present a new concept to solve the housing crisis, highlighted the missing data sets, calculated building area total for the data available and visualised this data in an interactive decision support tool.

## **5. Conclusion**

For the area of Lambeth, the next stages would be looking at buildings on a case-by-case basis in order to accurately obtain calculations as to the number of residential units each building could provide. To be able to accurately calculate areas for the rest of London would require that more datasets become available.

It is clear that there is not going to be an overnight fix to the situation, nor will that one solution on its own solve the housing problem. However, it has highlighted an opportunity that may be available, and that has a strong foundation for further research and development. Of course, it is unrealistic to state that all the land and properties that have been identified should be and will be overbuilt with housing. The main point that has been raised is that there is public land and property available, some of which could benefit from collaboration with private companies. By ensuring that data regarding governmental and local authority assets are kept up-to-date and at a specific standard level of detail will ensure transparency is achieved and potential assets can be identified. Maintenance of the datasets is key. This is essential for the continued development of this study from borough level to covering the whole of London, the South East and the rest of the UK.

## 6. References

- DEPARTMENT FOR COMMUNITIES AND LOCAL GOVERNMENT 2014. Local Government Transparency Code 2014. *In: GOVERNMENT, D. F. C. A. L. (ed.)*. London.
- FAINSTEIN, S. S. 2008. Mega-projects in New York, London and Amsterdam. *International Journal of Urban and Regional Research*, 32, 768-785.
- GLA 2012. London View Management Framework: Supplementary Planning Guidance. *In: AUTHORITY, G. L. (ed.)*. London.
- HERITAGE, E. & CABE 2007. Guidance on tall buildings. *In: HERITAGE, E. & CABE (eds.)*.
- HILL, D. 2013. London housing crisis: what is it, exactly? *Dave Hill's London Blog* [Online]. Available from: <http://www.theguardian.com/uk-news/davehillblog/2013/oct/28/london-housing-crisis> [Accessed 28th October 2013].
- JOHNSON, B. 2013. 2020 Vision - The Greatest City on Earth: Ambitions for London. *In: AUTHORITY, G. L. (ed.)*. London.
- KPMG & SHELTER 2014. Building the homes we need: A Programme for the 2015 Government. *In: KPMG (ed.)*. Online.
- NATIONS, U. 2012. World Urbanization Prospects: The 2011 Revision. New York: Department of Economic & Social Affairs.
- NLA 2014. London's Growing Up! - NLA Insight Study. *In: NLA (ed.)*. London: NLA - London's Centre for the Built Environment.
- ONS 2014a. Release: House Price Index, May 2014. Online
- ONS 2014b. Subnational Population Projections, 2012-based projections. *In: STATISTICS, O. O. N. (ed.)*.
- SAVILLS. 2014. London's housing need becomes more concentrated in lower mainstream - Is it time for developers and investors to look beyond prime? Available: [http://www.savills.co.uk/\\_news/article/72418/176668-0/5/2014/london-s-housing-need-becomes-more-concentrated-in-lower-mainstream---is-it-time-for-developers-and-investors-to-look-beyond-prime-](http://www.savills.co.uk/_news/article/72418/176668-0/5/2014/london-s-housing-need-becomes-more-concentrated-in-lower-mainstream---is-it-time-for-developers-and-investors-to-look-beyond-prime-).
- THE ECONOMIST 2014. The ascent of the city. *The Economist*. Print.
- WSP 2013. Delivering the London 2020 Vision.
- WSP 2014a. Build Homes Above Hospitals to Solve London's Housing Shortage. *In: GROUP, W. (ed.)*.
- WSP 2014b. BUILDING OUR WAY OUT OF A CRISIS - CAN WE CAPITALISE ON LONDON'S PUBLIC ASSETS TO PROVIDE HOMES FOR THE FUTURE?

## 7. Biography

*Joanna Foster is a recent MSc GIS graduate from UCL. She currently works at Harper Dennis Hobbs as a Consultant within the Retail Consultancy team.*

*Philippa Wood works at WSP |Parsons Brinckerhoff as a GIS Analyst within their Development team, and is also an alumnus of the UCL MSc course in GIS.*

# **Assessing the quality of OpenStreetMap building data and searching for a proxy variable to estimate OSM building data completeness**

Claire Fram<sup>1</sup>, Katerina Chistopoulou<sup>2</sup> and Claire Ellul<sup>3</sup>

<sup>3</sup> Dept. of Civil, Environmental and Geomatic Engineering, University College London

Gower Street, London, WC1E 6BT

Tel. +44 (0) 20 7679 4118 Fax +44 (0) 20 7380 0453

9 January 2015

## **1. Introduction**

OpenStreetMap (OSM) is an open data, geospatial information (GI) project that relies on contributions from volunteers to create a digital, on-line, map of the world. The use of OSM is also free and available under the Open Database License (Hecht, et al., 2013). As the amount of information available in the OSM database continues to grow, interest has grown regarding the application of OSM in industry and according to scientific standards (Haklay, 2010; Hecht, et al., 2013; Koukoletsos, et al., 2012). However the quality of OSM building data is largely unknown and thus the practical applications for OSM building data remain limited.

This study was taken on with the support and supervision of Risk Management Solutions (RMS) to investigate the quality of OSM building data with the purpose of assessing OSM build data's potential application in RMS products, specifically natural catastrophe exposure models.

To incorporate OSM building data into exposure modelling, the quality of OSM data must be known or the uncertainty of OSM data completeness must be quantified. Unless OSM data quality can be quantified, applying OSM data to any commercial products would introduce an unacceptable, unknown quantity of uncertainty.

Two objectives were defined in this study. The first objective is to understand OSM building data quality. Multiple case studies were used to reveal the similarities and differences between OSM data quality in different places. The second objective of this study is to test whether OSM building data quality might be estimated with a proxy variable. Identifying an appropriate proxy variable might offer RMS, and use-cases like RMS's, to quantify the quality of OSM building data in areas without official reference data.

---

<sup>1</sup> claire.fram.13@ucl.ac.uk

<sup>2</sup> Katerina.Christopoulou@rms.com

<sup>3</sup> c.ellul@ucl.ac.uk

## 2. OSM building data assessment

OSM building data quality was tested in three study areas Leeds, London and Sheffield using a *unit-based* assessment defined by Hecht et al. (2014). To highlight the heterogeneity of OSM completeness and accuracy, the case study areas were divided into smaller sub-sections. Within each sub-section, OSM data was compared against data from OS Street View. Using this approach, OSM data completeness is quantified at the spatial resolution of the sub-unit. “Spatial units” (Hecht, et al., 2013, p. 1073) for this study were defined as 1km<sup>2</sup> cells. This grid size was used in Koukoletos et al.’s study (2012).

Hecht et al.’s unit-based method (2013) for assessing OSM polygon completeness was applied. In this method the aggregate area of building footprints’ per-spatial unit was used to quantify data completeness. The 1km<sup>2</sup> grid was applied to each study area; buildings that were partially located in multiple spatial units were subdivided by the spatial unit boundaries (illustrated in Figure 1).

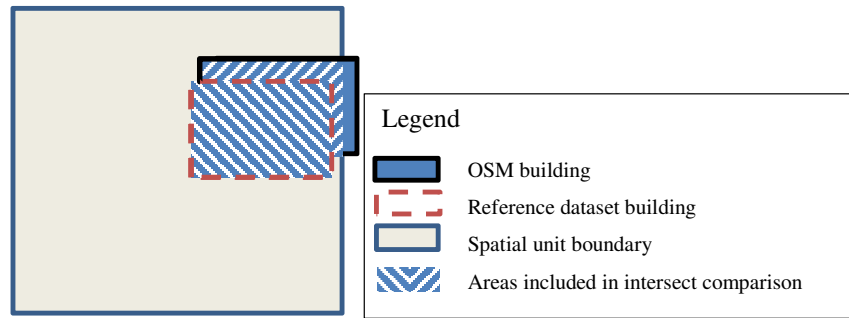


Figure 1: Example of completeness calculation based on footprint intersections with spatial unit boundaries

The OSM building footprint area within each cell was aggregated and compared with the aggregate OS Street View building footprint located in the same cell (Equation 1). The product of this equation serves to quantify OSM building completeness and can be compared across sub-units.

$$C_{Area} = \frac{\sum Building\ footprint_{OSM}}{\sum Building\ footprint_{Ref}} \times 100$$

Equation 1: Unit based method: area (Hecht, et al., 2013, p. 1076)

Of the three case study areas, Leeds had the lowest level of OSM building completeness, as a measure of aggregate footprint coverage per square kilometre. When the completeness percentages of all sub-units was averaged within each study area, Leeds had an average completeness ratio of 30%, while Sheffield has the highest at 75% (Table 2). London also had a low average completeness ratio (33%, Table 2). However, London is a larger study area. 1456 km<sup>2</sup> were included in London’s assessment of OSM data quality, compared to 469 km<sup>2</sup> in Leeds and only 292 km<sup>2</sup> in Sheffield. The OSM completeness ratio for each city is represented at a 1km<sup>2</sup> resolution in Figure 3, Figure 4 and Figure 5. Results describing the distribution of OSM completeness estimates by 1km<sup>2</sup> are seen in Figure 2.

In calculating the average OSM completeness ratios, each spatial unit ( $\text{km}^2$ ) was given equal weight. However, building density was not always equally distributed spatially. The number of buildings represented in each spatial unit ( $\text{km}^2$ ) is presented in Table 3.

If areas with high OSM quality are considered, only 13% of Sheffield's buildings are located high quality OSM areas. In Leeds, 2% of buildings are in high quality OSM areas, and only 3% of London buildings are in areas of high quality OSM. However, London has the largest actual count of buildings (29352) located in these high quality areas (Figure 2 and Table 3).

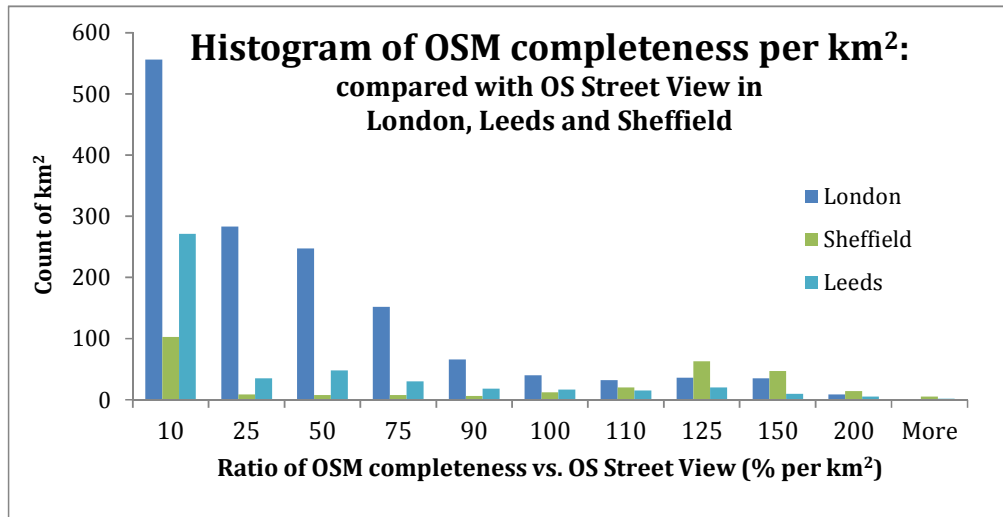


Figure 2: Histogram of OSM completeness ratio results for Leeds, London and Sheffield

Table 1: Results uses to calculate histogram of OSM area-completeness by km<sup>2</sup>

<b>Bin</b>	<b>London</b>		<b>Sheffield</b>		<b>Leeds</b>	
	<i>Count of Km<sup>2</sup></i>	<i>Percentage of total Km<sup>2</sup></i>	<i>Count of Km<sup>2</sup></i>	<i>Percentage of total Km<sup>2</sup></i>	<i>Count of Km<sup>2</sup></i>	<i>Percentage of total Km<sup>2</sup></i>
<b>10%</b>	556	38%	103	35%	271	7%
<b>25%</b>	283	19%	9	3%	35	10%
<b>50%</b>	247	17%	8	3%	48	6%
<b>75%</b>	152	10%	8	3%	30	4%
<b>90%</b>	66	5%	6	2%	18	4%
<b>100%</b>	40	3%	12	4%	17	3%
<b>110%</b>	32	2%	20	7%	15	4%
<b>125%</b>	36	2%	63	21%	20	2%
<b>150%</b>	35	2%	47	16%	10	1%
<b>200%</b>	9	1%	14	5%	5	0%
<b>More</b>	1	0%	5	2%	2	100%

Table 2: Average OSM completeness ratios (compared with OS Street View)

	London	Sheffield	Leeds
Average OSM completeness ratio	33%	75%	30%

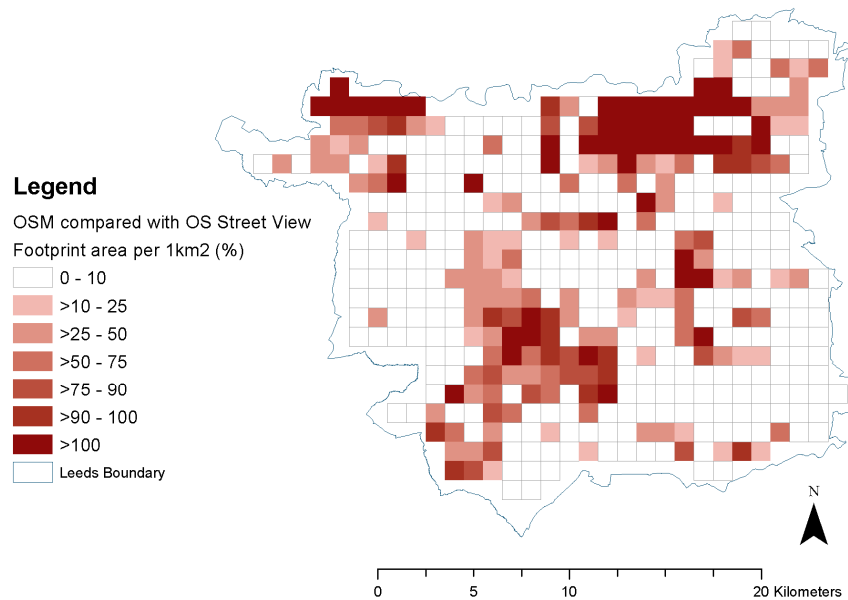
Table 3: Assessment of OSM completeness ratios in the context of building density

	London		Sheffield		Leeds	
	Count	%	Count	%	Count	%
Total OS Street View Building Count in study area	864916	100%	83310	100%	137632	100%
Total OS Street View buildings in grids with completeness ratios <=10%	364866	42%	5535	7%	85018	62%
Total OS Street View buildings in grids with completeness ratios >10%	500050	58%	77775	93%	52614	38%
Total OS Street View buildings in grids with completeness ratios >=90% and <=110%	29352	3%	10307	12%	3392	2%



## Leeds Aggregate Coverage Comparison

OSM vs OS Street View building footprint coverage



Data Sources: OpenStreetMap building data (downloaded 11 May 2014); Ordnance Survey data © Crown copyright and database right 2014

Figure 3: Leeds, map of aggregate footprint coverage comparison between OSM and OS Street View

## London Aggregate Coverage Comparison

OSM vs OS Street View building footprint coverage

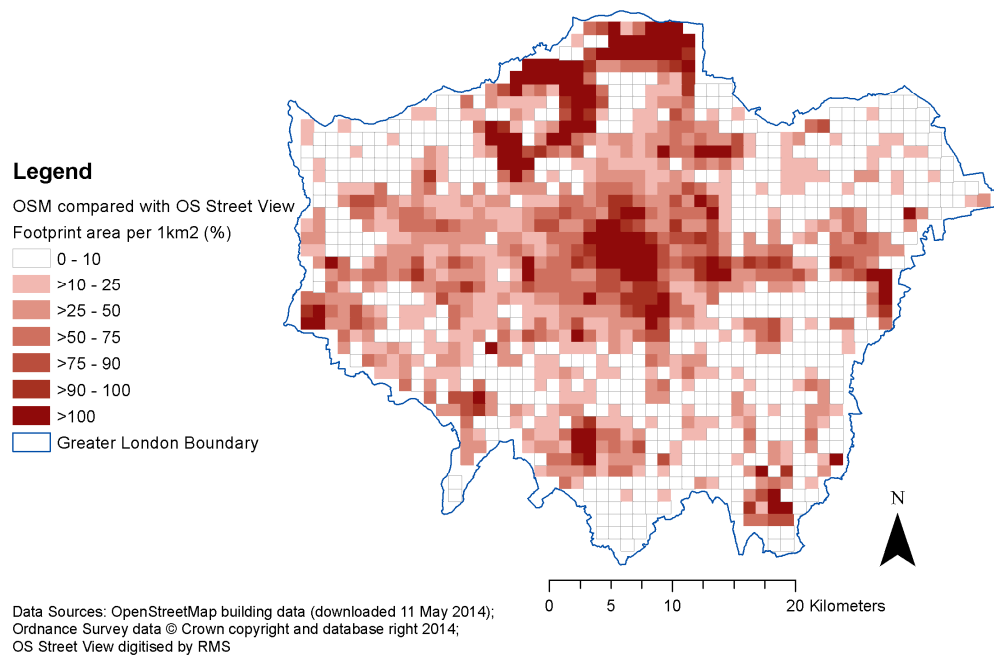
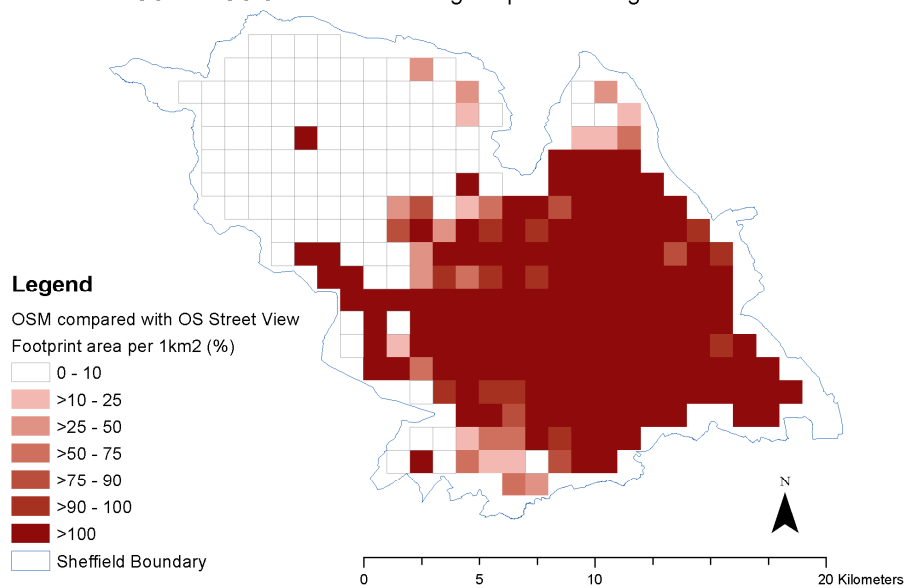


Figure 4: London, map of aggregate footprint coverage comparison between OSM and OS Street View

## Sheffield Aggregate Coverage Comparison

OSM vs OS Street View building footprint coverage



Data Sources: OpenStreetMap building data (downloaded 11 May 2014); Ordnance Survey data © Crown copyright and database right 2014  
OS Street View digitised by RMS

Figure 5: Sheffield, map of aggregate footprint coverage comparison between OSM and OS Street View

### 3. Testing a proxy for OSM building data completeness

In the applied, unit-based method a reference dataset was required to benchmark OSM building data quality. However, OSM building data has the most potential to reduce uncertainty in RMS exposure models in areas where reliable reference datasets are not available. While OSM building data quality is likely heterogeneous in all cities (see results in section 2), this section investigates if there is a relationship between OSM building data quality and an independent, proxy variable.

This section explores the relationship between LandScan gridded population data and OSM building data completeness. Population data was chosen as a potential proxy for OSM data completeness for two reasons. First, across international case studies (Germany, The United Kingdom and America), a strong correlation between population density and OSM road network density has been proven (Haklay, 2010; Hecht, et al., 2013). Second, population information is available globally at a 1km<sup>2</sup> resolution via LandScan gridded population data (Oak Ridge National Laboratory, 2014). The global coverage offered by this dataset allows for any relationship discovered between OSM completeness and population density to be tested and applied internationally.

To compare OSM data completeness with population density, a centroid was extracted from each 1km<sup>2</sup> grid of the unit-based assessment, and the completeness result metric was spatially joined with LandScan population density information. A simple correlation assessment was used to assess the feasibility of using population as a proxy variable for OSM area-completeness.

There was no significant correlation discovered between OSM area completeness and population density in any of the cities. In London correlation of population density to OSM completeness estimates had an R-squared value of 0.1753. The relationship R-squared value was 0.3044 and for Leeds it was 0.0010.

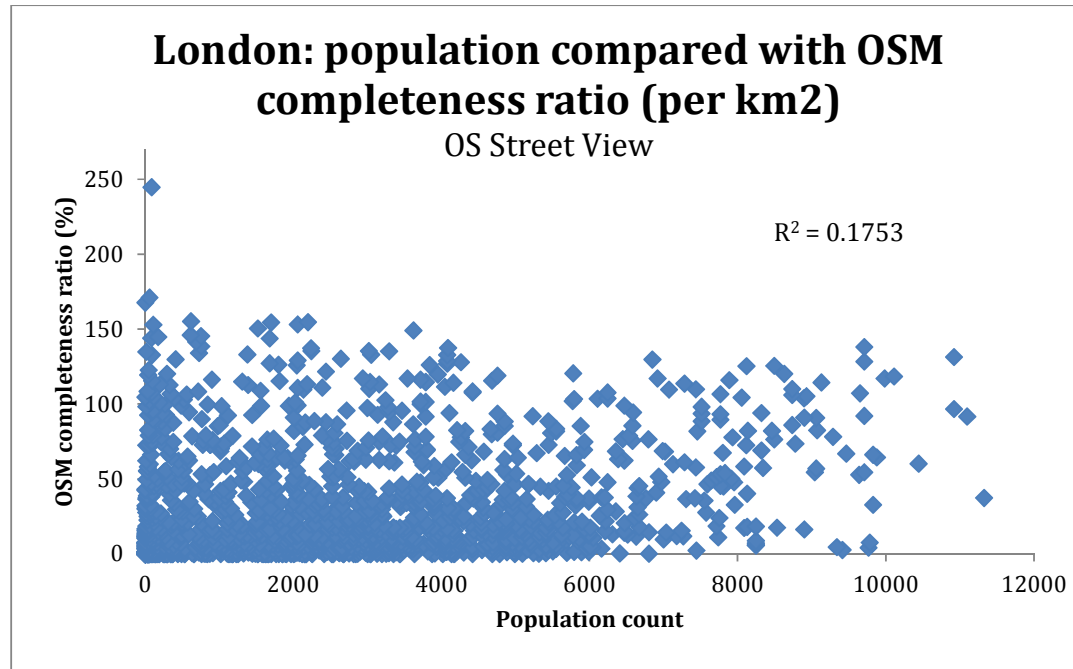


Figure 6: London, correlation between population density and OSM completeness (OS Street View as reference dataset)

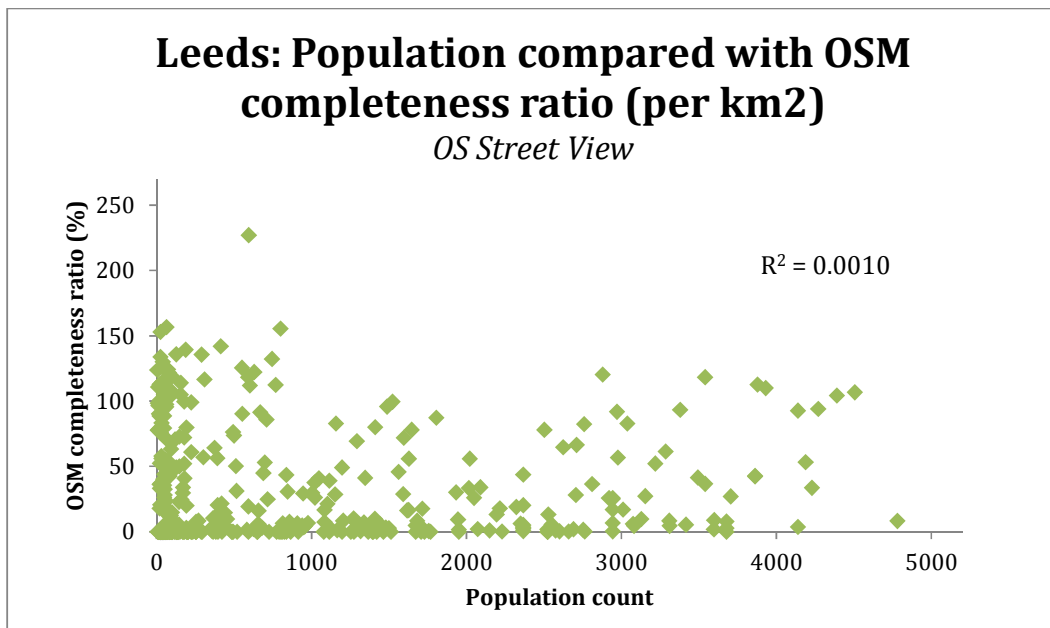


Figure 7: Leeds, correlation between population density and OSM completeness (OS Street View as reference dataset)

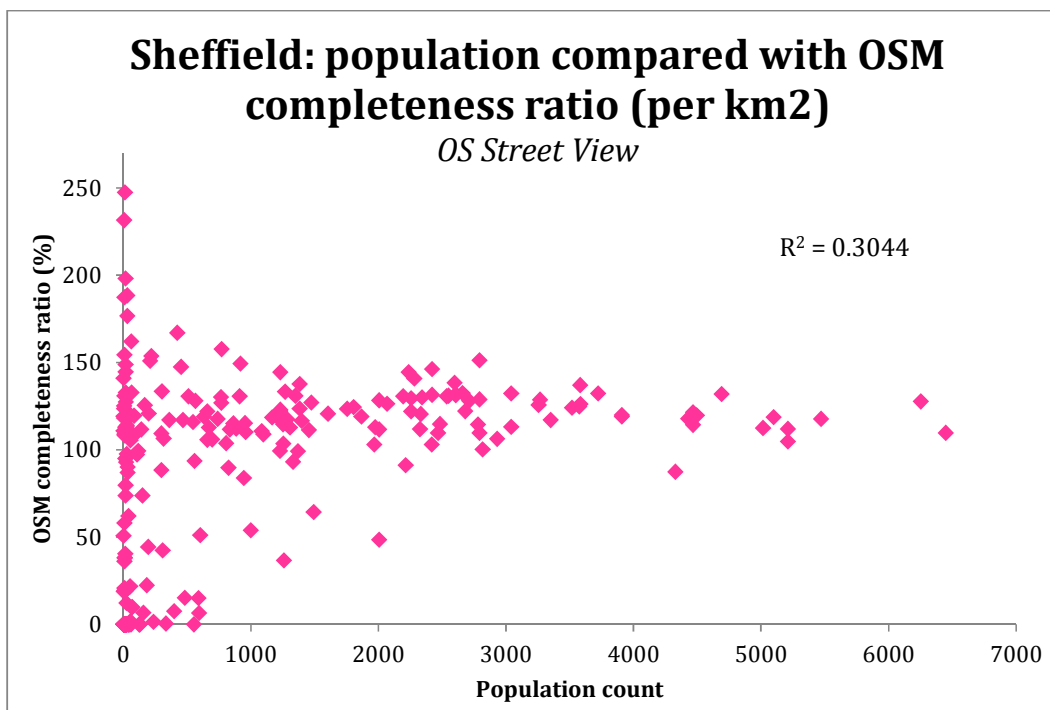


Figure 8: Sheffield, correlation between population density and OSM completeness (OS Street View as reference dataset)

#### 4. Conclusion and further research

In order for OSM building data to add value to products like RMS's natural catastrophe exposure model, or other commercial products, OSM building completeness must be understood. This study proved that OSM building data completeness within UK cities and between UK cities is variable. However, without a reference dataset, estimating OSM building completeness is not possible using population as a proxy variable. Alternative proxy variables should be sought. Therefore this study did not discover a viable method for estimating OSM completeness for regions with limited official data.

#### 5. Acknowledgements

This research was completed with the support of Risk Management Solutions. Katerina Christopoulou, my principal supervisor at Risk Management Solutions provided invaluable guidance to this study. I am very grateful for Katerina's early commitment to the project and persistent mentorship. Claire Elull, my academic advisor also provided a great amount of support to the completion of this study.

#### 6. Biography

Claire Fram received her MSc in Geographical Information Science from UCL in 2014. She is now a Graduate Specialist in GIS at Arup in London. Her research interests include: open data, data visualisation and data analytics.

#### 7. References

Haklay, M., 2010. How good is OpenStreetMap information? A comparative study of OpenStreet-Map and Ordnance Survey datasets for London and the rest of England. *Environment and Planning B: Planning and Design*, Volume 37, pp. 682-703.

Hecht, R., Kunze, C. & Hahmann, S., 2013. Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS International Journal of Geo-Information*, 11 November, Volume 2, pp. 1066-1091.

Jackson, S. P. et al., 2013. Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, Issue 2, pp. 507-530.

Koukoletsos, T., Haklay, M. & Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, 14(4), pp. 477-498.

Mooney, P., Corcoran, P. & Winstanley, A. C., 2010. *Towards Quality Metrics for OpenStreetMap*. s.l., ACM, pp. 514-517.

Oak Ridge National Laboratory, 2014. *LandScan*. [Online] Available at: [web.ornl.gov](http://web.ornl.gov) [Accessed 1 September 2014].

# Assessing geographic data usability in analytical contexts: Undertaking sensitivity analysis of geospatial processes

R. Frew<sup>1</sup>, G. Higgs<sup>1</sup>, M. Langford<sup>1</sup>, J. Harding<sup>2</sup>

<sup>1</sup> GIS Research Centre, WISERD, Faculty of Computing, Engineering and Science, University of South Wales, Treforest, Pontypridd CF37 1DL

<sup>2</sup> Ordnance Survey, Explorer House, Adanac Drive, Southampton SO16 0AS

April 6, 2015

## Summary

This paper addresses the comparative dearth of research on spatial data usability by employing sensitivity analyses to the findings from applying GIS-based accessibility models. Comparisons were made using approaches based on Euclidean distances and more sophisticated accessibility measures that utilise network travel distances: the latter incorporating measures of supply and demand by using innovative extensions to the Two-Step Floating Catchment Area method (2SFCA). To illustrate the sensitivity of findings from applying such models with a range of data sources, geographic accessibility to secondary schools was calculated for Output Areas in South Wales using a 2SFCA plug-in to ArcGIS<sup>TM</sup>. By using different permutations of spatial data, for both the supply- and demand-side parameters in such models, differences in walking distances and FCA scores were sought in order to comment on the usability of such data sources. Preliminary conclusions are made on the appropriateness of such data sets in relation to different types of network-based accessibility modelling tasks.

**KEYWORDS:** Usability; GIS-based accessibility models; spatial data; sensitivity analysis; E2SFCA.

## 1. Introduction

Sources of spatial data continue to expand with inevitable debates surrounding the provenance of such data and their usability for GIS-based tasks. There is therefore an increased scrutiny as to the quality and usability of such data and the respective advantages and limitations of both proprietary and crowd-sourced data.

Although the highest quality data often remains expensive to obtain for some users (for example high resolution LiDAR data, or Ordnance Survey MasterMap products), other data sets are becoming available without the need for expensive capital or revenue outlay. Recent reports (for example, Avery and Gittings, 2014) on the use of unmanned aerial vehicles to produce a variety of remotely-sensed data as well as the availability of various software solutions, both at low costs relative to traditionally-sourced equivalents are enabling new data-producers to emerge. Such trends are paralleled by the opportunity for data users to generate their own data for their own purposes. At the same time the quality of such data is being questioned in some quarters, reinforcing earlier debates surrounding the use of VGI (volunteer geographic information) and previous work in the field of data quality theory and assessment (Haklay, 2010; Zielstra and Zipf, 2010). However, there is still very little research into the usability of such data in relation to different types of GIS-based tasks, although Higgs et al (2012) did investigate the impacts of different approaches to measuring accessibility to green space using comparable sources of data. Few studies to date incorporate sensitivity analysis that involve the use of different sources of spatial data applied to different stages of a ‘typical’ GIS project (although see Jones (2010) for a notable exception).

This paper will report on the usability of a range of geographical data in one such application area: namely their use in network-based accessibility modelling. Based on these findings preliminary

assessments will be made on their usefulness in such tasks, using both relatively routine and more sophisticated methods of measuring accessibility.

Accessibility studies using GIS have become a well-established component of geographical studies concerned with measuring potential inequalities in provision of both public and private services and are beginning to be used by policy makers to inform decision making processes. Related fields include studies of the spatial distribution and optimisation of services in areas such as public health, welfare provision and environmental justice. Recent examples of such research include those concerned with examining the geographical distribution of alcohol outlets in Glasgow in relation to deprivation (Ellaway et al, 2010) and disparities in locations of sports facilities in Wales (Higgs et al, 2015).

## **2. Study approach**

### **2.1 Study area**

Two areas in South Wales were chosen for study: the city and county of Cardiff; and the neighbouring Vale of Glamorgan County. Cardiff is the largest city in Wales, although within its county boundary are villages located in the green belt that separates Cardiff from Newport (to the east) and the densely-populated Rhondda valleys to the north. The Vale of Glamorgan has several smaller population centres, with much of the area having rural characteristics despite proximity to major transport links (the M4 motorway) and to large towns and cities (such as Bridgend and Cardiff).

### **2.2 Geographic data**

The spatial data products chosen were typical of those commonly used in UK-based accessibility analysis studies, including Ordnance Survey MasterMap Integrated Transport Network™ (ITN) Layer and the additional Ordnance Survey Urban Path layer. OpenStreetMap (OSM) network data for South Wales was obtained from a third-party provider, as an example of the crowdsourced/VGI data that is now routinely available to GIS researchers. One further dataset was used to examine whether a product not designed for use as a network could approximate the results of the specifically-designed datasets. VectorMap District, available free-to-use from Ordnance Survey OpenData, was built into a network using standard, readily-available GIS tools (using Arc GIS). At the time there was no free-to-use network dataset using Ordnance Survey data, though in March 2015 Ordnance Survey launched the Open Road network dataset, also under their OpenData programme, and this will be subject to later analysis.

The accessibility assessment tasks were also conducted using Euclidean (straight line) distances.

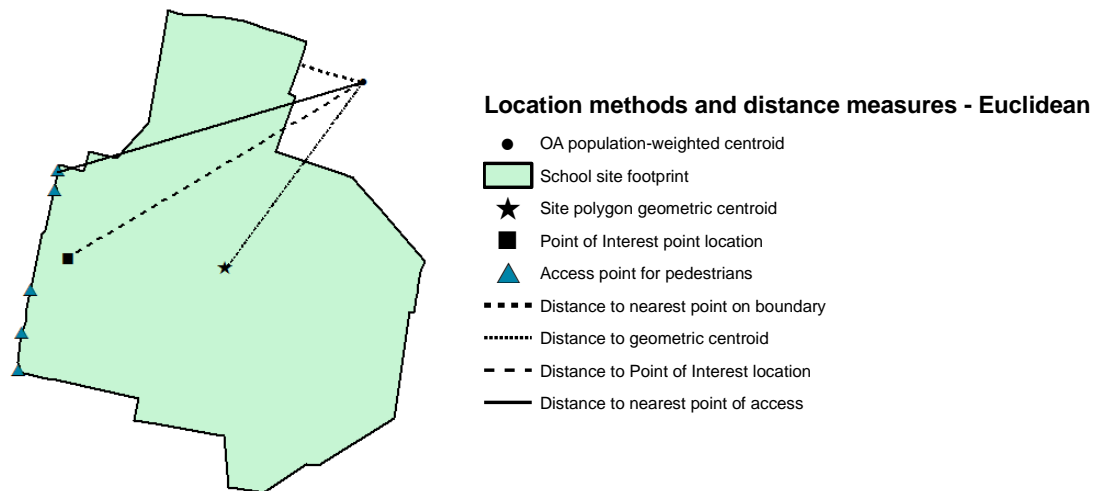
The relative accessibility of different locations within the study areas were assessed using different methods, with the processes subjected to sensitivity analysis in order to identify areas of similarity and difference. Several iterations of each analytical task were therefore performed, using permutations of the different datasets.

### **2.3 Location of supply features**

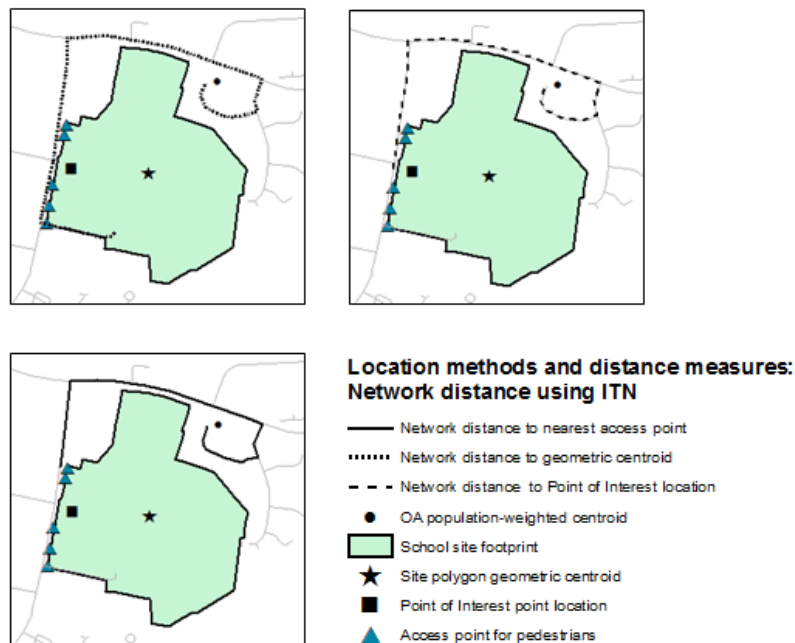
Accessibility studies use various methods to assess the accessibility (or inaccessibility) of demand to supply. In this study, the supply features were secondary schools (with travel-to-school journeys being the focus of many studies relating to active travel and children's health, child road safety, catchment areas, parental choice, etc). The location of such facilities is also subject to a degree of choice by the researcher. Accordingly, as part of the sensitivity analysis, different methods of locating these features were compared. Many studies use points to represent locations, and as secondary schools often occupy large sites, they are ideal for use in comparing different methods of representing polygons by points. The Ordnance Survey Points of Interest dataset was used as the initial point locations of the schools, and Ordnance Survey Sites dataset was used to extract the "footprint" of



entire school sites, including playing fields, etc. From the Sites dataset, three different location methods were compared: centroid (the geometric centroid of the entire site); access point (one or more way in to the school site for pedestrians); and boundary (any point on the perimeter of the site). Figure 1 illustrates how this choice may impact on travel distances, using a typical school site as an example, and Figure 2 shows how network distances may vary using the same variety of location methods.



**Figure 1** Four different approaches to measuring Euclidean distance from a location to a facility.



**Figure 2** Examples of network distance variations, from a point location to a local facility.

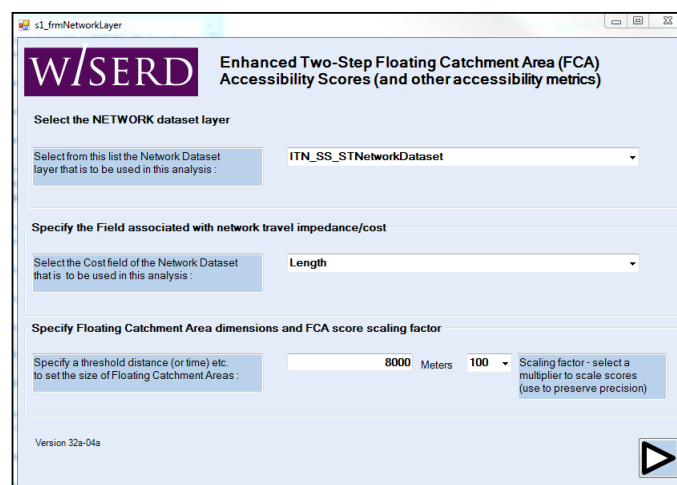
## 2.4 Location of demand

Various demand (population) representations are available to researchers. In this study, the method of locating the population included the use of 2011 UK Census Output Area (OA) population-weighted centroids. Other methods are available, both more detailed (at post code or address level, for example) or more generalised (for example, either of the census Super Output Area layers, both of which are aggregations of OAs). The method chosen uses readily-available and free-to-use data that is sufficiently detailed to allow differences to be identified between smaller areas while avoiding the increased computational loads and visualisation challenges resulting from the use of more detailed representations.

## 2.5 Methodology

This study calculated Euclidean and network distances with different permutations surrounding supply-side options. Comparisons were then made between the results of the various iterations of the accessibility models, both visually and statistically. One indication of similarity was achieved through comparison of Destination Overlap, where the identities of the supply facility nearest to each demand centre was compared for each network option (see Table 1).

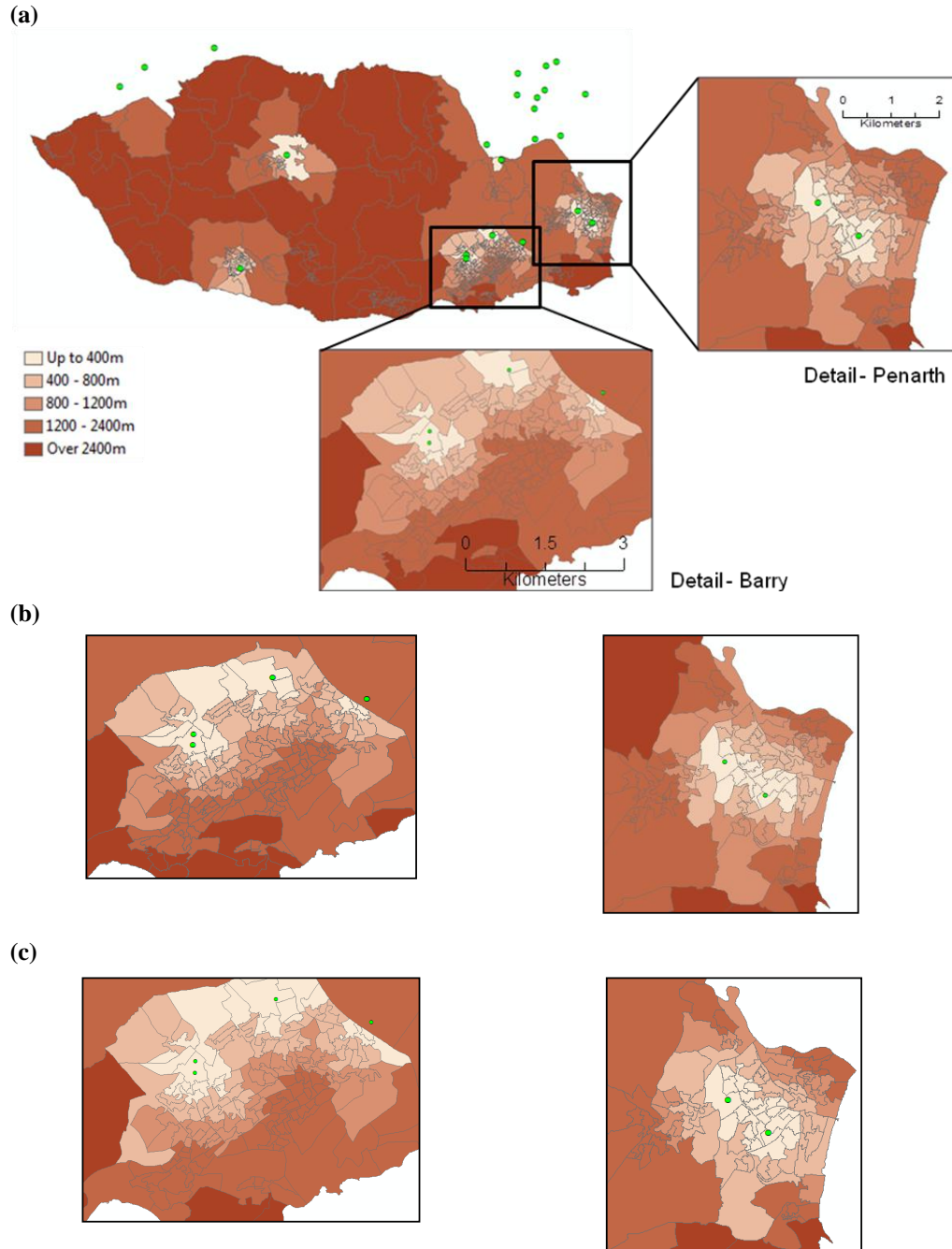
Euclidean distance ignores the actual travel route taken, and as with network distances takes no account of the capacity of the supply facility (in this case the number of school places available), nor the level of demand (in this case the secondary school-age population). Accordingly, a more sophisticated measure of accessibility was used based on the enhanced two-step floating catchment area method (E2SFCA). A tool developed by researchers at the University of South Wales was used in ArcGIS™ (Figure 3 shows the user interface of the plug-in). In order for the tool to be used effectively, data on pupil numbers/school roll was obtained from information published by the relevant local authority, and an estimate of the school-age population of each OA made from age categories contained in published 2011 census data. Although there was no convenient category of “secondary school age” in the census, there was information on 12 to 16 year olds, and an estimate was made of the numbers of students at school outside these age categories.



**Figure 3** Illustration of the E2SFCA plug-in tool first screen. Further screens offer the options of incorporating levels of supply and demand.

### 3. Preliminary Findings

Patterns of “worst” and “best” accessibility were broadly similar, but with differences stemming from the method of locating the supply feature (as in Figure 4) and the choice of network dataset.



**Figure 4** Examples of variations in distances from OAs to their nearest facility, depending on the method used to locate that facility (in this case, secondary schools in the Vale of Glamorgan).

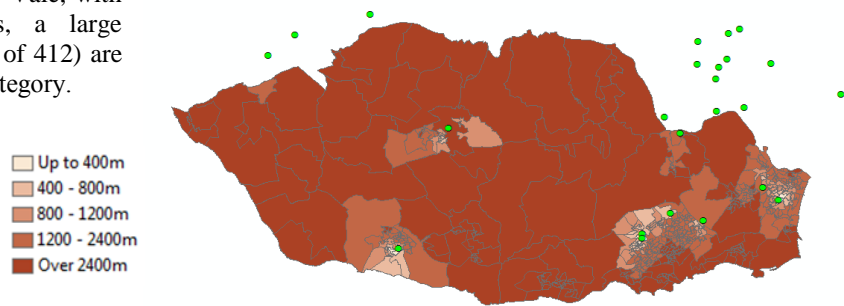
- (a) OS Sites dataset, OA to polygon centroid, Euclidean distance;
- (b) OS Sites dataset, OA to nearest access point, Euclidean distance;
- (c) OS Sites dataset, OA to nearest point on site boundary, Euclidean distance.

Differences in findings when applying alternative methods of measurement (i.e. between distance measures and the results of E2SFCA calculations, examples of which are shown in Figure 5) were evident, highlighting the impacts of using different approaches on the results from GIS-based models. Figure 5(b) illustrates the effect of supply and demand on accessibility, and the influence of the size of catchment area used in E2SFCA calculations. Preliminary findings also indicate urban/rural differences which also merit further investigation, an example of such differences is shown in the Distance Overlap figures of Table 1.

In contrast, other findings suggest that the use of different datasets or different network products may make no statistical difference to outcomes either in terms of distance or E2SFCA scores. All distance results and all E2SFCA results were significantly correlated (using Spearman's rank correlation) and Table 2 shows Wilcoxon Z scores for E2SFCA results from the Vale. Several paired comparisons (18%) have no significant differences between their scores at the 5% level, though the majority of comparisons (76%) have differences that are significant at the  $< .001$  level.

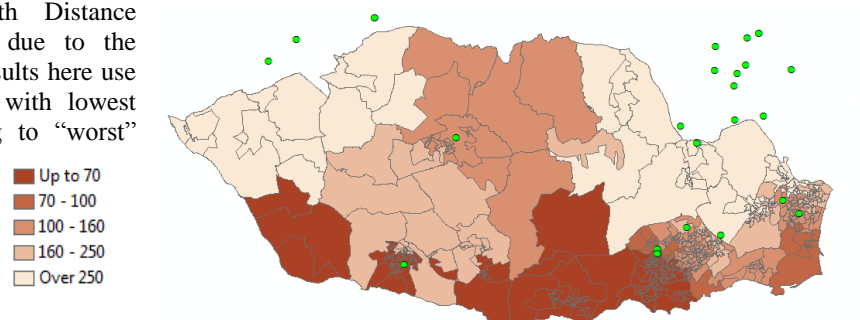
**(a) Distance measures for OAs to school centroids, ITN network.**

In rural areas such as the Vale, with relatively few facilities, a large number of OAs (150 out of 412) are in the “worst” distance category.



**(b) 2SFCA for OAs to school centroids, ITN network.**

Direct comparison with Distance measures are difficult, due to the different scales used. Results here use rounded quintile splits, with lowest E2SFCA score equating to “worst” accessibility.



**Figure 5** Example of differences in accessibility visualisations obtained when using ITN, comparing distance and E2SFCA measures: (a) Distance; (b) E2SFCA.

As part of this PhD research, further analysis of the results will be conducted with the intention of isolating the factors within the underlying data that contributed to the differences in accessibility scores within the same output areas. However, the extent of differences identified in this preliminary analysis leads us to suggest that researchers need to be made more aware of the implications of using different sources of data in ‘typical’ GIS tasks.

Practical issues with different datasets (for example cost, ease of download, availability, etc) along with their currency and update patterns, are also worthy of further study. Identifying methods whereby the usability of these spatial data sources can be made more transparent to researchers prior to their implementation in GIS analytical tasks is one of the ultimate aims of this research.

	Euclidean			ITN			UP			OSM			VMD		
	Cent	AccPt	Per	Cent	AccPt	Per	Cent	AccPt	Per	Cent	AccPt	Per	Cent	AccPt	Per
<b>Euclid</b>															
Cents		87.6	96.6	83.3	84.5	93.2	86.4	89.6	94.7	81.6	78.6	78.4	83.5	81.1	80.8
AccPts	86.9		86.2	80.8	87.9	83.0	84.0	92.0	83.0	85.7	89.3	85.0	85.9	89.3	84.0
Perim	84.9	92.7		80.6	83.3	92.7	83.7	87.4	93.2	93.5	82.8	92.2	92.7	82.5	92.7
<b>ITN</b>															
Cents	82.0	76.6	74.8		84.0	84.0	94.9	87.9	84.0	93.5	81.1	80.6	99.5	84.5	84.0
AccPts	76.5	81.2	76.0	92.9		87.4	87.4	93.2	86.6	84.0	97.6	87.6	83.5	95.1	87.4
Perim	72.8	73.0	75.5	92.5	88.2		86.9	87.9	97.1	83.3	86.7	96.4	82.5	86.4	98.1
<b>UP</b>															
Cents	89.2	87.7	87.5	90.0	85.7	83.2		90.3	87.9	90.0	82.8	80.6	94.9	87.4	84.0
AccPts	88.6	87.9	88.0	89.3	89.1	85.5	93.7		89.1	87.1	93.7	86.7	86.9	92.7	87.1
Perim	89.0	86.5	90.2	89.0	86.4	85.1	91.6	93.2		85.0	86.4	94.4	85.4	85.9	95.6
<b>OSM</b>															
Cents	71.1	80.3	80.1	88.8	85.0	84.8	84.2	84.0	84.8		80.8	80.7	93.7	84.2	83.5
AccPts	81.1	76.2	74.5	90.1	96.9	85.7	81.5	84.9	82.2	74.9		87.4	83.5	95.1	85.2
Perim	74.9	79.6	81.3	82.0	80.7	80.6	75.7	74.7	77.2	82.0	80.9		81.1	86.7	94.9
<b>VMD</b>															
Cents	57.0	62.7	61.8	71.4	68.7	65.6	64.9	64.3	66.3	66.8	60.3	60.4		84.0	83.0
AccPts	55.7	64.0	63.0	70.4	71.3	67.3	66.9	68.4	66.6	66.7	58.8	61.7	78.7		89.3
Perim	59.1	65.9	67.5	71.3	67.6	66.5	68.4	68.5	70.1	66.9	57.1	69.8	67.7	72.1	

**Table 1** Destination overlap (using distances), comparing Cardiff and Vale Secondary Schools. Out of 105 comparisons, Vale had higher figures in 86 (82%). The 19 comparisons in which Cardiff figures were higher are highlighted for ease of identification.

Below diagonal - Overlap (%) for Cardiff schools. Above diagonal - Overlap (%) for Vale schools.

	Euclidean			ITN			UP			OSM			VMD		
	Cents	AccPt	Perim	Cents	AccPt	Perim	Cents	AccPt	Perim	Cents	AccPt	Perim	Cents	AccPt	Perim
<b>Euclid</b>															
Cents		148	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
AccPts	-1.445		< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
Perim	-13.084	-12.683		< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
<b>ITN</b>															
Cents	-13.358	-13.342	-11.001		874	< .001	< .001	020	< .001	452	167	< .001	266	189	< .001
AccPts	-13.331	-13.292	-10.872	-159		< .001	154	< .001	< .001	041	974	< .001	122	322	< .001
Perim	-15.158	-15.140	-14.682	-9.272	-7.610		< .001	< .001	< .001	< .001	< .001	935	< .001	< .001	< .001
<b>UP</b>															
Cents	-13.564	-13.529	-11.134	-4.202	-1.426	-8.613		009	< .001	070	138	< .001	< .001	< .001	< .001
AccPts	-13.189	-13.130	-11.082	-2.333	-3.615	-9.355	-2.611		< .001	003	< .001	< .001	< .001	< .001	< .001
Perim	-15.083	-15.063	-14.656	-8.004	-7.659	-3.871	-10.661	-12.052		< .001	< .001	581	< .001	< .001	491
<b>OSM</b>															
Cents	-13.131	-13.083	-11.097	-7.752	-2.046	-8.272	-1.814	-2.955	-6.980		027	< .001	945	573	< .001
AccPts	-13.252	-13.207	-11.040	-1.383	-0.033	-8.786	-1.481	-0.050	-7.871	-2.213		< .001	785	002	< .001
Perim	-15.126	-15.111	-14.690	-10.387	-9.698	-0.082	-10.225	-11.140	-5.51	-12.051	-11.169		< .001	< .001	089
<b>VMD</b>															
Cents	-14.803	-14.762	-13.962	-1.113	-1.546	-7.766	-4.907	-5.197	-6.264	-0.069	-2.273	-8.590		< .001	< .001
AccPts	-14.731	-14.712	-14.083	-1.312	-0.595	-8.108	-4.115	-4.158	-7.266	-0.564	-3.100	-9.338	-5.138		< .001
Perim	-15.547	-15.529	-15.318	-10.528	-9.935	-8.111	-10.312	-10.995	-0.889	-8.956	-10.796	-1.702	-6.486	-7.448	

**Table 2** Differences in E2SFCA scores between networks and location method for Vale Secondary Schools. Wilcoxon Z scores are shown below diagonal, statistical significance above diagonal (black = sig at <.001 level; green = sig at .01 level; amber = sig at .05 level; red = not significant at .05 level).

#### 4. Acknowledgements

The study reported here forms part of an Ordnance Survey-sponsored PhD research programme. The plug-in to ArcGIS™ was developed during Phase 1 of the WISERD project funded by the ESRC and HEFCW (ESRC Grant Reference: RES-576-25-0021). However any views expressed herein do not necessarily represent those of these organisations.

#### 5. Biography

Robin Frew is a final year PhD student at the University of South Wales.

Professor Gary Higgs is currently Director of the GIS Research Centre in the Faculty of Computing, Engineering and Science, University of South Wales and a co-Director of the Wales Institute of Social and Economic Research, Data and Methods (WISERD). Over-arching research interests are in the application of GIS in social and environmental studies, most recently in the areas of health geography and emergency planning.

Dr Mitch Langford is a Principal Lecturer in the Faculty of Computing, Engineering and Science, University of South Wales. His current research interests include dasymetric mapping, population modelling, and geospatial analysis within the fields of healthcare, social equality and environmental justice.

Dr Jenny Harding is a Principal Research Scientist at Ordnance Survey (GB) with particular interests in user focused research, geography and geographic data usability. Her role includes leading research in these areas both internally within Ordnance Survey and externally in collaborative projects with universities.

#### References

Avery S and Gittings B (2014). Do it yourself drones – experimenting with low cost UAVs. *GIS Professional*, October, 12-14.

Ellaway A, Macdonald L, Forsyth A, Macintyre S (2010). The socio-spatial distribution of alcohol outlets in Glasgow city. *Health & Place*, 16(1), 167-172.

Haklay M (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37, 682-703.

Higgs G, Fry R, Langford M (2012). Investigating the implications of using alternative GIS-based techniques to measure accessibility to green space. *Environment and Planning B: Planning and Design*, 39, 326-343.

Higgs G, Langford M and Norman P (2015). Accessibility to sport facilities in Wales: A GIS-based analysis of socio-economic variations in provision. *Geoforum*, 62, 105-120.

Jones S (2010). Open geographical data, visualisation and dissemination in public health information. *AGI Geocommunity '10*. Available at: <http://www.agi.org.uk/storage/geocommunity/presentations/SamuelJones.pdf> (Accessed 5 February 2014).

Zielstra D and Zipf A (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal. Available at [http://agile2010.dsi.uminho.pt/pen/shortpapers\\_pdf/142\\_doc.pdf](http://agile2010.dsi.uminho.pt/pen/shortpapers_pdf/142_doc.pdf) (Accessed 18 April 2013).

# Profiling Burglary in London using Geodemographics

C G Gale<sup>\*1</sup>, A D Singleton<sup>†2</sup> and P A Longley<sup>‡3</sup>

<sup>1</sup>UCL Department of Civil, Environmental & Geomatic Engineering

<sup>2</sup>University of Liverpool Department of Geography and Planning

<sup>3</sup>UCL Department of Geography

## Summary

A geodemographic classification provides categorical summary class assignments of neighbourhood areas based on salient population characteristics and built environment attributes. The regional London Output Area Classification (LOAC) is an example of such a classification, created using the same methodology as the national 2011 Output Area Classification. Police.uk data coded to LOAC provides an alternative perspective on burglary rates in London, with dwellings in different geodemographic clusters having experienced stark differences in the rate of burglaries. We conclude LOAC benefits from a greater predictive ability when compared to a national classification for differentiating socio-spatial structure, thus providing a more detailed insight into the variations of burglary across London.

**KEYWORDS:** Geodemographics, Burglary, Crime, London, OAC

## 1. Introduction

Geodemographic classifications are summary indicators of the social, economic, demographic and built characteristics of a small area zonal geography. They are designed to facilitate comparison between locations, for example, highlighting similarity in patterns of population structure between different parts of a country, or contrasting crime rates by coding Open Data sources. Within the UK, there is a lineage of freely available small area geodemographic classifications. The most recent example at the national level is the 2011 Output Area Classification, or 2011 OAC (ONS, 2014). This was built using 2011 UK Census data and output areas (OAs), the smallest spatial element of UK Census geography and primary unit of dissemination for the last two UK Censuses.

## 2. 2011 London Output Area Classification

A criticism of national classifications such as the 2011 OAC is that they do not adequately accommodate local or regional structures that diverge from national patterns. This is particularly evident in London where 85% of OAs belong to three of the eight 2011 OAC Super Groups. This has been deemed unsatisfactory by some users who only require a geodemographic perspective of London, such as the Greater London Authority (GLA). Consequently, the GLA commissioned the creation of the 2011 London Output Area Classification (LOAC). LOAC was created using the same inputs and methodology to the 2011 OAC, however, with a geographic extent limited to London. Further details of the classification are available from (Longley and Singleton, 2014).

## 3. LOAC and crime

The utility of LOAC as a tool for stratifying policy interventions within London can be explored using an illustrative example in the context of recorded crime. Data for individual recorded crimes have been made available on a monthly basis since December 2010 from the Police.uk website.

---

\* chris.gale@ucl.ac.uk

† alex.singleton@liverpool.ac.uk

‡ p.longley@ucl.ac.uk

Crimes are presented at a record level by crime category alongside a georeference of a location proximal to where the crime was recorded as occurring. In total 4 million crimes were recorded in London from December 2010 to July 2014.

The majority of recorded crimes are geocoded by police forces using a variety of methods such as tagging by mobile GPS receivers or address referencing. However, such data in raw form present a high risk to individual disclosure (Kounadi et al., 2014), and as such, it is not possible to publicly release crime data of this level of precision within a UK context given legislative constraints (Singleton and Brunsdon, 2014). As such, publically accessible crime data released through Police.uk are anonymised so that no individual crime event location is identifiable (Tompson et al., 2014). Crimes are allocated to a nearest centroid point of a pre-defined zonal geography (Tompson et al., 2014) which represent a collection of streets. This geography was created using Voronoi polygons drawn around street segment centroids and points of local relevance. To ensure privacy, polygons were merged if necessary to ensure each contained at least eight addresses (Singleton and Brunsdon, 2014; Tompson et al., 2014). Data made available by Police.uk uses these centroids as the recorded location of any crimes that occur within each polygon. As such, multiple crimes can be recorded at a single spatial location.

Such disclosure controls make the coding of crime events by the LOAC typology problematic, given uncertainty introduced by the intersecting tessellations of the Voronoi polygons and the 2011 OA geography. To understand the extent to which the aggregation of crime events impact LOAC cluster assignment, an estimation of the Voronoi zonal geography for London was created. Using all recorded crime events in London since December 2010, a total 94,667 Voronoi polygons were created, which represents an approximation of the zonal geography used to report crime by Police.uk (Singleton and Brunsdon, 2014). Figure 1 shows a subset of the Voronoi polygons and crime event centroids overload on LOAC, illustrating how the eight LOAC Super Groups intersect each polygon. The creation of this geography enabled the evaluation of two different methods of assigning crime events to LOAC Super Groups. The first method assigned a LOAC Super Group on the basis of the Output Area into which the recorded crime centroid fell. This has the advantage of being simple to compute, however, ignores that Voronoi polygons and the 2011 OA geography do not perfectly nest. As such, a second method was implemented that overlaid the 94,667 Voronoi polygons onto the boundaries of London's 25,053 OAs. The proportion of each OA within the Voronoi polygons were calculated. The total number of recorded crimes assigned to each Voronoi polygon were then re-assigned proportionally to each Output Area, and by association, each LOAC Super Group. For example, if 120 crime events were recorded in a Voronoi polygon which had 81% of its area belonging to 'London Life-Cycle' and 19% to 'Ageing City Fringe', then by multiplying these proportions ( $0.81 \times 120$  and  $0.19 \times 120$ ) it can be estimated 97 crimes might have occurred in 'London Life-Cycle' and 23 within 'Ageing City Fringe'. An assumption of this method is that crime distributions would be uniform across an OA, which in reality may not be the case given variable land use patterns.

The distribution of crimes across the eight LOAC Super Groups using these two allocation methods are shown in Table 1, and indicate few notable differences. As such, for the remaining analysis, the first method was implemented where crime centroids as supplied by Police.uk were used to assign LOAC clusters as this represented the least data manipulation and assumption about the geography of residential structure. The Police.uk website uses a number of different categories to classify crime. However, not all crime types would sensibly be profiled using LOAC, given that the typology relates to residential population characteristics. As such, a crime category that might be considered appropriate to explore is burglary, where offences would be more prevalent in residential locations (Maguire and Bennett, 1982). Calculating rates of burglary requires a denominator, and residential population is one of the most commonly used attributes (Andresen, 2006), however, established research has also shown using different denominators can impact observed patterns of crime (Boggs, 1965). Therefore, others have argued that using the number of occupied households may be considered as a preferable denominator for burglary rate calculations due to population movements during the working day and weekends (Harries, 1981; Ratcliffe, 2010).



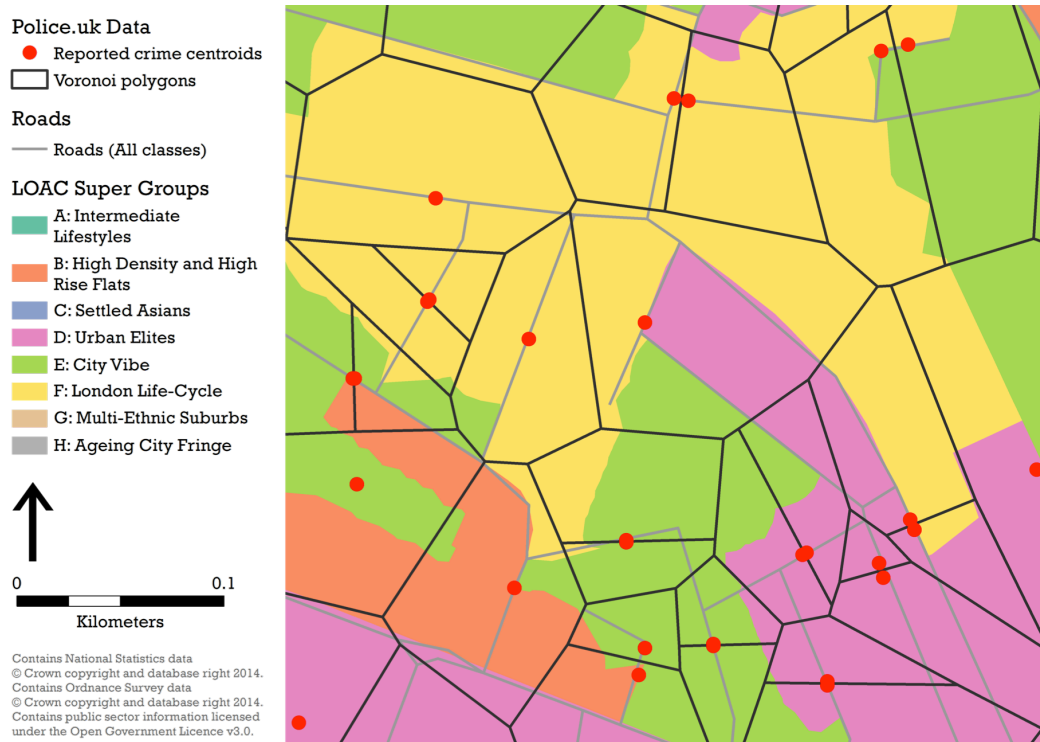


Figure 1: Police.uk Voronoi polygons and crime event centroids overlaid on the LOAC Super Groups

Table 1: Percentage of recorded crime assigned to LOAC Super Group using centroid location and the proportional assignment method

	Recorded crimes based on centroid locations	Recorded crimes based on proportional assignment	Difference
<b>A: Intermediate Lifestyles</b>	10.25%	10.26%	0.02%
<b>B: High Density and High Rise Flats</b>	13.88%	13.56%	-0.31%
<b>C: Settled Asians</b>	11.38%	11.09%	-0.28%
<b>D: Urban Elites</b>	16.36%	16.51%	0.16%
<b>E: City Vibe</b>	15.28%	15.48%	0.20%
<b>F: London Life-Cycle</b>	8.69%	8.52%	-0.18%
<b>G: Multi-Ethnic Suburbs</b>	18.34%	19.10%	0.76%
<b>H: Ageing City Fringe</b>	5.84%	5.47%	-0.36%

Burglary rates were calculated using a denominator of taxable dwellings, as a surrogate for the total number of households, in London in 2011, made available through the ONS Neighbourhood Statistics website. These are presented as index scores in Figure 2. These scores illustrate the propensity for burglaries to occur by LOAC clusters relative to the London average. A score of 100 is a rate the same as the London average of 98 burglaries being committed per 1,000 dwellings between December 2010 and July 2014, a score of 200 is twice the average, and 50 is a half. Dwellings within OA classified into the Super Groups ‘C: Settled Asians’, ‘D: Urban Elites’, ‘E: City Vibe’ and ‘G: Multi-Ethnic Suburbs’ LOAC Super Groups all have higher relative rates of burglary compared to the London average. For example, OA classified into the Super Group ‘C: Settled Asians’ exhibit burglaries at a rate 25% higher than the London average, whereas dwellings within ‘H: Ageing City Fringe’ have a likelihood of burglaries being committed at a rate 9% less than the London average. The ‘C: Settled Asians’ and ‘H: Ageing City Fringe’ LOAC Super Groups are both predominantly found in the outer Boroughs of London which are more suburban areas, yet display very large variation in burglary rates. A similarly divergent pattern also exists for clusters predominantly found within inner London, where the ‘D: Urban Elites’ and ‘E: City Vibe’ Super Groups have higher burglary rates than the London average, yet burglary in the ‘B: High Density and High Rise Flats’ cluster are 30% less likely.

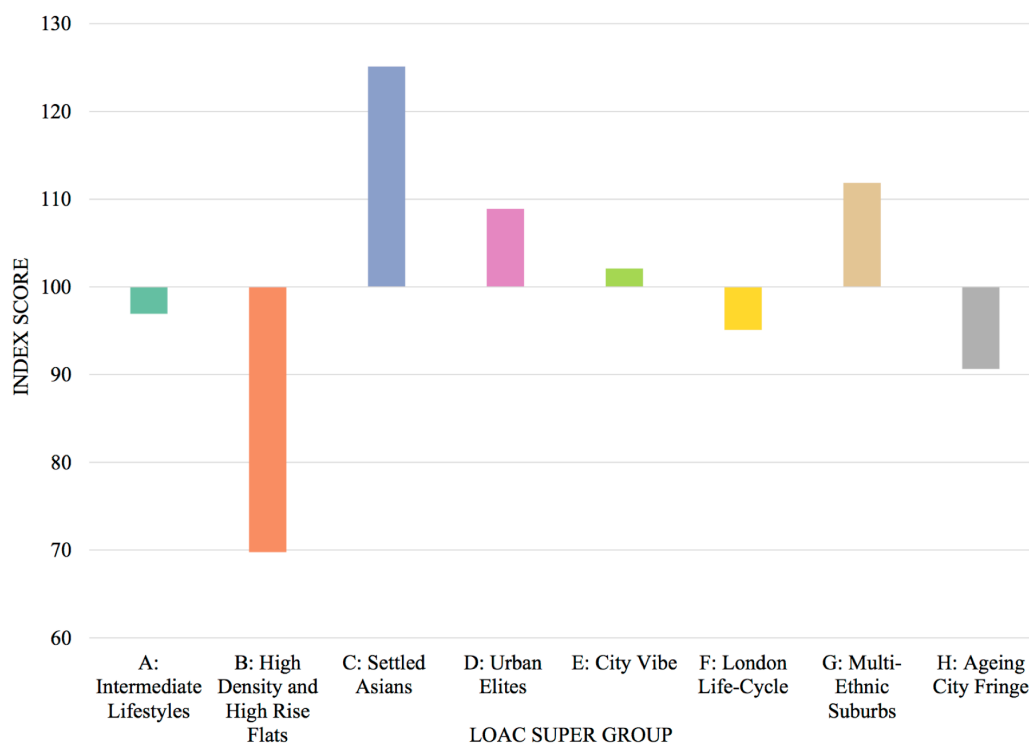


Figure 2: Index scores of burglary rates in LOAC Super Groups between December 2010 and July 2014 standardised using the total number of households

#### 4. Conclusion

The stratification of burglary rates using LOAC illustrates stark differences between each Super Group within London. The rate of burglaries range from being 25% higher than the London average in the ‘C: Settled Asians’ Super Group to 30% lower in ‘B: High Density and High Rise Flats’. The utility of LOAC for such public policy targeting applications highlights how regional classifications

can provide a level of detail not possible at the national level. More generally, the combination of demographic, socio-economic and physical conditions of London as represented by LOAC, rather than traditional administrative geographies, provides a unique perspective and method of summarising burglary statistics; thus providing a simplified way of comparing criminal activity across the socio-spatial variations found in London. Looking prospectively, the example of profiling crime data with LOAC is just one example of how it can be used to derive insight from the growing wealth of Open Data sources available for London.

## 5. Acknowledgements

This work was supported by EPSRC grants EP/J004197/1 (Crime, policing and citizenship (CPC) - space-time interactions of dynamic networks) and EP/J005266/1 (The uncertainty of identity: linking spatiotemporal information between virtual and real worlds) and ESRC grants ES/K004719/1 (Using secondary data to measure, monitor and visualise spatio-temporal uncertainties in geodemographics) and ES/L011840/1 (Retail Business Datasafe).

## 6. Biography

**Dr Chris Gale** is a Research Associate on the EPSRC funded Crime, Policing and Citizenship project at University College London.

**Dr Alex Singleton** is a Reader in Geographic Information Science at the University of Liverpool.

**Professor Paul Longley** is Professor of Geographic Information Science at University College London.

## References

- Andresen, M. A. (2006) 'Crime Measures and the Spatial Analysis of Criminal Activity', *British Journal of Criminology*, 46(2), pp. 258–285.
- Boggs, S. L. (1965) 'Urban Crime Patterns', *American Sociological Review*, 30(6), pp. 899–908.
- Harries, K. D. (1981) 'Alternative Denominators in Conventional Crime Rates', In Brantingham, P. J. and Brantingham, P. L. (eds.), *Environmental Criminology*, London, Sage, pp. 147–167.
- Kounadi, O., Bowers, K. and Leitner, M. (2014) 'Crime Mapping On-line: Public Perception of Privacy Issues', *European Journal on Criminal Policy and Research*, pp. 1–24.
- Longley, P. A. and Singleton, A. D. (2014) *London Output Area Classification: Final Report*, Greater London Authority, [online] Available from: <https://londondatastore-upload.s3.amazonaws.com/Vik%3D2011+LOAC+Report.pdf> (Accessed 13 December 2014).
- Maguire, M. and Bennett, T. (1982) *Burglary in a dwelling: the offense, the offender and the victim*, London, Heinemann Educational Books.
- ONS (2014) 'Methodology Note for the 2011 Area Classification for Output Areas', Office for National Statistics, [online] Available from: <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/methodology-and-variables/methodology.pdf> (Accessed 24 August 2014).
- Ratcliffe, J. (2010) 'Crime Mapping: Spatial and Temporal Challenges', In Piquero, A. R. and Weisburd, D. (eds.), *Handbook of Quantitative Criminology*, Springer New York, pp. 5–24.

- Singleton, A. and Brunsdon, C. (2014) 'Escaping the pushpin paradigm in geographic information science: (re)presenting national crime data', *Area*, 46(3), pp. 294–304.
- Tompson, L., Johnson, S., Ashby, M., Perkins, C. and Edwards, P. (2014) 'UK open source crime data: accuracy and possibilities for research', *Cartography and Geographic Information Science*, pp. 1–15.

# Assessment of social vulnerability under three flood scenarios using an open source vulnerability index

Garbutt K<sup>\*1</sup>, Ellul C<sup>†2</sup> and Fujiyama T<sup>‡2</sup>

<sup>1</sup> Centre for Urban Sustainability and Resilience, Department of Civil, Environmental and Geomatic Engineering, University College London

<sup>2</sup> Department of Civil, Environmental and Geomatic Engineering, University College London

February 18, 2015

## Summary

This paper utilises an open source flood vulnerability index to assess changes in social vulnerability within the English county of Norfolk under three flood scenarios. Open source demographics data are combined with flood zone data and GIS analysis of accessibility to key services to create the flood vulnerability index. The impact of flooding was found to be disproportionately distributed amongst those areas recording a high level of social vulnerability before flood risk was included. Analysis suggests those at risk of flooding are more likely to be elderly, poor and have long-term health problems.

**KEYWORDS:** Vulnerability, Open Source, Demographics, Flood Risk, Vulnerability Index

## 1. Introduction

Floods pose an environmental and fiscal challenge for the United Kingdom. More than five million homes and businesses are at risk of flooding (DEFRA, 2013) and an average of £1 billion in flood damages is incurred each year (EA, 2013). Flooding can have a far-reaching and long-lasting impact on a community (Bennet, 1970; Milojevic *et al.*, 2011). Those in society most often deemed vulnerable: the elderly, poor or unemployed, for example, often see their level of vulnerability increase during hazard events as both risk and exposure increases. The features of a person's life that makes them vulnerable are often intensified: the loss of income following a flood exacerbating poverty, for example. A greater knowledge of the spatial distribution of vulnerability within communities is therefore key to understanding how a population may be impacted by a hazard event (Cutter & Emrich, 2006). Highlighting those who are exposed to a hazard, as well as those who are potentially more vulnerable due to their circumstances, can aid emergency response and risk reduction strategies (Nelson, *et al.*, 2007).

This paper utilises an open source vulnerability index (OS-VI), previously created by the authors (Garbutt *et al.*, 2014), to assess changes in social vulnerability within the English county of Norfolk under three flood scenarios. Norfolk has a substantial coastline and a lengthy history of flooding, including being severely impacted by the 1953 North Sea Flood, as well as a substantial elderly (23%) and potentially vulnerable population (DCLG, 2011). The revised OS-VI presented herein, which incorporates updated demographics data recently released by the Office for National Statistics, provides a place-based assessment of vulnerability and when combined with flood risk scenarios can provide

---

\* k.garbutt.12@ucl.ac.uk

† c.ellul@ucl.ac.uk

‡ taku.fujiyama@ucl.ac.uk

non-governmental organisations (NGO) and local authorities with valuable context and guidance when planning for or responding to flood emergencies.

## 2. Background

The British Red Cross (BRC) works throughout the UK and internationally preparing communities for disasters, supporting post-disaster recovery, delivering and teaching first aid and assisting individuals with health and social care needs. Since 2010 the BRC has operated a Mapping Team that utilises Geographical Information Systems (GIS) and the growing wealth of spatial data to support the organization's work. For the BRC, vulnerability analysis and mapping provides context to hazards and provides information that can guide service delivery and the provision of community resilience building programs.

The development of vulnerability indices is an increasingly common method to capture the geographical distribution of vulnerability across regions, countries or sub-national areas (see: Cutter *et al.*, 2003; Johnson *et al.*, 2012). Vulnerability indices are regularly used, in conjunction with needs assessments and on-the-ground research, to target service provision and justify resource allocation (Flanagan *et al.*, 2011). Further, such indices are increasingly being coupled with hazard assessments to create integrated risk analyses and feed into emergency response strategies (COES, 2010; Dunning & Durden, 2011; Siagian, *et al.*, 2013). To produce such practical analysis and assessment of vulnerability, many socio-economic elements that influence vulnerability must be assigned measureable numeric indicators (Atteslander *et al.*, 2008). However, past work on measuring and mapping vulnerability has been limited by a focus on income-related indicators, a lack of consideration of accessibility, the production of large resolution indices (county or country scale), and the reliance on proprietary data and/or methodologies, with limited attention paid to open source data (Garbutt *et al.*, 2014).

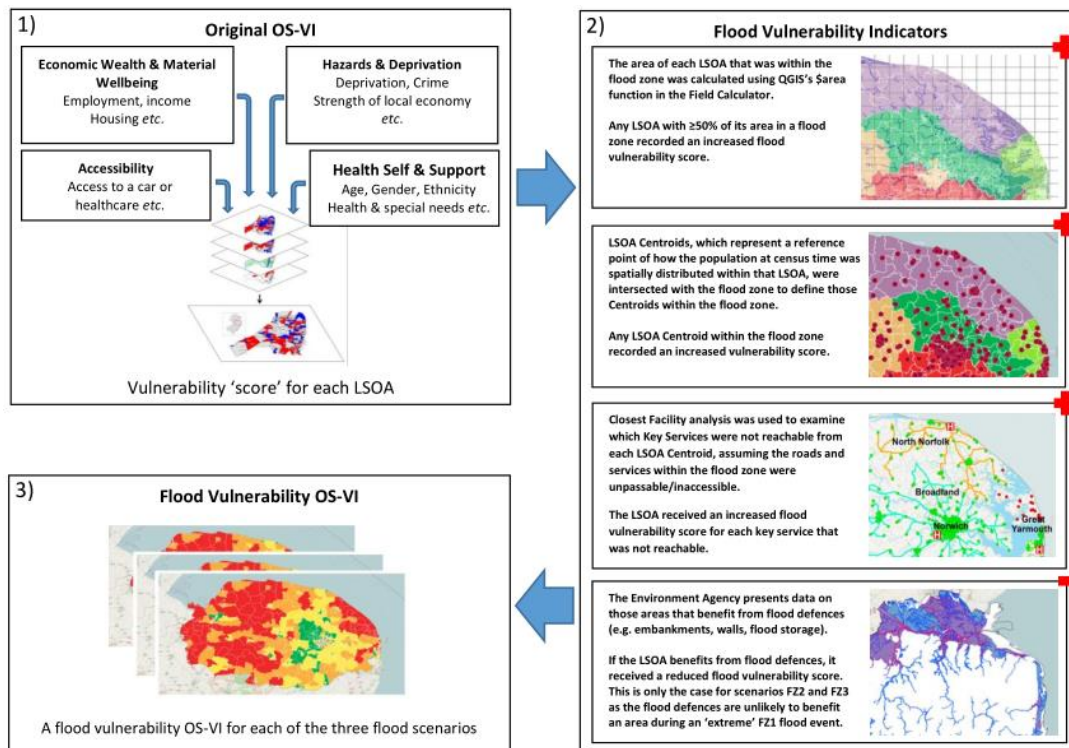
## 3. Methodology

The authors previously created an Open Source Vulnerability Index (OS-VI), a deductive indicator-based index that incorporated 53 different vulnerability indicators shown to influence vulnerability (Garbutt *et al.*, 2014). Unlike many vulnerability indices, the OS-VI was produced using open source demographic data only and incorporated indicators of flood risk as well as accessibility and the loss of capabilities and access to key services (health facilities and food stores) when determining an area's level of social vulnerability. The OS-VI provided a mechanism whereby quality open source data on the core drivers of vulnerability could be used to create a vulnerability index with a sufficiently small resolution to examine vulnerability at the community level. Here, the OS-VI is used as a starting point to examine the impact of three flood scenarios (presented in table 1) and is supplemented with spatial analysis to generate four flood vulnerability indicators for each of the 539 Lower Layer Super Output Areas (LSOA) within Norfolk (see figure 1).

**Table 1** Flood Scenarios (adapted from Environment Agency, 2014)

Scenario	Probability
Very Low (FZ1)	Land assessed as having less than 1 in 1,000 (0.1%) annual probability of flooding. This scenario is described as an 'extreme flood'.
Low (FZ2)	Land assessed as having between a 1 in 200 (0.5%) annual probability of flooding.
Moderate (FZ3)	Land assessed as having a 1 in 100 (1%) or greater annual probability of flooding.

The results of the four flood vulnerability indicators are combined with the social vulnerability indicators within the original OS-VI to produce an overall vulnerability score for each LSOA in Norfolk. The results are mapped and the changes in vulnerability examined.



**Figure 1** Flow Chart outlining the methodology used to develop the flood vulnerability OS-VI presented herein. 1) Use of original OS-VI; 2) production of flood vulnerability indicators; 3) combination of both to create a Flood Vulnerability OS-VI for each flood scenario under analysis.

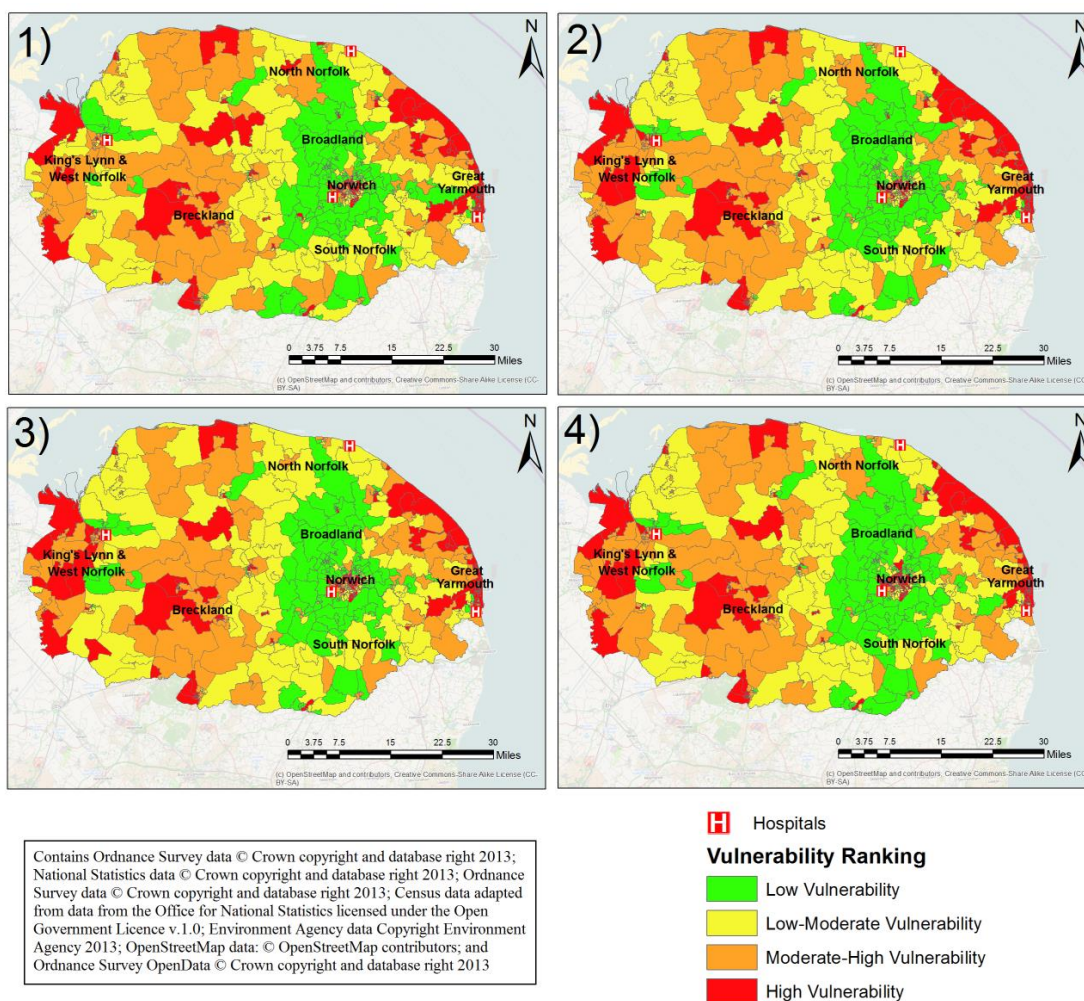
#### 4. Results

As can be seen in figures 2 and 3, the major clusters of LSOA with high vulnerability ratings across the original OS-VI and all three flood scenarios loosely match those areas with greater exposure to the flood hazard. Although changes are slight, vulnerability ratings of areas already deemed highly vulnerable within the original OS-VI were found to increase as the impact and extent of flooding increased across the three flood scenarios examined. The impact of flooding was found to be disproportionately distributed amongst those LSOA recording a *high* or *moderate-high* vulnerability rating: roughly 66% of LSOA with the majority of their area within the flood zone and 76% with their Centroid within a flood zone recorded a *high* or *moderate-high* vulnerability rating.

Of those highlighted as at risk of flooding, our analysis suggests residents are also more likely to live alone and be aged 65+; have an income below the national median; lack central heating in their home; have bad or very bad health and limited actions due to a long-term health problem/disability; provide in excess of 50 hours care to another per week. An underlying relationship between the presence of flood hazard in an area and its socio-economic and health vulnerabilities is therefore suggested, although the relationship is unclear and further study is needed.

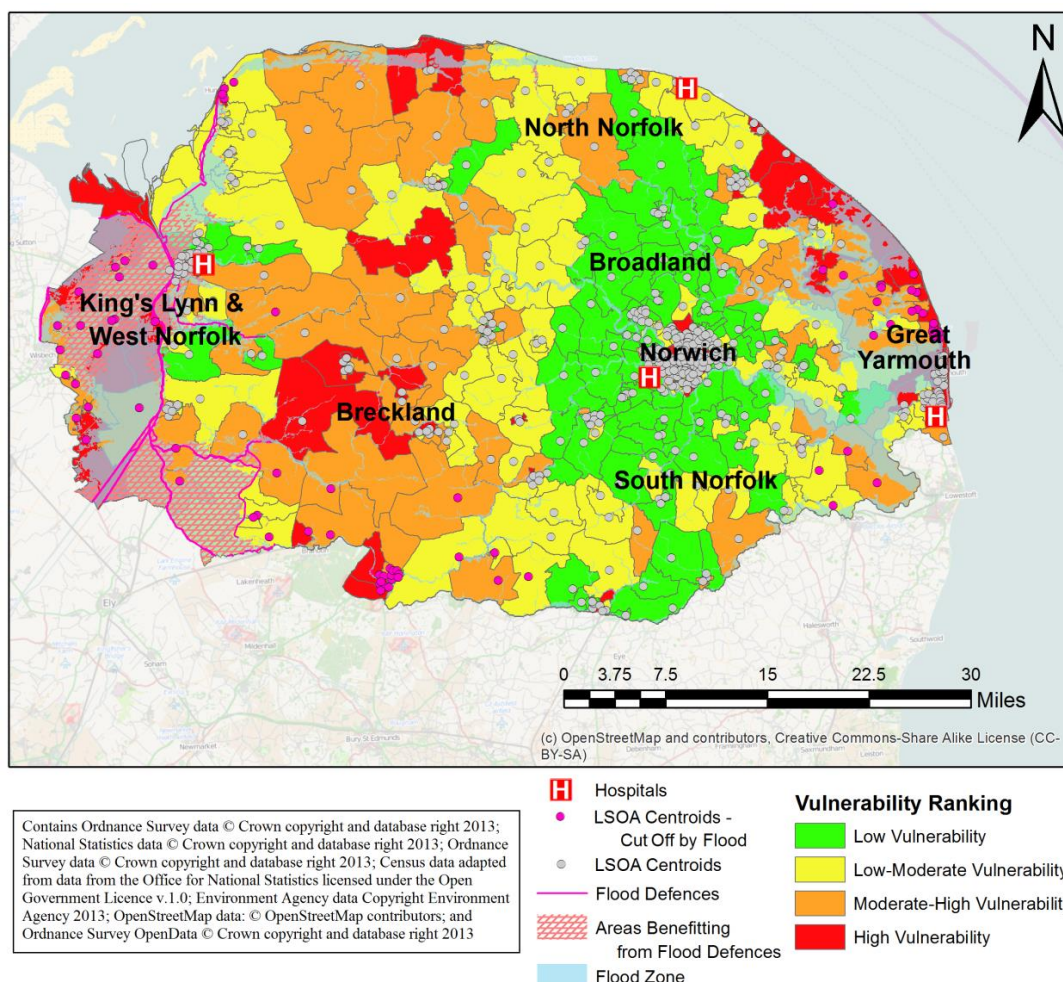
In addition, analysis of changes in accessibility found that those populations highlighted above are also more likely to live in an area where travel time to key services is in excess of the national average. Travel time to key services was severely impacted under all flood scenarios. For example, under scenario FZ1, the local authority of Kings Lynn recorded a 223% increase in average travel time, from 21 minutes to 68 minutes, and the local authority of Great Yarmouth, which was identified as being the most vulnerable local authority in Norfolk as well as the most affected by flooding, recorded 43% of its LSOA as completely cut-off from healthcare facilities.





**Figure 2** Vulnerability Indices. 1) Original OS-VI; 2) Vulnerability Index under FZ1 scenario; 3) Vulnerability Index under FZ2 scenario; 4) Vulnerability Index under FZ3 scenario.





**Figure 3** Vulnerability Index under FZ3 'extreme flood' scenario featuring flood data overlay.

## 5. Discussion

The results presented above demonstrate the potential cascading impact of a flood hazard as it impacts an already vulnerable population: exacerbating pre-existing vulnerabilities, limiting capabilities and restricting accessibility and access to key services. The OS-VI and its use here to examine the potential impact of flooding was found to be a useful supplementary tool for the BRC, with the measurement and visualisation of vulnerability and potential hazard impact providing context and aiding response service planning and capability assessment.

To ensure its use by the BRC, open data and software was used exclusively to reduce organizational expenditure and provide quality data that can be freely disseminated. The demographic data used was deemed the best available at the scale and resolution required and the GIS software used, QGIS, was also deemed appropriate for the analysis required. One limitation noted was the assumption that, under each flood scenario, all roads within the flood zone would be impassable and thus roads and key services within each zone were restricted. This was due to the lack of flood depth information within the available flood data. Future work would use topographical data and digital elevation models to calculate flood depth and improve the accessibility metrics within the indices.

It was also noted that the discrete county boundary used to restrict analysis to within the county of

Norfolk likely led to limitations within accessibility routing as key services outside of the study zone may have been a viable option if the restriction was not in place. At a county level the analysis remains valid, but for organisations like the BRC who regularly work across boundaries, future work will address this scale effect by examining service locality and the catchment zones of neighbouring services that are utilised by populations within the study zone.

## **6. Conclusion**

The approach presented here utilises vulnerability indices built upon quality open data to examine the potential impact of flooding on social vulnerability. The original OS-VI takes into account a broad range of social and economic indicators to highlight hotspots of social vulnerability. The addition of varying hazard scenarios allows for the examination of how vulnerability changes during an emergency. The indices produced can be extended to the national level whilst retaining the relatively small LSOA resolution. The methods used are scalable and adaptable and the use of open data allows all parties involved to easily coordinate and share information, potentially improving local knowledge and reducing vulnerability (Trujillo *et al.* 2000). The OS-VI and its use here to examine flood risk represents the first step in imagining a dynamic and customisable disaster risk platform that can be updated as new data is made available and adapted and utilised by the BRC and others to identify pockets of vulnerable communities and improve emergency response.

## **7. Biography**

Kurtis Garbutt is a postgraduate research engineer at the Centre for Urban Sustainability & Resilience at University College London. Kurtis received a full EPSRC scholarship to work with the British Red Cross on development of a GIS tool to support the work of NGOs and improve understanding of urban resilience.

Claire Ellul is a lecturer and course tutor at UCL. Claire's research focuses on spatial data management and infrastructures, in particular the creation, maintenance and use of metadata, as well as big data performance optimization and the use of topology in GIS, in particular 3D GIS.

Taku Fujiyama is a lecturer and leader of the Resilience Research Group at UCL. Taku's research focuses on the resilience of infrastructure, primarily transport systems, as well as the understanding of user behaviour of transport environments, particularly amongst the most vulnerable, and the design and operation of railway infrastructure.

## **8. Acknowledgements**

This work was supported by the Economic and Social Research Council and represents part of an EngD project currently being undertaken at University College London. The authors thank Andrew Braye and the Mapping Team at the British Red Cross for guidance. Portions of this study have been presented in an article currently under review by the journal *Environmental Hazards: Human and Policy Dimensions*. Data was provided by Ordnance Survey under ©Crown copyright and database right 2014; Office for National Statistics licensed under the Open Government Licence v.1.0; Environment Agency under Copyright Environment Agency 2014; and OpenStreetMap under ©OpenStreetMap contributors.

## **References**

- Bennet G., 1970. Bristol floods 1968. Controlled survey of effects on health of local community disaster. *Br Med J*, 3, 454-458.
- COES (California Office of Emergency Services). 2010. State of California multi-hazard mitigation plan. Sacramento: California Office of Emergency Services.
- Cutter, S.L., Boruff, B.J. & Shirley, W.L., 2003. Social vulnerability to environmental hazards. *Social Science Quarterly*, 84, pp.242-261.
- Cutter, S.L. & Emrich, C.T., 2006. Moral Hazard, Social Catastrophe: The Changing Face of

- Vulnerability along the Hurricane Coasts. *Annals of the American Academy of Political and Social Science*, 604(1), p.110.
- DCLG, 2011 (Department for Communities and Local Government). *English Indices of Deprivation 2010*.
- DEFRA, 2013 (Department for Environment Food & Rural Affairs). Reducing the threats of flooding and coastal change. Available at: <https://www.gov.uk/government/policies/reducing-the-threats-of-flooding-and-coastal-change>.
- Dunning, M.C., and Durden, S. 2011. Social vulnerability analysis methods for corps planning. Washington, DC: U.S. Army Corps of Engineers, Institute for Water Resources.
- EA (Environment Agency), 2013. Investing in the Future: Flood and Coastal Flood Risk Management in England – A long-term investment plan. Available at: <http://www.environment-agency.gov.uk/research/library/publications/108673.aspx>.
- EA (Environment Agency), 2014 Flood Map for Planning (from Rivers and the Sea). Available at: <http://apps.environment-agency.gov.uk/wiyby/37837.aspx>
- Flanagan, B.E., Gregory, E.W., Hallisey, E.J., Gerd, J.L.H- & Lewis, B. 2011. A social vulnerability index for disaster management. *Journal of Homeland Security & Emergency Management*. 8(1), pp. 3-27.
- Garbutt, K., Ellul, C. & Fujiyama, T. 2014 Mapping Social Vulnerability to Flood Hazard in Norfolk, England *Environmental Hazards: Human and Policy Dimensions* (under review).
- Johnson, D.P., Stanforth, A., Lulla, V. & Lubet, G. 2012. Developing an applied extreme heat vulnerability index utilizing socioeconomic and environmental data. *Applied Geography*, 35(1-2), pp.23–31. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S014362281200032X> [Accessed March 17, 2013].
- Milojevic, A., Armstrong, B., Kovats, S., Butler, B., Hayes, E., Leonardi, G., & Wilkinson, P. 2011. Long-term effects of flooding on mortality in England and Wales, 1994-2005: controlled interrupted time-series analysis. *Environ Health*, 10(1), 11.
- Nelson, C., Lurie, N., Wasserman, J., Zakowski, S., & Leuschner, K. J. (2007). Conceptualizing and defining public health emergency preparedness. RAND Working Paper – May 2008. National Emergency Training Center.
- Trujillo, M., Ordonez, A. & Hernandez, C. (2000) Risk mapping and local capacities: lessons from Mexico and Central America. Oxfam.
- Siagian, T.H., Purihadi, P., Suhartono, S. & Ritonga, H. 2013. Social vulnerability to natural hazards in Indonesia: driving factors and policy implications. *Natural Hazards*, 70(2), pp.1603–1617.

# **MAPPING OF SPATIAL DISTRIBUTION OF TUBERCULOSIS CASES IN KEBBI STATE, NIGERIA**

**2008-2011**

**Usman Lawal Gulma**  
[Usmangulma38@yahoo.com](mailto:Usmangulma38@yahoo.com)  
**Department of Geography,**  
**Adamu Augie College of Education, Argungu,**  
**Kebbi State, Nigeria.**

## **ABSTRACT**

The World Health Organization has declared tuberculosis a global emergency in 1993. It has been estimated that one third of the world population is infected with Mycobacterium tuberculosis, the causative agent of tuberculosis. The identification of clusters in space-time is of great interest in epidemiological studies. The objective of this paper was to identify and map the spatial distribution of Tuberculosis during the period 2008-2011 in Kebbi State. Kernel Density Analysis tool in ArcGIS Spatial Analyst was employed to map the trend of TB cases over the period. The results revealed that the highest occurrence of 2,220 cases was in 2009 while year 2011 recorded the least cases of 1179 across the state. It was further revealed that Birnin Kebbi LGA with a population of 268,620 recorded the highest cases of 1,639. However, Suru LGA with population of 148,474 recorded only 70 cases over the period. In conclusion, TB cases were unevenly distributed in the state but high cluster rates were identified in the four emirate headquarters of Gwandu, Yauri, Argungu and Zuru. Increasing the number of diagnostic and treatment centers were recommended in order to reduce the number of cases across the state.

**Key Words:** Tuberculosis, ArcGIS, Population, Cluster, Kernel Density

## 1. Introduction

Tuberculosis (TB) is an infectious disease caused by the bacillus *Mycobacterium tuberculosis* and spreads through air by a person suffering from TB. The 1990 World Health Organization (WHO) report on the Global Burden of Disease ranked TB as the seventh most morbidity-causing disease in the world, and expected it to continue in the same position up to 2020 [WHO, 1996]. In 2001, the WHO estimated that 1.86 billion persons (32% of the world population) were infected with TB. Each year, 8.74 million people develop TB and nearly 2 million die. This means that someone somewhere contracts TB every four seconds and one of them dies every 10 seconds [Dye et al, 1999]. TB is one of the oldest human diseases that still affect large population groups, mainly in marginal areas and comprising vulnerable groups impacted by extreme poverty, malnutrition, and crowded housing. These groups are prone to infection by the tuberculosis bacilli and to acquiring active TB Baker et al (2011).

The World Health Organization (WHO) reported in 2010 that there were an estimated 9.4 million incident cases (range 8.9 million–9.9 million) of TB globally, equivalent to 137 cases per 100,000 populations, and that 1.1 million of those cases also tested positive for human immunodeficiency virus (HIV). The mortality of HIV-negative patients with TB was estimated at 1.3 million, this being equivalent to 20 deaths per 100,000 people. The incidence of TB patients in Asia was 55% and 30% in Africa; smaller proportions of cases occurred in the Eastern Mediterranean Region (7%), the European Region (4%), and the Americas Regions (3%) (WHO, 2010).

At present, geographic information systems (GISs) are among the most useful tools in epidemiology, as they can be used to identify geographical areas and population groups with a higher risk of sickness or premature mortality and which therefore require higher preventive care or health information and monitoring of diseases in time and space.

In the case of TB, various researchers have used GIS to study this infectious disease. Moonan et al. [2004] used GIS to identify the geographic locations of TB transmission and incidence in the United States of America during 1993 to 2000. In India, Tiwarin et al. [2006] carried out a geospatial investigation of TB occurrence in the Almora district using GIS and the SCAN statistics program. Nunes [2007] in Portugal detected spatial and temporal clusters during 2000–2004 by using SCAN. The above-mentioned authors agree that GIS and SCAN are useful tools for vigilance against TB.

According to United States embassy in Nigeria (2010), Nigeria ranked 10<sup>th</sup> among the 22 high burden TB countries of the world. Kebbi state, the study area ranked 17<sup>th</sup> out of 36 states of Nigeria within the same period. This trend is alarming considering the fact that millennium development goal (MDG) is set to achieve the reduction of TB cases to half by the year 2015.

## 1.2 Objectives

The main purpose of this paper was:

1. To analyse the spatial distribution of tuberculosis (TB) cases by area in Kebbi state, Nigeria over a four-year period (1998-2011), using geographical information systems (GIS) technique.

2. To demonstrate capability of GIS in disease mapping and surveillance.  
It is expected that the results obtained by this study would assist policy makers in decision making.

## 2. Methodology

Density analysis takes known quantities of some phenomena and spreads it across the landscape based on the quantity that is measured at each location and the spatial relationship of the locations of the measured quantities.

Kernel Density calculates the density of point features around each output raster cell. In other words, it calculates a magnitude per unit area from point or polyline features using a kernel function to fit a smoothly tapered surface to each point or polyline.

Conceptually, a smoothly curved surface is fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the Search radius distance from the point. Only a circular neighborhood is possible. The volume under the surface equals the Population field value for the point, or one if NONE is specified. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center. The kernel function is based on the quadratic kernel function described in Silverman (1986).

## 3. Result

The results of the analysis revealed that the prevalence of TB cases in the study area could be attributed to high population densities per square kilometer. This is evident in the 2008 analyzed map which shows Zuru local government area which is the 4<sup>th</sup> in terms of number of cases but has the highest spatial density of the disease. This demonstrated the need for the employment of GIS for TB analysis in order to map and highlight the most affected areas for timely intervention and decision making.

LGA	2008 CASES	2009 CASES	2010 CASES	2011 CASES	TOTAL CASES	POPULATION
ALIERO	70	87	64	43	264	67078
AREWA	60	68	32	110	270	189728
ARGUNGU	318	268	141	191	918	200248
AUGIE	42	23	21	19	105	116368
BAGUDO	74	96	35	61	266	238014
B/KEBBI	356	474	295	514	1639	268620
BUNZA	42	28	10	20	100	123547
DANDI	71	72	26	59	228	146211
DANKO/WASAGU	63	54	20	74	211	265271
FAKAI	62	32	21	29	144	119772
GWANDU	42	52	25	45	164	151077
JEGA	126	106	61	138	431	197757
KALGO	17	19	12	25	73	84928
KOKO/BESSE	168	141	58	95	462	154818

MAIYAMA	9	33	17	30	89	173759
NGASKI	49	69	25	52	195	126102
SAKABA	59	28	15	21	123	91728
SHANGA	22	45	46	76	189	127142
SURU	25	17	11	17	70	148474
YAURI	269	272	138	254	933	100564
ZURU	259	236	106	197	798	165335

Table 1: TB Cases (Source National TB Control Programme Kebbi State)

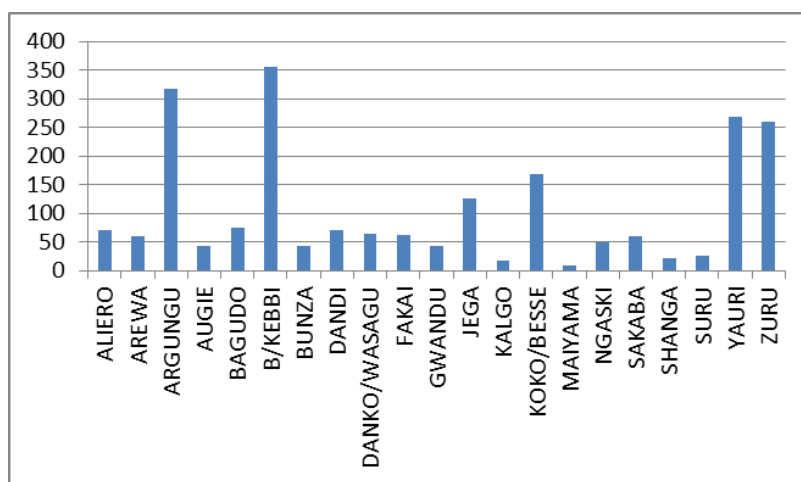


Figure 1: Graph of TB cases 2008

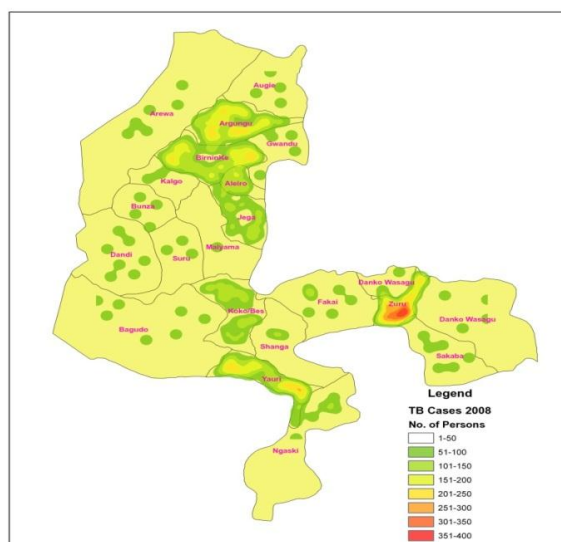


Figure 2: Map of Kebbi state showing density of TB Cases 2008

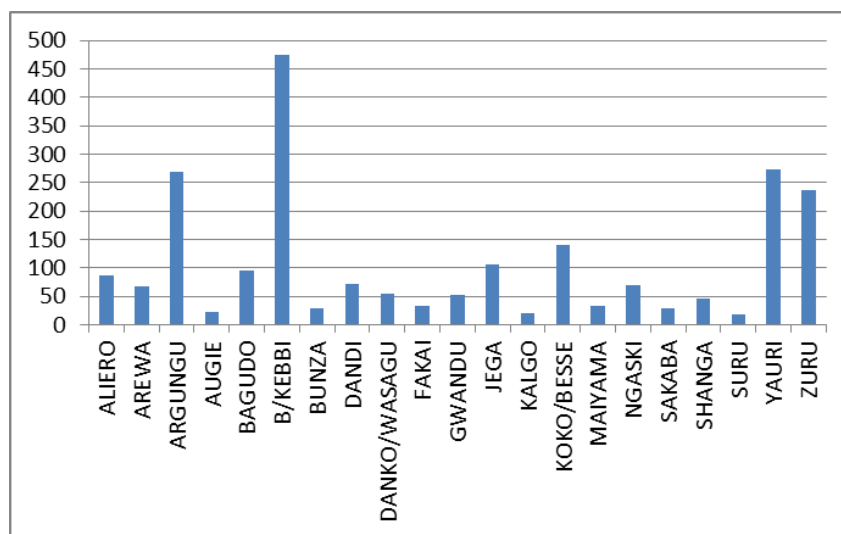


Figure 3: Graph of TB cases 2009

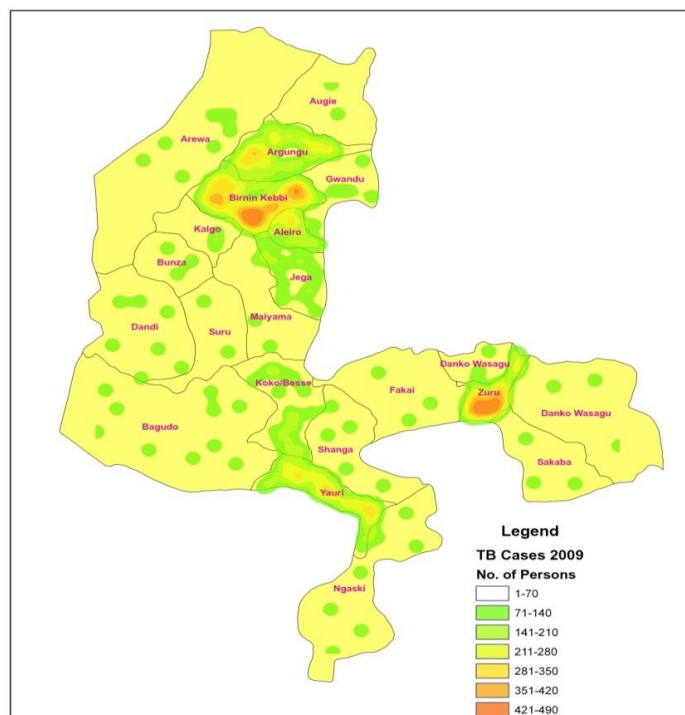


Figure 4: Map of Kebbi state showing density of TB Cases 2009



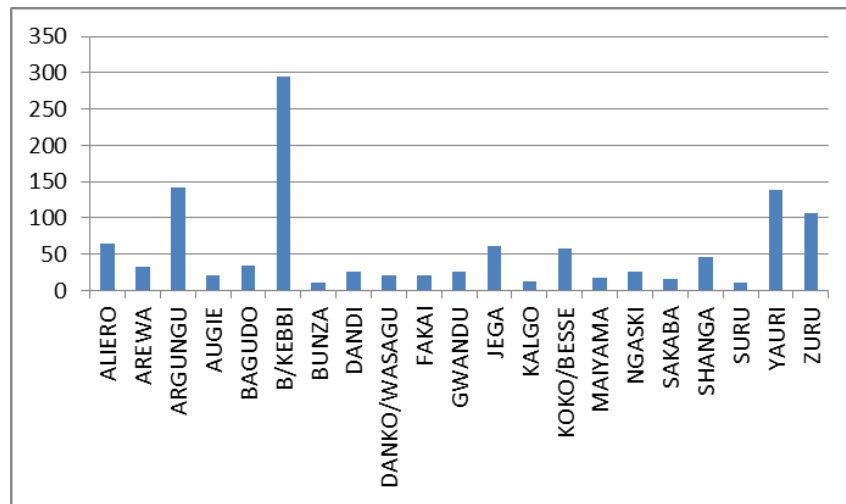


Figure 5: Graph of TB cases 2010

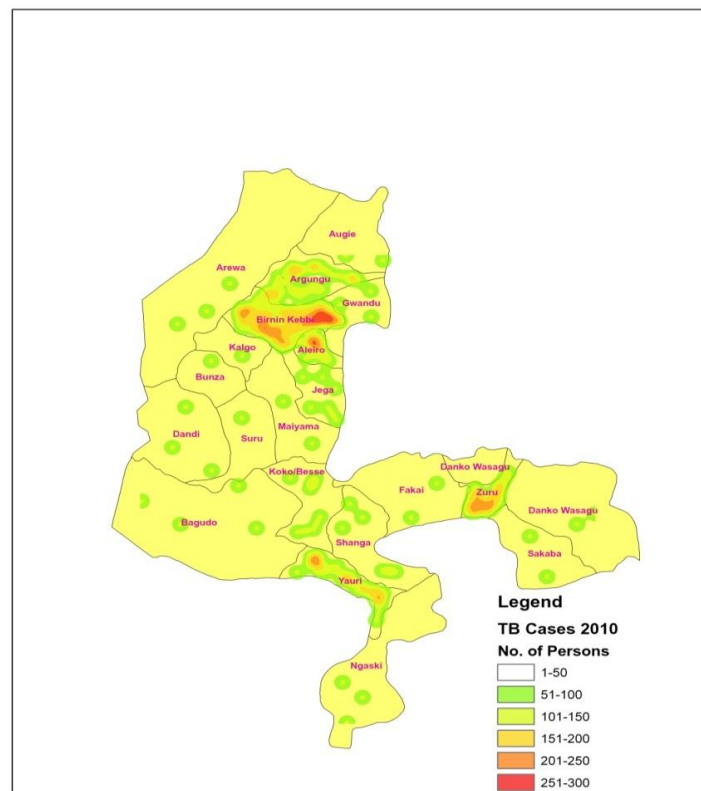


Figure 6: Map of Kebbi state showing density of TB Cases 2010

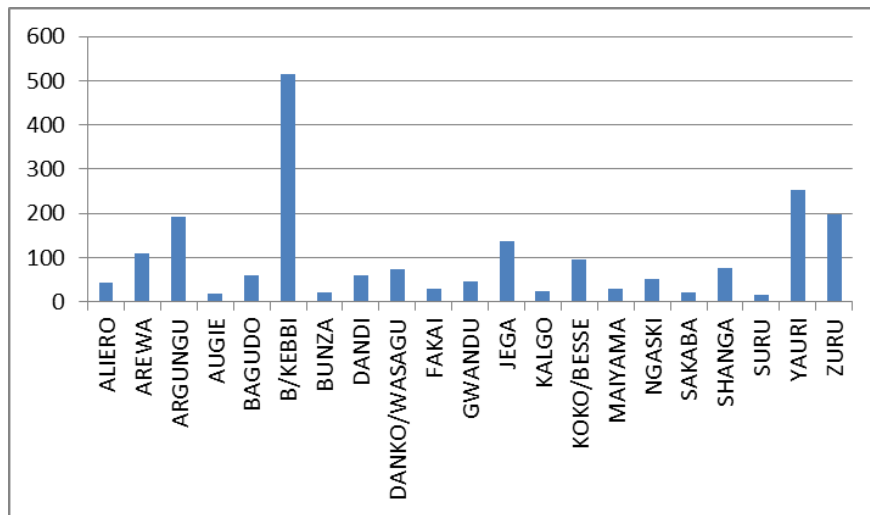


Figure 7: Graph of TB cases 2011

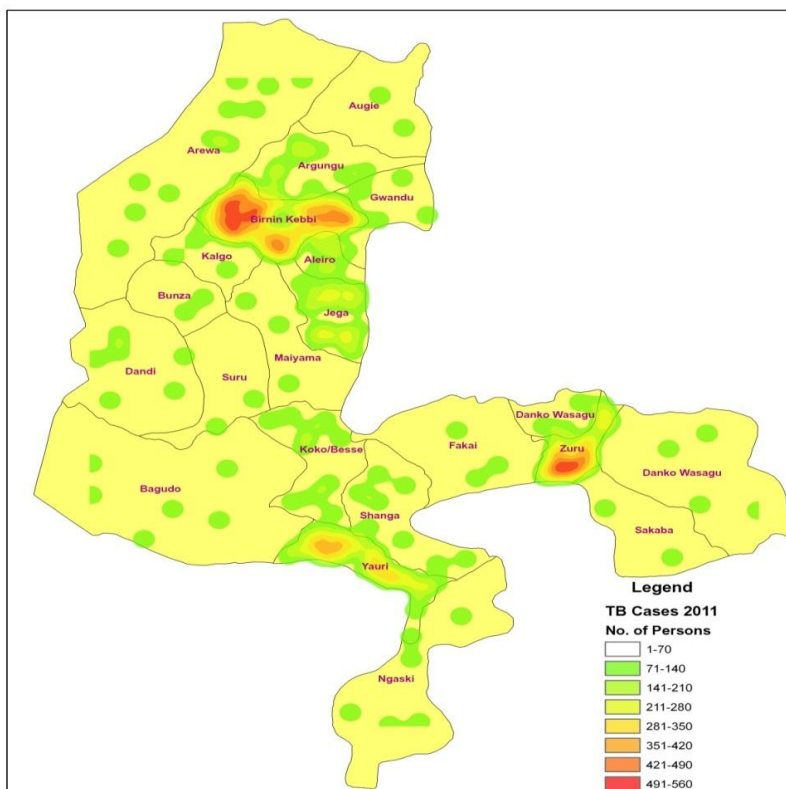


Figure 8: Map of Kebbi state showing density of TB Cases 2011

**Acknowledgement**

The author wishes to acknowledge the profound contribution of the state coordinator of national tuberculosis control programme in Kebbi state, Nigeria for providing data for this research. I equally acknowledged the contributions of the authors whose works were reviewed in this paper. I thank you very much.

**Biography of the Author**

Usman Lawal Gulma is a lecturer in Geography at Adamu Augie College of Education, Argungu, Kebbi state, Nigeria. He holds a Bsc degree (Geography) and a Master of (GIS) all from the Usmanu Danfodiyo University, Sokoto, Nigeria. He is also a member of international research and development institute. He has a number of publications in national and international journals to his credit. The author is currently a research postgraduate student in the school of Geography, University of Leeds, United Kingdom and he is interested in spatial analysis.

## References

- B W Silverman (1986), *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.
- C Nunes (2007), *Tuberculosis incidence in Portugal: Spatio-temporal clustering* International Journal of Health Geographics Vol. 6 No. 30.
- Dye C, Scheele S, Dolin P, Pathania V, Raviglione M (1999), *Global Burden of Disease estimated incidence, prevalence, and mortality by country*. Journal of Medical Association 1999, 282:677-686.
- M G Baker, K Venugopal, and C P Howden (2011), *Household Crowding and Tuberculosis*, World Health Organization Regional Office for Europe, Copenhagen, Denmark.
- N Tiwari, C M Adhikari, A Tewari and V Kandpal (2006), *Investigation of Geo-spatial hotspot for the occurrence of Tuberculosis in Almora District* International Journal of Health Geographics VOL.5 NO.33.
- P K Moonan, M Bayona and T N Quitugua (2004), *Using GIS Technology to identify areas of tuberculosis transmission* International Journal of Health Geographics Vol. 3 No. 23.
- United States Embassy in Nigeria (2012), Nigeria Tuberculosis Facts Sheet <http://Nigeria.usembassy.gov> retrieved on 26/01/2014.
- World Health Organization (2010), *Global Tuberculosis Control*, WHO Report, Geneva, Switzerland.

# The Influence of Familiarity on Route Choice: Edinburgh as a Case Study

Maud van Haeren<sup>\*</sup> and William Mackaness<sup>†</sup>

The University of Edinburgh School of Geosciences

November 5, 2014

## Summary

Automatically generated routing instructions are provided by Satnav and Internet based mapping services in order to assist us in getting to unfamiliar places. Instructions from these devices are based around least cost algorithms, described on a street- by- street basis. Taking no account of what we might already know, the instructions are long, difficult to remember and require effort to interpret. If we could opportunistically route the person via known areas, the recognition process would be easier, the instructions could be fewer, and the users would find greater comfort in travelling through spaces familiar to them. In this paper we model a user's heterogeneous familiarity of the city such that it modifies a cost surface, resulting in directions that route the user via familiar spaces. A familiarity index was created based on historical GPS based trajectories. Participant route choice was found to be closer to outputs from the model than simple shortest path.

**KEYWORDS:** Familiarity, Shortest Path, Cost Surfaces, Navigation, Pedestrian Wayfinding.

## 1. Introduction

This research aimed to measure familiarity and its influence on route choice made by pedestrians in the City of Edinburgh. Increasingly, satnav devices and smartphones are used to assist in getting to unfamiliar places (Savage et al., 2011; Schmid, 2008; Zandbergen and Barbeau, 2011). The street by street information given by these devices can appear counterintuitive and contain information that is hard to remember. Yet people do not take the shortest path. Various factors might influence their choice; it is certainly known that people like to take advantage of places they know, as this requires less cognitive effort and feels easier to navigate (Demirbas, 2001; Schmid et al., 2010; Gale et al., 1990; Lovelace and Hegarty, 1999; Li, 2006; Papinski et al., 2009; Pahlavani and Delavar, 2014). Regardless of the complexity or simplicity of the environment people are habitual, preferring to retrace routes they travelled before or know rather than exploring new ones (Golledge, 1999). The aim of this research was to improve navigational solutions for pedestrians in urban environments by taking into account familiarity of their environment. This in turn, required us to: 1) create a quantitative measure of familiarity and 2) optimally incorporate this within a shortest- path algorithm. So how might we measure familiarity and how well can such a model predict such route choices? Here we present the model, and comment on its predictive capacity which was evaluated through quantitative methods and street level experiments.

---

<sup>\*</sup> [s1363550@ed-alumni.net](mailto:s1363550@ed-alumni.net)

<sup>†</sup> [william.mackaness@ed.ac.uk](mailto:william.mackaness@ed.ac.uk)

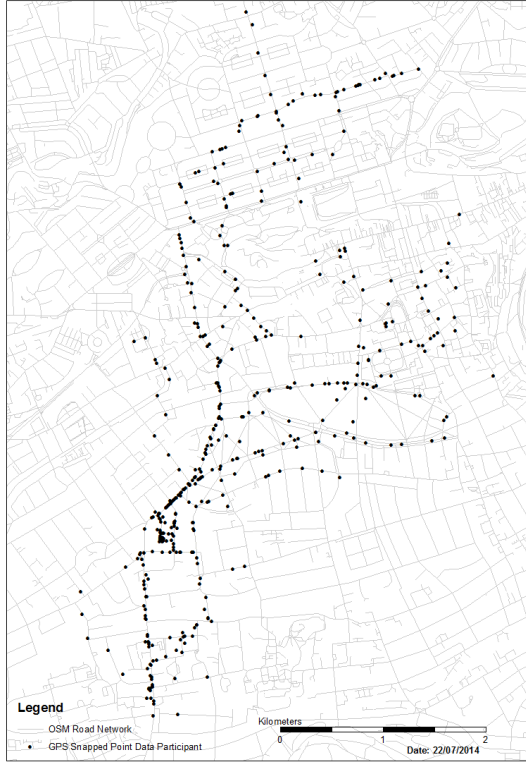
## 2. Methodology

Spatial familiarity can be quantified as a product of revisit times along routes (Pahlavani and Delavar, 2014; Gale et al., 1990). Additionally other studies have looked at using the duration of staying in a certain place as a qualifier to a familiarity index (Meness and Moreira, 2007; Schmid and Richter, 2006). An index based on revisit times of road segments is relatively easy to implement based on the assumption that the more the individual travels along a route, the stronger the representation of the important elements of the environment within their cognitive map (Golledge, 1999; Montello, 1993).

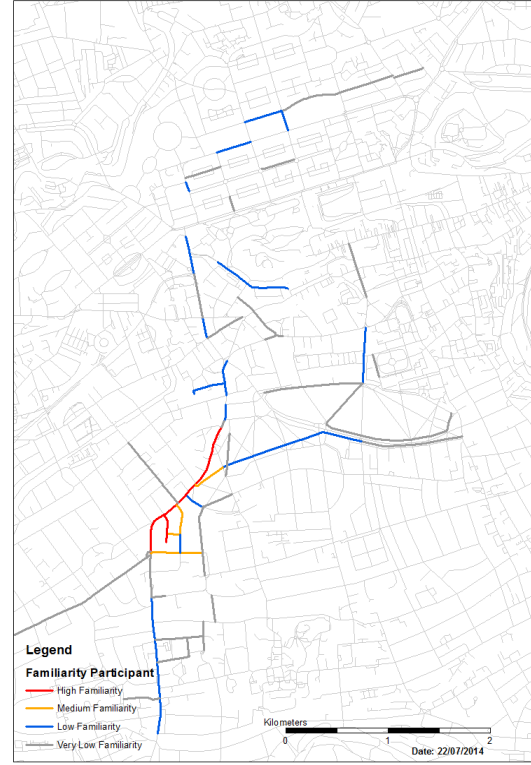


**Figure 1** Research study area – City of Edinburgh, UK

In order to quantify familiarity, information about people's whereabouts was collected to create spatial user profiles. The research was focused solely on walking pedestrians. All participants lived and worked in the city of Edinburgh (Figure 1). Ten participants agreed to record their GPS trajectory data and this information was used to identify familiar streets. GPS data was recorded via smartphones which were considered to be sufficiently accurate in capturing location (Zandbergen and Barbeau, 2011; von Watzdorf and Michahelles, 2010). The familiarity index was based on a minimum of seven days' worth of travel. Additionally the participant was asked to participate in lab based experiments following data collection in order to record preferred routes on maps, and for follow up interviews. GPS measurements were extracted using Google's Location History functionality (Google, 2014). Due to changes in sampling rate, WiFi black spots and urban canyoning, some of the familiarity maps were rather patchy in nature (Joshi, 2001; Meness and Moreira, 2007; Ochieng et al., 2003; von Waltzdorf and Michahelles, 2010; Zandbergen and Barbeau, 2011). Figure 2 is an example of a participant's history with its corresponding familiarity network based on revisit times in Figure 3.



**Figure 2** Patchy GPS Coverage Participant 1



**Figure 3** Mapping Familiarity to the Network

When computing just travel distance, Dijkstra’s algorithm is a popular choice for calculating the shortest distance between two points in a network (Wise, 2002; Worboys and Duckham, 2004; Sonnier, 2006; Lloyd, 2010). Here we focus on **two** criteria: *travel distance* and *familiarity*. Thus it is a conflicting bi- criterion network problem (Gen and Lin, 2005) – solving for a minimum travel distance and a maximum of familiarity. There are several approaches to solving a multi- criteria path optimization. A traditional way of solving this optimization (and the approach taken here) was first, to linearly weight all independent criteria so that they result in one value for each edge of the network and then second, solve by using Dijkstra’s method (Corley and Moon, 1985; Malczewski, 2011). Street level data from OpenStreetMap – OSM (OpenStreetMap, 2014) was used to create the underlying network serving as input in the Network Analyst tool. OSM was chosen because, in urban contexts, it is both topologically accurate and more complete than Ordnance Survey’s ITN data (Haklay, 2008).

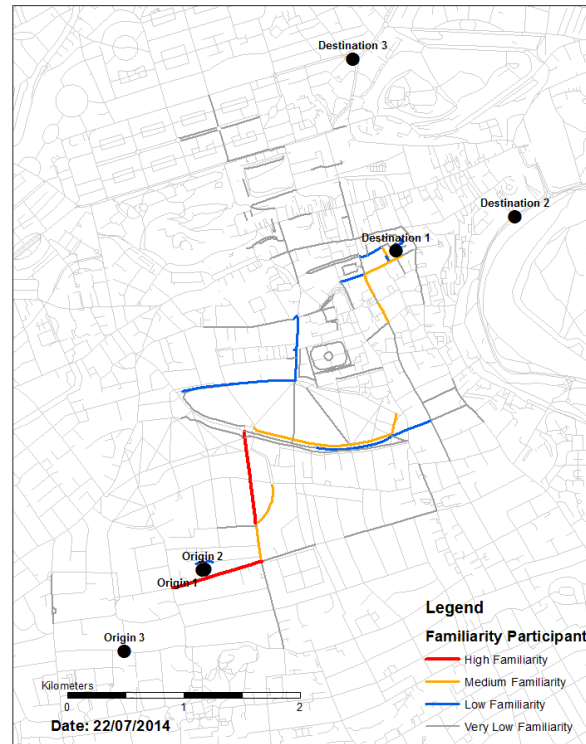
GPS date stamps were used to count the number of days a participant had revisited a specific street by grouping them based on date of measurement and counting the number of different dates. This count was used to create a Familiarity Index (FI), Equation 1, which is then multiplied by the value of the underlying street network – in this case the length of the road in meters (Corley and Moon, 1985; Malczewski, 2011; Dijkstra, 1959; Ochieng et al., 2003). Thus the cost of the street was reduced, in effect making previously visited edges (streets) more ‘attractive’ as compared with unvisited/ less visited streets. To the best of knowledge of the authors, defining a Familiarity Index in this manner has not been done before. The street network with these newly calculated total edge weights (Figure 3) served as the basis for network analysis and the application of Dijkstra’s Algorithm (Dijkstra, 1959; van Haeren, 2014).

$$FI = 1 - \left( \frac{\text{Revisit Days Count}}{\text{Total Number Tracked Days}} \right) \quad (1)$$



### 3. Evaluation

A route that was far from familiar regions of the city would not be expected to make large deviations in order to pass through that familiar region; a route falling entirely within a region well known to the pedestrian, might be strongly influenced by their degree of knowledge of it. Therefore to validate the accuracy of the model, origin and destination pairs were chosen that variously crisscrossed familiar parts of the network in order to assess how the paths deviated from shortest path, i.e. taking account of the familiar. Based on visual examination of their GPS- trajectory data, three types of routes were distinguished: 1. Familiar place to familiar place; 2. Familiar place to unfamiliar place; and 3. Unfamiliar place to unfamiliar place.

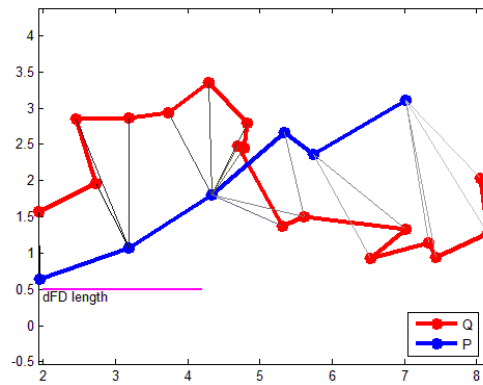


**Figure 4** Choosing Origin and Destination Points

To evaluate the veracity of the outputs from the model, participants were asked to draw their own preferred routes on paper. Participants were then asked open ended questions as to their reasons for drawing a specific route (Papinski et al., 2009; Gale et al., 1990). For each of the three different origin – destination situations, three routes per participant were generated: 1) SP (shortest path): using Dijkstra’s algorithm to solve for edge weights of distance only, 2) SPF (shortest path with familiarity): using Dijkstra’s algorithm to solve for a combined edge weight of distance and familiarity and 3) HC (human choice) route as drawn by the participants on paper maps.

In order to quantitatively compare the different outputs, differences in length and differences in the Discrete Fréchet Distance were measured. The Fréchet Distance measures resemblance between lines and takes into account the course of the line (Alt and Godau, 1995; van Haeren, 2014). Figure 5 shows two hypothetical lines and the corresponding Discrete Fréchet Distance. Summation of these distances enables comparison between the three outputs (Danziger, 2011; Eiter and Mannilla, 1994).

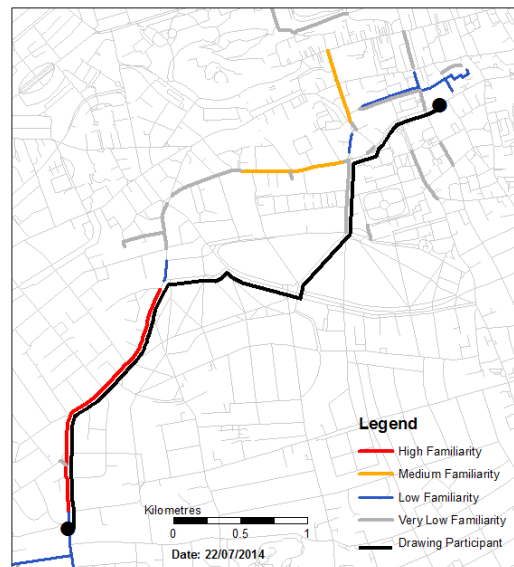




**Figure 5** Discrete Fréchet Distance of hypothetical lines P & Q

#### 4. Refinement of the Methodology

Visually comparing the drawn routes against the GPS trajectory data of each individual revealed that people did not always choose routes that took maximum advantage of their familiarity with an area. Figure 6 shows the drawn route takes a detour from what might be expected based on measured familiarity. This could be because 1) there are factors other than familiarity influencing choice, 2) that the model of familiarity was deficient, or 3) that people say one thing, and do another! Golledge (1999) and Montello (1993) both suggest that learning of the environment takes place at the eye level perspective - travel experience gained at the 'environmental scale' (Montello, 1993) whereas learning takes place via the examination of layouts - the 'figural scale' (Montello) using photographs or maps; and that these are different learning processes. Choice patterns and decision processes are not identical at these different perspectives or spatial scales, due to factors such as time and effort (Montello, 1993). This appears to suggest that asking people to draw on paper maps is not the best way of validating the outputs from the model. Therefore a second small sample experiment was set up to improve on evaluation of the outputs. Three of the ten participants were asked to walk from a known origin to a known destination and these were then compared with other outputs.



**Figure 6** Comparing a participant's familiarity map with what they drew (in the mid-section, one would expect the black line to follow the familiarity path more closely)

## 5. Results

In order to search for patterns, the results of these comparisons were grouped based on the type of origin - destination combination (Situation 1, 2 and 3). Table 1 shows the averaged results for the ten participants in percentage length difference: Column 1 indicates that they are willing to walk 15% longer distances as compared with the shortest path in order to remain in familiar territory.

**Table 1** Length Differences in Percentages

	SP & SPF	SP & HC	SPF & HC	Situation Average
Situation 1	104%	101%	100%	101%
Situation 2	120%	112%	93%	108%
Situation 3	122%	106%	102%	110%
Average	115%	107%	98%	-
	<i>SPF longer</i>	<i>HC longer</i>	<i>SPF longer</i>	

Table 2 shows the average Discrete Fréchet Distance in meters for the ten participants. The smallest Fréchet Distance, and therefore the smallest shape difference, can be seen when comparing the shortest path with the path calculated based on the Familiarity Index (column 1).

**Table 2** Average Discrete Fréchet Distance in Metres

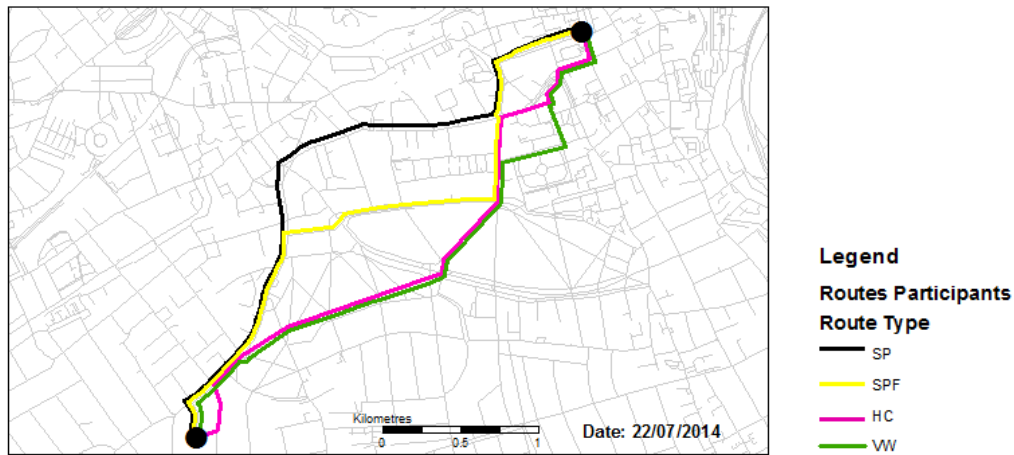
	SP & SPF	SP & HC	SPF & HC	Situation Average
Situation 1	1147	1432	1870	1483
Situation 2	868	513	809	730
Situation 3	707	793	842	781
Average	907	913	1174	-

From the interviews, three ideas emerged as to why participants did not take the shortest path. Firstly, people prefer walking through green spaces as much as possible. Secondly, people take routes which “feel” simple and require less thinking, or are directly related to familiarity. Thirdly, people try to avoid things like busy streets or steep gradients. One could of course argue that all three factors are implicitly reflected in the historical trajectories of the participants. These results furthermore suggest that “cognitive ease” of following familiar routes might not be the only reason as to why certain route choices are made and that many more factors influence this decision making process (Penn, 2001).

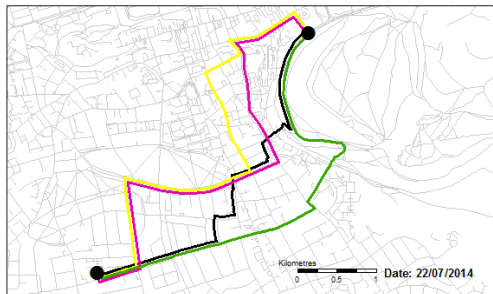
Three participants completed a Validation Walk (VW): first walking from a familiar origin to a familiar destination (Situation 1) based on recollection from memory, and then walking from the same familiar destination back to a familiar origin based on what was calculated by Dijkstra’s algorithm based on shortest distance in metres only. Figure 7, 8 and 9 show the results of the VW compared to the earlier calculated and drawn routes for the three participants: none of the participants has the exact same VW as HC. Table 3 shows the averaged results for the VW. The difference in absolute length is a bit less for the Validation Walk, meaning that the SPF predicts length better. The Discrete Fréchet Distance also shows that the Validation Walk is more similar to the prediction of the SPF than the route drawn by the participants (HC).

**Table 3** Results comparison Validation Walk and Human Choice (experiment 2)

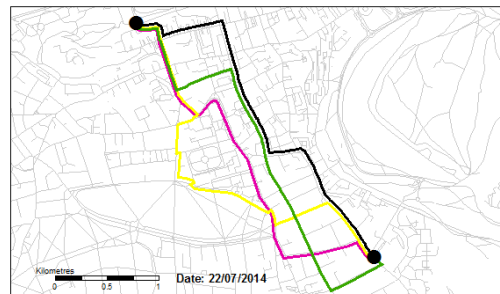
	SPF & HC	SPF & VW
Length Comparison (%)	102%	99%
Discrete Fréchet Distance (m)	714	588



**Figure 7** Participant 1 Validation Walks



**Figure 8** Participant 3 Validation Walks



**Figure 9** Participant 2 Validation Walks

## 5.1 Privacy Considerations

Undertaking this research brought into focus issues of privacy – as much can be inferred from both an individuals’ digital footprint, and an individuals’ data shadow (information others have generated about them) (Koops 2011; Kuebler et al. 2013). People’s tolerance for privacy invasion varies with culture and age (Thomas et al. 2013), and people’s understanding of how such information is shared (Post and Woodrow 2008). Privacy issues are of increasing concern, given the increasing precision with which LBS devices record location, and the ability to infer activities based on trajectory analysis (Android 2014; Hanson 2005).

## 6. Conclusion

The research has demonstrated that familiarity can be quantified and used as an edge weight multiplier to modify outputs from Dijkstra’s Shortest Path Algorithm. The ‘Shortest Path Familiarity’ predicted the actual route more accurately when looking at the comparison of the Validation Walk cases versus the Human Choice cases. From the results, the following conclusions can be drawn 1) People are willing to travel 15% further if it means moving through familiar territory; 2) People see the world differently through maps as compared to the world they experience.

The validity of these conclusions is restricted to the constraints in the data set, sample size and other limitations (van Haeren 2014). To further explore these conclusions additional research is recommended: 1) development of a richer model of familiarity that incorporates data collection over a longer period of time, and gives priority to recency of visitation; 2) deeper understanding of the ‘in filling’ process by which people ‘connect’ familiar spaces together and how this influences their route choice; and 3) Inclusion of motive for choice of route (social dimension, time of day, urgency, and satisfying other ‘on route’ tasks).

## 7. Acknowledgements

We would like to acknowledge all ten participants who have dedicated their time and data to this research.

## 8. Biography

Maud van Haeren recently completed the MSc GIS at The University of Edinburgh and is now living and working in The Netherlands. William Mackaness is a senior lecturer at The University of Edinburgh in the School of GeoSciences.

## 9. References

- Alt H and Godeau M (1995). Computing the Fréchet Distance Between Two Polygonal Curves. *International Journal of Computational Geometry & Applications*, 5(1&2), 75 – 91.
- Android (2014) ‘Google Android Location APIs, [online], available: <http://developer.android.com/google/play-services/location.html>
- Corley H W and Moon I D (1985). Shortest Path in Networks with Vector Weights. *Journal of Optimization Theory and Applications*, 46(1).
- Danziger Z (2011). Discrete Fréchet Distance Calculations for MATLAB.
- Demirbas G U D (2001). Spatial Familiarity as a Dimension of Wayfinding. Unpublished thesis, Bilkent University.
- Dijkstra E W (1959). A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1, 269 – 721.
- Gale N, Golledge R G, Halperin W C, Couclelis H (1990). Exploring Spatial Familiarity, *The Professional Geographer*, 42(3), 299 – 313.
- Gen M and Lin L (2005). Multiobjective Hybrid Genetic Algorithms for Bicriteria Network Design Problem. *Graduate School of Information, Production & Systems Waseda University Japan*.
- Golledge R G (1999). Wayfinding Behaviour: Cognitive Mapping and Other Spatial Processes, Baltimore, Maryland USA: The Johns Hopkins University Press.
- Google (2014). Google Location History [online], available: <http://maps.google.com/locationhistory> [accessed 02/05/2014].
- Haklay M M (2008). How Good is OpenStreetMap information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the rest of England. *Under review in Environment & Planning B*.
- Hanson, J. (2005) ‘Mobile Data Device and Method of Locating Mobile Data Device’. Google Patents [online], available: <http://www.google.co.uk/patents/US20050153681> [accessed 02/05/2014].
- Joshi R R (2001). A New Approach to Map Matching for In- Vehicle Navigation Systems: The Rotational Variation Metric. *Translated by Oakland (CA), USA*.

- Koops, B. J. (2011) 'Forgetting Footprints, Shunning Shadows. A Critical Analysis of the "Right to be Forgotten" in Big Data Practice', *SCRIPTed*, 8(3), 229-256.
- Kuebler, K., Palm, D. and Slavec, A. (2014) 'The Ethics of Personal Privacy and Location Based Services' in *Confronting Information Ethics in the New Millenium*.
- Lloyd C D (2010). Spatial Data Analysis: An Introduction for GIS users. Oxford: Oxford University Press.
- Lovelace K L and Hegarty M (1999). Elements of Good Route Directions in Familiar and Unfamiliar Environments. *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, Berlin: Springer, 65 – 82.
- Malczewski J (2011). Local Weighted Linear Combination. *Transactions in GIS*, 15(4), 439 – 455.
- MATLAB R2013a (2013).
- Meness F and Moreira A (2007). Using GSM CellID Positioning for Place Discovering. Portugal: University of Minho, Department of Information Systems.
- Montello D R (1993). Scale and Multiple Psychologies of Space. *Spatial Information Theory: A Theoretical Basis for GIS*, Berlin: Springer- Verlag, 312 – 321.
- Ochieng W Y, Quddus M, and Noland R B (2003). Map- Matching in Complex Urban Road Networks. *Revista Brasileira de Cartografia*, 55(02).
- OpenStreetMap (2014). OpenStreetMap, [online], available: <http://www.openstreetmap.org/> [accessed 02/05/2014].
- Pahlavani P and Delavar M R (2014). Multi- Criteria Route Planning Based on Driver's Preferences in Multi- Criteria Route Selection. *Transportation Research Part C: Emerging Technologies*, 40, 14 – 35.
- Papinski D and Scott D M (2011). A GIS- based Toolkit for Route Choice Analysis. *Journal of Transport Geography*, 19(3), 434 – 442.
- Papinski D, Scott D M and Doherty S T (2009). Exploring the Route Choice Decision- Making Process: A Comparison of Planned and Observed Routes Obtained using Person- Based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(4), 347 – 358.
- Penn, A. (2001). Space Syntax and Spatial Cognition. *Proceedings 3<sup>rd</sup> International Space Syntax Symposium Atlanta*.
- Savage N S, Baranski M, Chavez M E and Hollerer T (2011). I'm Feeling LoCo: A Location Based Context Aware Recommendation System. Translated by Vienna: Springer.
- Savage N S, Chun W, Chavez M E and Hollerer T (2012). Seems Familiar: An Algorithm for Inferring Spatial Familiarity Automatically. *Santa Barbara, California: Computer Science Department*, University of California.
- Schmid F (2008). Knowledge- Based Wayfinding Maps for Small Display Cartography. *Journal of Location Based Services*, 2(1), 57 – 83.
- Schmid F (2010). Personal Wayfinding Assistance. *Dissertation zur Erlangung des Grades eines Doktors der Ingenieurwissenschaften*, Bremen: Universitat Bremen.

- Schmid F and Richter K F (2006). Extracting Places from Locational Data Streams. *UbiGIS 2006 – Second International Workshop on Ubiquitous Geographical Information Services*.
- Sonnier D L (2006). Parallel Algorithms for Multicriteria Shortest Path Problems.
- Thomas, L., Little, L., Briggs, P., McInnes, L., Jones, E. and Nicholson, J. (2013) ‘Location Tracking: views from the older adult population’, *Age Ageing*, 42(6), 758 – 63.
- van Haeren M (2014). Technical Report Part II: A Personalised Routing Algorithm. *Unpublished thesis (MSc Geographical Information Science)*, The University of Edinburgh.
- von Watzdorf S and Michahelles F (2010). Accuracy of Positioning Data on Smartphones. *LocWeb*.
- Wise S (2002). GIS Basics. London: Taylor & Francis.
- Worboys M and Duckham M (2004). Fundamental Spatial Concepts. *GIS: A Computing Perspective*, second ed. London: Boca Raton, 426.
- Zandbergen P A and Barbeau S J (2011). Positional Accuracy of Assisted GPS Data from High Sensitivity GPS- enabled Mobile Phones. *Journal of Navigation*, 64(03), 381 – 399.

# Real time coupled network failure modelling and visualisation

Harris N<sup>1</sup>, Robson C<sup>1</sup>, Barr S<sup>1</sup> and James P<sup>1</sup>

<sup>1</sup> Newcastle University, School of Civil Engineering and Geoscience, Newcastle-upon-Tyne, NE1 7RU

November 11, 2014

## Summary

This paper, presents an approach to real-time spatio-temporal analysis of infrastructure network performance by developing an open source geovisualisation tool coupled with infrastructure network failure models in order to simulate, visualise and analyse how spatial infrastructure networks respond over time to major perturbations and failures.

**KEYWORDS:** Network, geovisualisation, failure modelling, real-time simulation

## 1 Introduction

Exploring how geospatial networks behave in near real time is becoming increasingly relevant to a number of sectors and a range of different applications, including transport planning and for infrastructure systems engineering to understand potential hazards (Leu et al., 2010). The spatial distribution, time dynamics and attributes of the flows on such networks leads to an information “richness” that makes analysing network behaviour a valuable exercise (Tominski et al., 2012). However, understanding the spatio-temporal flows across a network is not straight forward using traditional static graphs and diagrams as they fail to capture the intrinsic dynamics of these systems (Adrienko and Adrienko, 2011). Geovisualisation technologies offers considerable potential in this regard, not only for the single point-in-time assessment of network performance, but also in terms of recognising and understanding the often hidden patterns of network evolution over time and space that lead to particular configurations of interest.

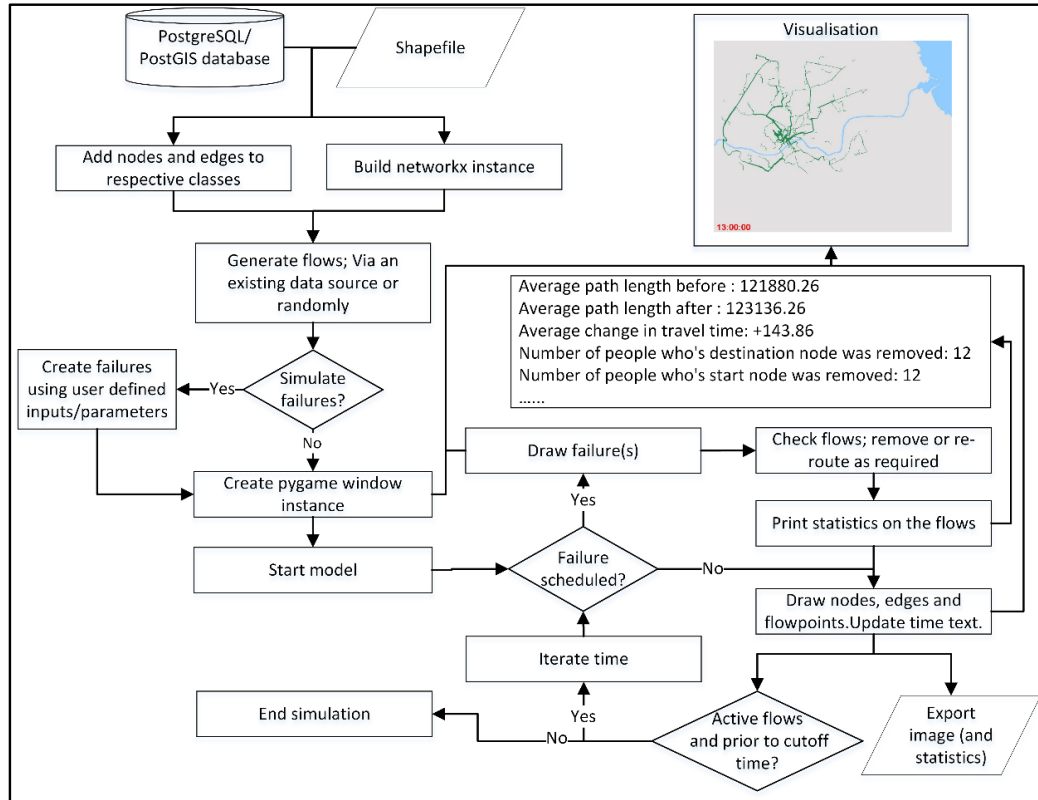
Real-time geovisualisation of network dynamics is particularly important in assessing the vulnerability of critical spatial infrastructure networks to perturbations. While significant progress has been made in developing graph-theoretic failure models for complex infrastructure networks, these often characterise the failure dynamics of a network in topological terms (Bompard et al., 2011; Boccaletti et al., 2006). While informative, this approach offers little insight into the explicit vulnerability of the network as a function of its spatial structure, or an obvious means by which to understand the temporal dynamics of the infrastructure network failure. In this paper, we present an approach that addresses the lack of real-time spatio-temporal analysis of infrastructure network performance by developing an open source geovisualisation tool coupled with infrastructure network failure models to simulate, visualise and analyse how spatial infrastructure networks behave over time.

## 2 Coupled network simulation and geovisualisation tool

### 2.1 Network model parameterisation

The tool has been developed in python using NetworkX (NetworkX, 2014) and can be run using either a static shapefile, or a spatial network stored in a separately developed interdependent spatial network database schema (Barr *et al.*, 2013) as input. The input data generates a network instance attached to which are the flow/movement attribute values of interest, such as time, distance or cost, while a separate set of node and edge classes are used to record/store dynamic attributes that represent the evolving response of

the system(Figure 1). The flows and movements assigned consist of a start time, start node and end node and thus form origin-destination pairs. Initial routes between origin-destination pairs are generated by calculating the weighted shortest or least cost path. The assignment of flows and their corresponding start time allows, at the simulation stage, the dynamics of how flows are affected spatially over time to be investigated.



**Figure 1:** Flow diagram illustrating the visualization tool developed.

## 2.2 Network failure modelling

Spatial infrastructure networks are exposed to a wide range of events which can affect component performance, from natural hazards through to breakdowns and stress due to demand (Demšar *et al.*, 2008). The ability to model such failures has been integrated within our tool in the form of three failure models; namely, (i) targeted failures such as failures ranked by node degree (number of incident edges), (ii) failures due to stress on the network for instance failures on edges exceeding a certain flow limit and (iii) failures due to natural hazards such as floods (spatial footprints represented in shapefiles for example). These failure methods can be applied to both nodes and edges, with the perturbation of both being possible within a single simulation run. During a failure simulation run all active flows or movements are checked and if their shortest/least cost path intersects the failed node(s)/edge(s) an alternative shortest/least cost path is re-calculated if possible and the affected flows assigned to the new route. At the same time any changes to the network structure are also recorded along with corresponding failure metrics (Figure 1).

## 2.3 Network visualisation



Visualisation of the network and its evolution of flows as a result of the failure model applied is undertaken using PyGame, an open source software package designed for developing games in OpenGL. Initially, static imagery is added to the PyGame canvas to provide geographical context and initial configuration of the network rendered from the network class. Once the simulation commences, the network are re-rendered according to the user specified simulation time step (e.g., every 30 seconds). Within the visualisation engine the dynamics of the changing network flows can be represented by either adjusting the size/thickness or colouring of edges/nodes to depict the number of flows across each network primitive at a given time period i.e. the last 10 minutes. The actual flows themselves can also be visualised, with a point plotted at each epoch.

### **3 Tyne and Wear traffic analysis**

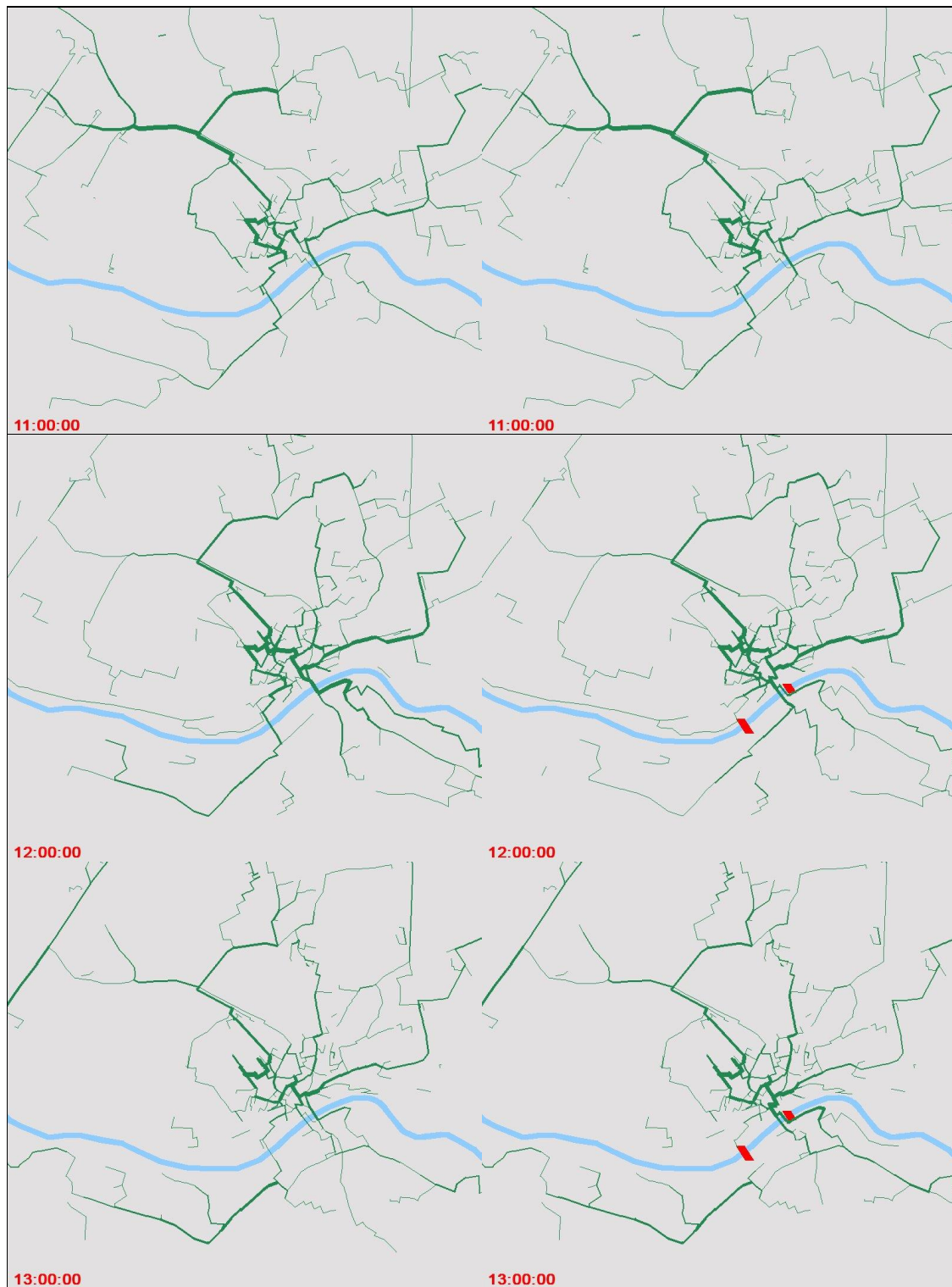
The developed tool was used to analyse the Tyne and Wear road network during the day of the Tyne and Wear derby (Newcastle United vs Sunderland). Flows were extracted from recorded geotagged tweets on this day; when a single Twitter user tweeted more than once throughout the day from different locations these were used to form a route that was mapped to the road network to create a network flow. From the twitter data 1337 flows were created and applied to the road network. An initial simulation was run to investigate hourly flows across the network (Figure ), that showed flows increased in the city centre before dying down as the games kicks off at 12:45 and then increased again around 14:30 as the fans begin to leave the football ground. Visible pinch-points occurred in traffic routes over the River Tyne, so a failure analysis at these locations was undertaken to analyse the effect of closing these roads. The failures were introduced at 11:00. Figure 3 shows the evolution of the routes post-failure showing previously low flow stretches of road experiencing dramatic flow increases in order to accommodate the lack of accessibility due to the failure of the principles bridges across the river Tyne.

### **4. Conclusion**

Visualising the spatial behaviour geospatial infrastructure networks has the potential to offer new insights into the dynamics of the networks which we rely upon for our quality of life and economic prosperity. The tool we have developed allows us to gain a better understanding of network behaviour when exposed to perturbations, enabling changes in flows across system to be assessed spatially over time, and thus allow spatially those parts of the system that may require adaption to be recognised. Future work will extend the tool to incorporate directly sensor flow data and to integrate this tool into a real time decision theatre environment.



**Figure 2:** The loading of flows over the road network from 10:00 to 16:00.



**Figure 3:** Comparing the traffic loads if two main road arteries were to be closed during the same period.

## 5. Biographies

Neil Harris is a researcher in Geomatics at Newcastle University. Whose research centres around the collection, management, analysis and visualisation of geospatial sensor and social media data.

Mr Craig Robson is currently studying for a Ph.D. in spatial infrastructure network modelling at Newcastle University.

Dr Stuart Barr is a Senior Lecturer in Geographic Information Science at Newcastle University.

Mr Philip James is a Senior Lecturer in Geographic Information Science at Newcastle University.

## References

- Adrienko, N. and Adrienko, G. (2011) 'Spatial Generalization and Aggregation of Massive Movement Data', *Visualization and Computer Graphics, IEEE Transactions on*, 17, (2), pp. 205-219.
- Barr, S. L., Alderson, D., Robson, C., Otto, A., Hall, J., Thacker, S. and Pant, R. (2013) 'A National Scale Infrastructure Database and Modelling Environment for the UK', *International Symposium for Next Generation Infrastructure*. Wollongong, New South Wales, Australia, pp.
- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. U. (2006) 'Complex networks: Structure and dynamics', *Physics Reports*, 424, pp. 175-308.
- Bompard, E., Wu, D. and Xue, F. (2011) 'Structural vulnerability of power systems: A topological approach', *Electric Power Systems Research*, 81, pp. 1334-1340.
- Demšar, U., Špatenková, O. and Virrantaus, K. (2008) 'Identifying Critical Locations in a Spatial Network with Graph Theory', *Transactions in GIS*, 12, pp. 61-82.
- Leu, G., Abbass, H. and Curtis, N. (2010) 'Resilience of ground transportation networks: a case study on Melbourne', *33rd Australian Transport Research Forum Conference*. pp. 14.
- NetworkX (2014) *NetworkX: Overview*. Available at: <https://networkx.github.io/>. (Accessed: 24/10).
- Tominski, C., Schumann, H., Andrienko, G. and Andrienko, N. (2012) 'Stacking-Based Visualization of Trajectory Attribute Data', *Visualization and Computer Graphics, IEEE Transactions on*, 18, (12), pp. 2565-2574.

# Football fan locality- An analysis of football fans tweet locations

Harris N<sup>1</sup>, James P<sup>1</sup>

<sup>1</sup> Newcastle University, School of Civil Engineering and Geoscience, Newcastle-upon-Tyne, NE1 7RU

November 11, 2014

## Summary

This paper looks at the validity of using social media as a dataset for spatial analysis and demonstrates the use of geo-located tweets to investigate the locality of football club fan-bases

**KEYWORDS:** Social media, Locality analysis, Football

## 1. Introduction

Twitter is a free social networking and micro-blogging service that enables its millions of users to send and receive messages of up to 140 characters (tweets). Twitter to date has over 284 million users and processes over 500 million tweets a day. Many recent events have been documented live via Twitter users on the ground. For example the hashtag #OccupyCentral became a live feed on information of the recent protest in Hong Kong with 700+ tweets a minute being sent at the peak of the protests (Boehler, 2014).

Approximately 3% of tweets also include location information in the metadata of the tweet. This amounts to 15 million tweets a day that contain location (Dredze *et al.*, 2013). Twitter has been used to map the effects of an earthquake (Yin *et al.*, 2012), the spread of disease (Lan *et al.*, 2012) and many other applications.

Tweets using the official club hashtags from football clubs provide a potentially rich data source to test the assumptions often made about the location of supporters. This is a hotly debated topic on forums and in pubs and assumptions about fan bases such as Manchester United supporters being predominantly based in the South East and their rivals Manchester City being locally based are widely quoted in the press (MEN, 2007). It is also a common assumption that the locality of football fans is inversely proportional to success, with teams that win trophies attracting fans from across the UK. (Dudley, 2014).

### 1.1 Football Tweets

Data is collected by listening on the official hashtag of each club in the football league e.g. Manchester United's official hashtag is #mufc. The location, the team in question and the time, are captured and stored in a PostGIS relational database. To date we have recorded several million football tweets spanning the last 3 seasons.

### 1.2 Locality

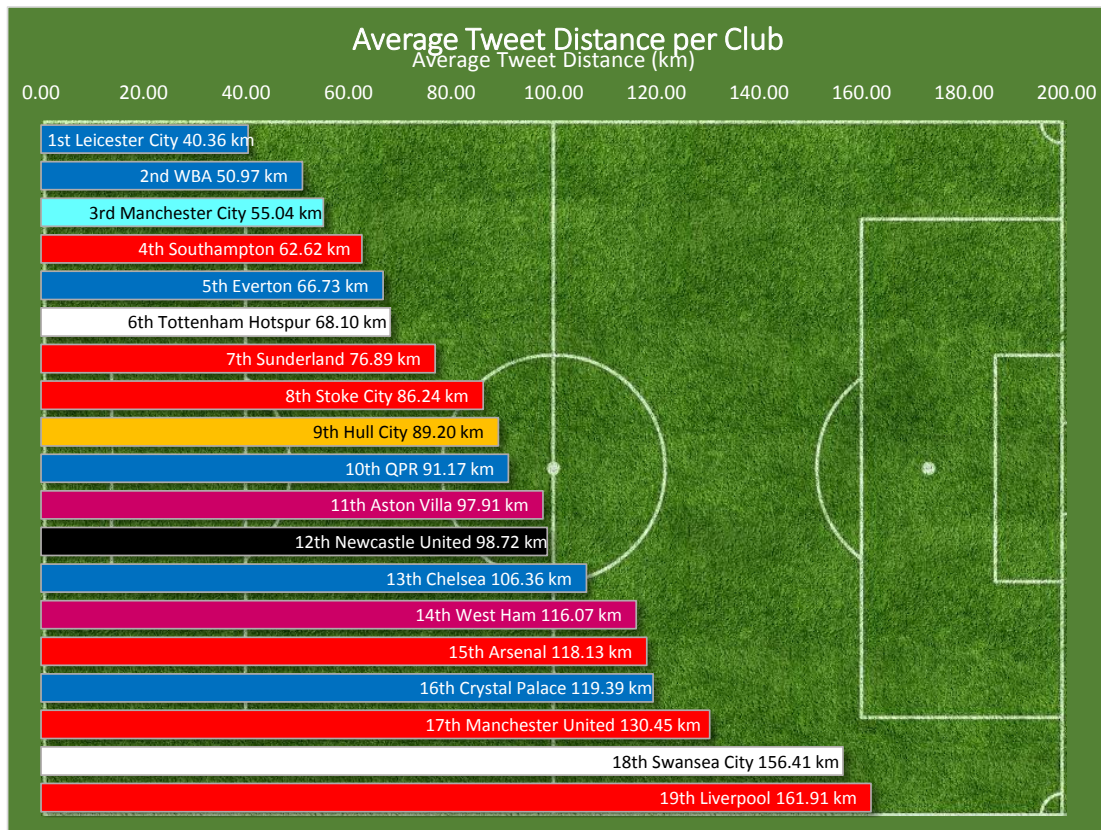
In order to understand the locality of football fans a definition of “local” is needed. In the food industry the term “local” is often used and definitions exist for the term local. CAMRA (2014) defines a local ale as one produced within 30 miles of the pub. *Farmer's Markets in Clark County* (2014) only count produce within 3 hours travel time as local. Waitrose (2014) considers local recipes as recipes using ingredients from the same county.

## 2. Analysis

In order to analyse the tweets a subset of the data was created using only tweets made during the October, 2014 international break (6/10/14 00:00 -16/10/14 00:00). This would discount tweets from fans travelling to and from away football matches and reduce tweets about contentious incidents in a game like a dubious red card. A total of 3000 tweets were collected for the Premier League and 1500 for the Championship.

## 2.1 Tweet distance as a Euclidean measure

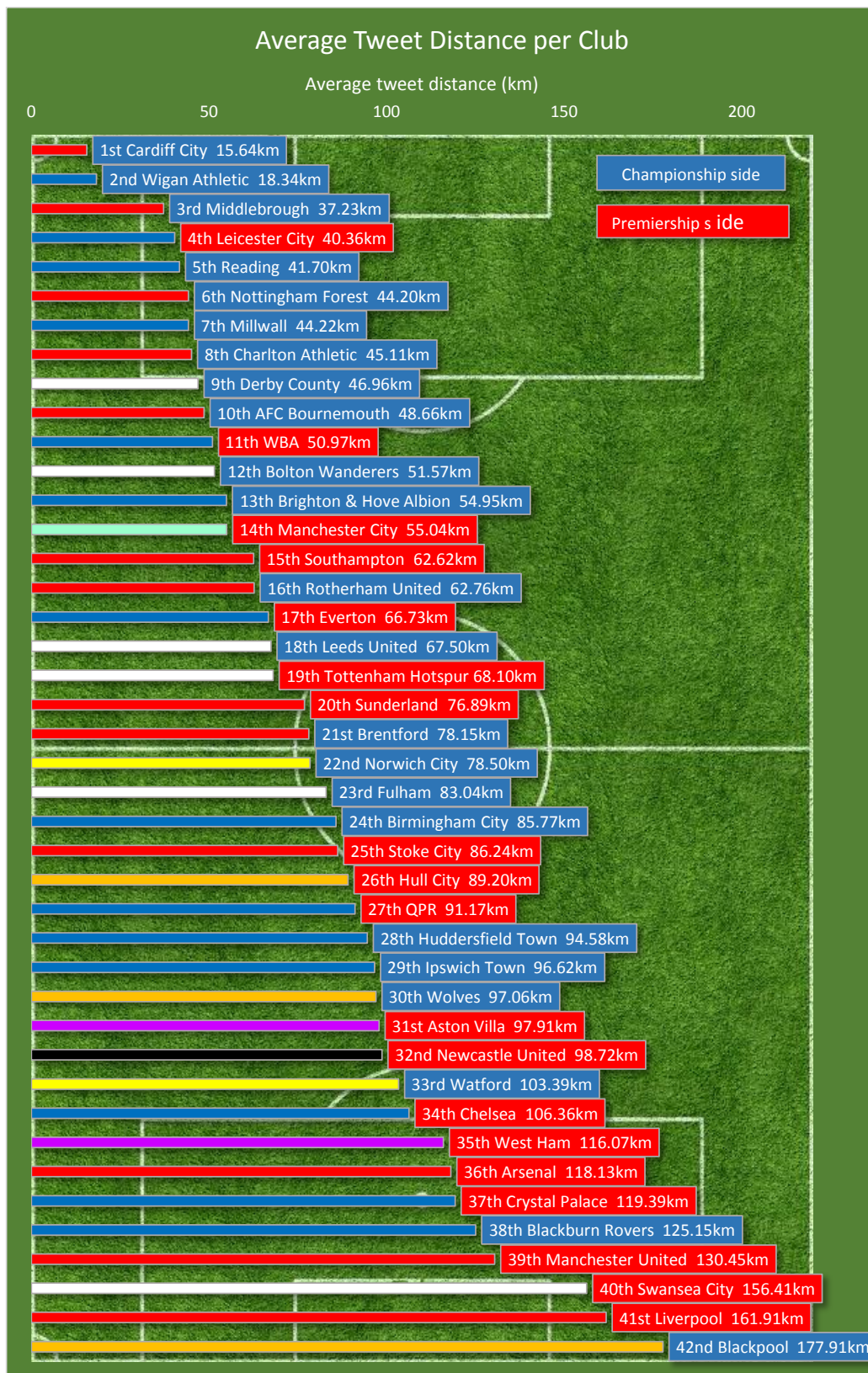
Initially the distance between tweet and club was analysed, with the home ground of each club acting as their reference point. Figure 1 shows the results for Premiership clubs. The results of this analysis are inconclusive although they suggest that there is some truth in the assumptions with the bottom of the table filled with the “glory clubs” of Manchester United, Arsenal and Liverpool. More unexpectedly, it also includes Crystal Palace and Swansea City. Swansea’s low score is probably due to their generic official hashtag of “#swans”. At the top of the table are some of the teams that have been recently promoted. Manchester City sits third perhaps suggesting that there is some truth in them being locally supported.



**Figure 1: Average Tweet distance per club**

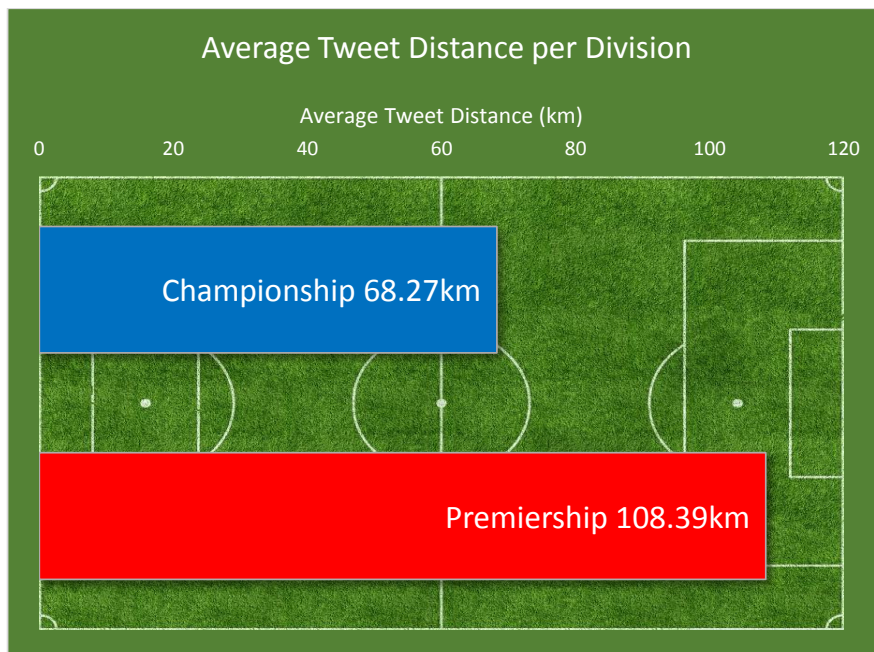
Another assumption often quoted by football pundits (*Football club and fan relationships better in lower leagues*, 2014) is that Premiership teams attract supporters outside the immediate geographic area whereas teams lower down the division are more likely to be locally supported. To test this hypothesis the Championship tweets from the same weekend were analysed in the same fashion and compared to the Premiership clubs. Figure 2 goes some way to supporting this assumption with only one Premier League side making it into the top 10 and only 7 in the top half of the table. The worst performing club is surprisingly Blackpool FC which may be due to the fact that at the time of data collection Blackpool were heavily in the news as a result of public disagreements between the owner, manager and fans (SkySports, 2014).





**Figure 2** Average tweet distance per club

To further test this hypothesis the average distance from tweet to club per division was computed and the results shown in Figure 3. Here the Championship clearly out scores the Premier League rivals with a difference in distance of 40km in average between the two divisions.

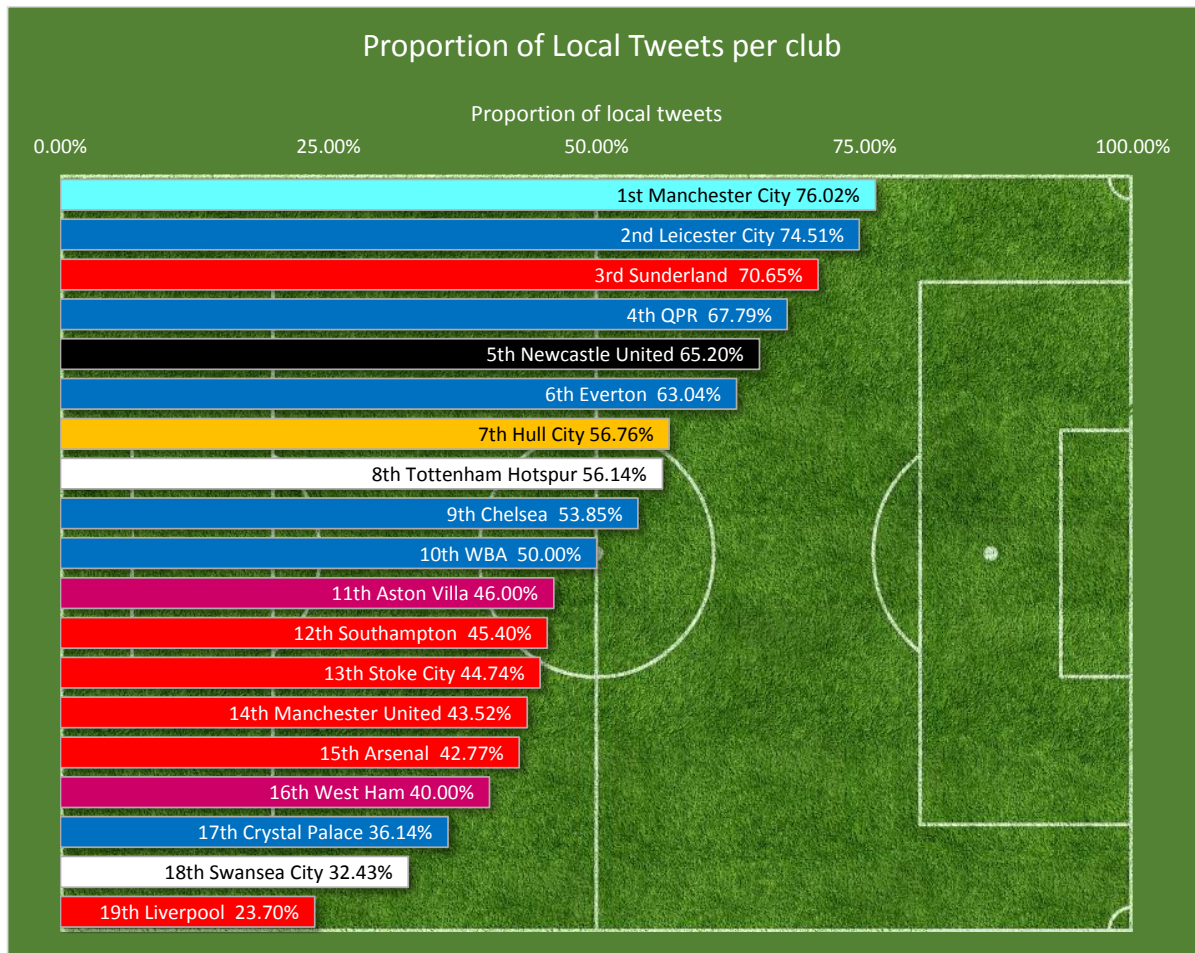


**Figure 3** Average tweet distance per division

## 2.2 Proportion of local tweets

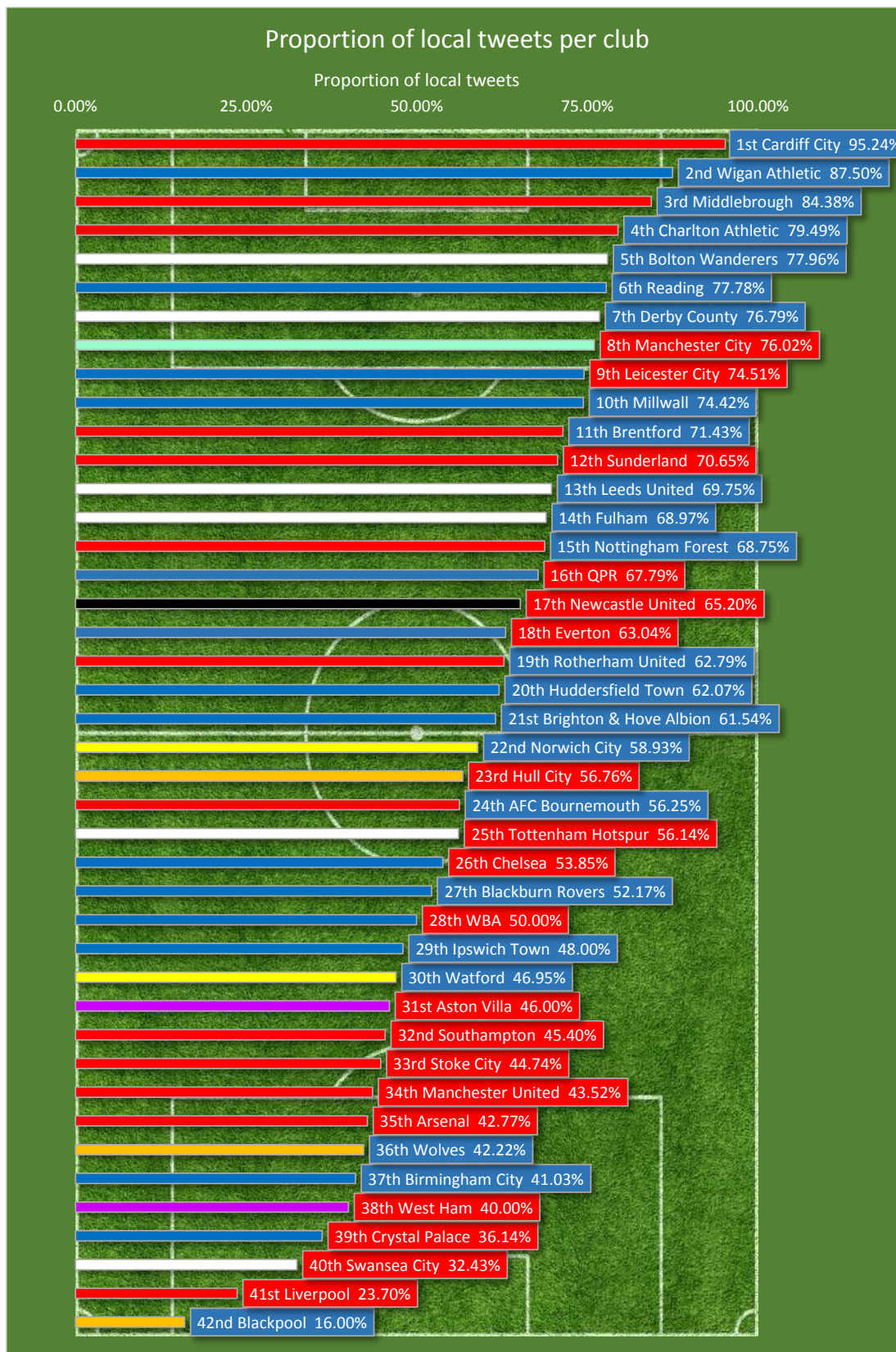
This basic analysis applies a 30 mile buffer around the club as a measure of locality, however if a club is more isolated, like Norwich City, it is possible to be more than 30 miles from the club but it still be the nearest football league club. Therefore a tweet was considered local if it was within 30 miles of the ground **or was the nearest** football league club. Figure 4 shows the results of this test. Again the results are inconclusive with Liverpool and Arsenal and Manchester United in the lower half but accompanied by some other teams like West Ham and Crystal Palace. At the top of the table we have Manchester City and a smattering of some of the smaller clubs.





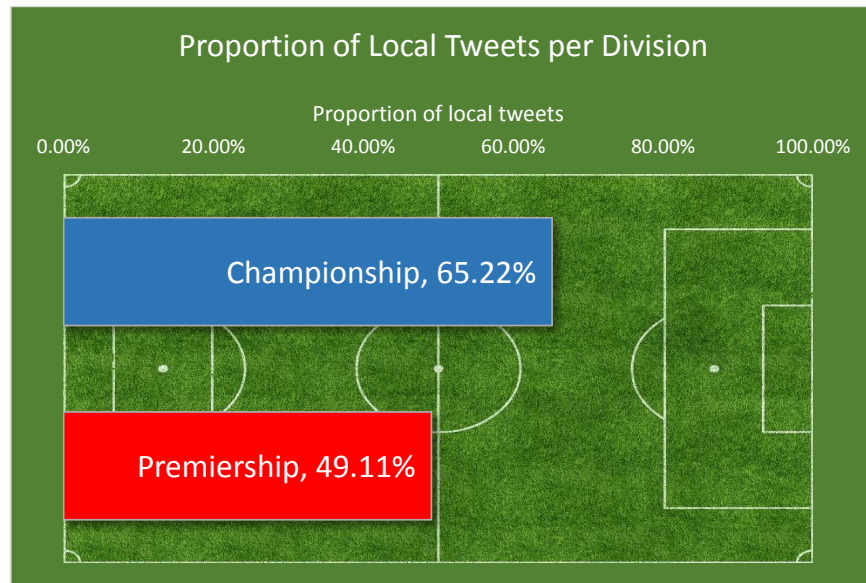
**Figure 4** Proportion of local tweets

Again this test was repeated including Championship sides with the results shown in Figure 5. This supports the hypothesis that teams in lower divisions attract a local fan base with only two Premier League clubs now in the top 10 and only 6 in the top half of the table.



**Figure 5** Proportion of local tweets

Finally the proportion of local tweets per division was looked at and the results show in Figure 6. Again the Championship scored significantly higher than the Premier League.



**Figure 6** Proportion of local tweets per division

### 3. Conclusions

This analysis of the tweets for this one weekend broadly supports the hypothesis that Championship teams have a more local fan base than their Premier League rivals.

When the results from both the distance and proportion tests are combined (Table 1) the results suggest that there is some truth to the assumption that the more successful teams have a wider fan base. The top end of the table is dominated by lesser teams with Manchester City and Everton being the exceptions, both of which are a city's second team which may result in their high placing. At the foot of the table, whilst it is Manchester United which has the reputation for the most widespread fan base, it is actually Liverpool who performed worst in both tests. The two notable anomalies in the bottom five are Swansea City and Crystal Palace. It is unclear from the tweet data itself why Crystal Palace does so poorly.

**Table 1:** Summary of results per club

Pos	Club	Rank on Average Distance	Rank on Proportion of local tweets	Lowest rank score	Average Rank
1	Leicester City	1	2	1	1.5
2	Manchester City	3	1	1	2
3	Sunderland	7	3	3	5
4	Everton	5	6	5	5.5
5	WBA	2	10	2	6
6	QPR	10	4	4	7
7	Tottenham Hotspur	6	8	6	7
8	Southampton	4	12	4	8
9	Hull City	9	7	7	8
10	Newcastle United	12	5	5	8.5
11	Stoke City	8	13	8	10.5
12	Chelsea	13	9	9	11
13	Aston Villa	11	11	11	11
14	West Ham	14	16	14	15

15	Arsenal	15	15	15	15
16	Manchester United	17	14	14	15.5
17	Crystal Palace	16	17	16	16.5
18	Swansea City	18	18	18	18
19	Liverpool	19	19	19	19

Given that the Twitter data is a simple point data set any number of further analyses could be applied to it depending on the nature of the selected concept of “local”. Looking at the proportion of supporters within a region e.g. the South East or considering network travel time rather than distance may also yield some interesting results.

Lovelace *et al.* (2014) discuss the limitations of using social media as a data source in spatial analysis and whilst it is not without its uncertainties it is important not to dismiss it outright as a concept. The sheer volume of tweets provides a rich data source for the research community. Further information on football tweets can be found at: <http://ceg-sense.ncl.ac.uk/footballtweet>.

#### **4. Biographies**

Neil Harris is a researcher in Geomatics at Newcastle University. Whose research centres around the collection, management, analysis and visualisation of geospatial sensor and social media data.

Phil James is a Senior Lecturer in GIS. His research centres on the use of Geospatial data to solve and support Engineering problems. Phil is the academic lead on Newcastle's Urban Observatory that integrates real time urban sensors across multiple scales with geospatial data and engineering models using the Cloud.



## 5. References

Boehler, P. (2014) *How Hong Kong's #OccupyCentral became a global topic on Twitter*. Available at: <http://www.scmp.com/news/hong-kong/article/1604409/infographic-how-hong-kongs-occupycentral-became-global-topic-twitter> (Accessed: 3/11/2014).

CAMRA (2014) *CAMRA LocAle*. Available at: <http://www.camra.org.uk/locale> (Accessed: 6/11/2014).

Dredze, M., Paul, M.J., Bergsma, S. and Tran, H. (2013) 'Carmen: A Twitter Geolocation System with Applications to Public Health'.

Dudley, B. (2014) 'Sing when you're losing', *Supporters Not Customers*. Available at: <http://supportersnotcustomers.com/2014/04/22/sing-when-youre-losing/>.

*Farmer's Markets in Clark County* (2014). Available at: <http://ext100.wsu.edu/clark/agriculture/small-farms/farmersmarkets/> (Accessed: 6/11/2014).

*Football club and fan relationships better in lower leagues* (2014). Available at: <http://www.staffs.ac.uk/news/football-club-and-fan-relationships-better-in-lower-leagues-staffs-uni-researcher-suggests-tcm4233521.jsp> (Accessed: 6/11/2014).

Lan, R., Lieberman, M.D. and Samet, H. (2012) 'The picture of health: map-based, collaborative spatio-temporal disease tracking'. pp. 27-35.

Lovelace, R., Malleon, N., Harland, K. and Birkin, M. (2014) 'Can Social media data be useful in spatial modelling?', *GISRUK* Glasgow.

MEN (2007) *Ben salutes Blue Manchester*. Available at: <http://www.manchestereveningnews.co.uk/sport/football/football-news/ben-salutes-blue-manchester-1117827> (Accessed: 6/11/2014).

SkySports (2014) *Sky Bet Championship: Blackpool chairman Karl Oyston insists protests won't work* (Accessed: 6/11/2014).

Waitrose (2014) *The Waitrose Small Producers' Charter*. Available at: [http://www.waitrose.com/content/waitrose/en/home/inspiration/about\\_waitrose/the\\_waitrose\\_way/small\\_producers\\_charter.html.html](http://www.waitrose.com/content/waitrose/en/home/inspiration/about_waitrose/the_waitrose_way/small_producers_charter.html.html) (Accessed: 6/11/2014).

Yin, J., Lampert, A., Cameron, M., Robinson, B. and Power, R. (2012) 'Using Social Media to Enhance Emergency Situation Awareness', *IEEE Intelligent Systems*.

# Do Geospatial & Heritage standards work and do they work together?

Glen Hart,

University of Nottingham

November 2014

## Summary

This research compared the ability of three geospatial and heritage standards to meet the needs of a heritage organisation concluding that the standards in either isolation or combination did not fully meet the requirements and that the nature of the standards made it difficult for them to work together. The work recommends the development of micro-standards to overcome these difficulties.

**KEYWORDS:** Standards, Heritage, GML, CIDOC-CRM, MIDAS

## 1 Introduction

The efficient exchange and reuse of data is vital to the future growth of the digital economy but is still a major challenge. Standards are important elements in addressing this issue by the geospatial and heritage communities among others. It is therefore worth asking how helpful, compatible and mutually supportive standards written by different communities are. Qualitative research was conducted through studying the Royal Commission on the Ancient and Historic Monuments of Scotland (RCAHMS), and with a specific focus on data standards, rather than standards involved with delivering services and APIs. The work was funded by the Horizon Digital Economy Institute at the University of Nottingham.

## 2 RCAHMS and Standards

RCAHMS identifies, surveys and analyses the historic and built environment of Scotland. It obtains information on the historic environment through its own surveys and investigative work, and from archaeological consultancies. Data is made available to professional bodies and the general public in a number of formats including ESRI Shapefile, CSV and Excel but not standards such as GML. The design of the RCAHMS Database has been heavily influenced by MIDAS, a heritage standard created in the 1990's.

RCAHMS' interests centre around three main concepts: Monument or Sites, Event and Collection. Of these the first was the main focus of the research. Monuments are the primary interest and represent things such as historic buildings, statues, gardens, standing stones, earth works, battlefields, and find spots. A monument's location is important as are dates, monument (place) names and classification. Events record monument investigations, such as a site survey using Lidar or architectural assessment. Collections are largely documents and photography concerning monuments.

RCAHMS' systems are not perfect; RCAHMS recognises the deficiencies and is working to improve them. For example monument names are inefficiently held and represented making it difficult to query them, and where a monument has multiple names there is no way to differentiate or classify them as

preferred, official, historic etc. Dates are not consistently implemented, there is no mandated way of representing location in terms of geometry, and there is also only a fairly limited way to reference one monument as being part of another. A greater adoption of standards could help to resolve the current deficiencies and promote greater data interoperability. Three standards were examined as being appropriate to RCAHMS: Geography Mark-Up Language (GML), CIDOC-Common Reference Model (CRM) and MIDAS Heritage.

GML is an important OGC standard for geospatial data transfer and modelling. GML is implemented as XML and has schemas that support topics such as geometry, topology and temporal models. The geometry and topology models are hugely influential having been used to implement spatial extensions to (o)rdbms' and triplestores including their query languages (SQL and SPARQL). GML is rather complex and non-governmental users often use popular less formal standards such as GeoJSON.

Produced by the CIDOC Documentation Standards Working Group within the International Council of Museums, CRM is a domain standard that promotes the exchange of heritage information. CRM's scope "can be summarised in simple terms as the curated knowledge of museums." (Le Boeuf et al, 2013) and "is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to." Like GML, CRM provides a language or vocabulary to enable a user to model their system, rather than be very specific about absolute structure. CRM is a complex standard that has most recently been expressed as an OWL ontology. The high-level abstract structure and depth of hierarchy make for a standard with "sometimes bewildering complexity" (Carlisle et al, 2014). It does nonetheless provide a good description of the informational needs of the heritage community.

MIDAS is a heritage standard developed by English Heritage in conjunction with many other UK heritage organisations. Published in 1998 it was revised in 2012 as MIDAS Heritage. MIDAS has been influenced by CRM. MIDAS does not tie itself to any particular implementation technology or philosophy. MIDAS is categorised into a number of Information Themes, which in turn are broken down into Information Groups. The main themes identify the primary interests in monuments and artefacts, what has happened to them and how they are documented. The supporting themes then provide the where, when and by whom. This is very similar to CRM, but much clearer and easier to digest.

### **3 Discussion**

MIDAS provides the best description of the Heritage community's data needs. CRM provides greater detail in terms of structure. CRM and MIDAS lack specificity in some areas such as positional information, place names and time. GML provides well-defined geometry types and a formal way to specify coordinate reference systems. In certain areas all the standards lack specificity, for example no detail is given as to how digital identifiers, place names or classification should be represented. In all cases what is missing is not so much a mandate that says that this classification system should be used, or these approved place names are the ones, but structurally how they should be represent and the semantics associated with those structures.

Ideally using these standards together as well as supplementing them with the missing detail would provide a good solution. However the heritage standards were not designed to work with GML and vice versa. MIDAS and CRM are essentially compatible and it is possible to map MIDAS onto CRM. GML and CRM are competing implementations, both designed to provide "the data model". Most sensibly would be for CRM to provide the domain framework and to utilise those aspects of GML where CRM is deficient such as geometry. This is more or less what Arches, an open source solution



for managing heritage data, does. It implements the CRM model but also supports GML as an import and export format (along with a number of others) and represents geometry using the OGC standard (Well-known text). Arches hides unwanted complexity rather than address the problem at source and a considerable amount of the GML standard is effectively redundant. The simple aspects of GML will be reusable but other aspects are not as transferable. The way GML defines a Period (such as the Ice Age) is too simple for heritage organisations where a period is not simply a tag associated with a time range, but a more complex interaction between culture, geography and time. The temporal model is more complex than currently required by RCAHMS and where the complexity is required by others does not always meet expectations resulting in the model being ignored (Parcero-Oubiña, 2012) (Plumejeaud, 2010) or modified (Siyuan et al, 2007) (Quak & Vries, 2005).

These observations point to improvements to the way standards are developed. There is a need to identify and standardise the way that some basic things such as place names and classifications are exchanged. Secondly, we should avoid wherever possible constructing large complex standards. This will be harder to do for domain standards where the domains themselves are necessarily complex; but here MIDAS addresses this complexity through sensible modularisation, conversely CRM supports a complex and unnecessarily deep hierarchy. GML is attempting to provide supporting ‘horizontal’ standards to ‘vertical’ domains such as heritage, but does so in an all-encompassing package. It might be better if a trend in ontology development were followed where there is a move away from large ‘upper’ ontologies such as DULCE and SUMO in favour of lightweight micro-ontologies (Hart & Dolbear, 2013) which could be mirror by moving away from large horizontal standards in favour of smaller micro-standards. Hence the well-defined and popular aspects of GML such as geometry could be disaggregated into compact micro-standards that can be easily adopted by domain standards. Lastly, it would be beneficial if such standards are not tied to particular technologies such as XML but instead have an abstract definition accompanied by several technology profiles.

#### 4 References

- Carlisle, P. K., I. Avramides, A. Dalgity, and D. Myers. 2014, “The Arches Heritage Inventory and Management System: A Standards-Based Approach to the Management of Cultural Heritage Information.” Paper presented at the CIDOC Conference: Access and Understanding – Networking in the Digital Era, Dresden, Germany.
- Hart, G., Dolbear. C. 2013, *Linked Data: A Geographic Perspective*, CRC Press, ISBN-10: 1439869952, ISBN-13: 978-1439869956
- Le Boeuf P., Doerr M., Ore C. E., Stead S. (Eds), 2013, *Definition of the CIDOC Conceptual Reference Model, Version 5.0.1(draft)*, [http://cidoc-crm.org/official\\_release\\_cidoc.html](http://cidoc-crm.org/official_release_cidoc.html)
- Lee E. (Ed.), *MIDAS Heritage*, 2012, *The UK Historic Environment Data Standard*, V1.1, <http://www.english-heritage.org.uk/publications/midas-heritage/>
- Parcero-Oubiña C, Fábrega-Álvarez P, Vicent-García M, Uriarte-González A, Fraguas-Bravo A, del-Bosque-González I, Fernández-Freire C, Pérez-Asensio E, 2012, *Conceptual basis for a cultural heritage data model for INSPIRE*, *Proceedings of the AGILE'2012 International Conference on Geographic Information Science*, Avignon, ISBN: 978-90-816960-0-5
- Plumejeaud C, Mathian H, Gensel J, Grasland C, 2010, *Spatio-temporal analysis of territorial changes from a multi-scale perspective*, *International Journal of Geographical Information Science*, 25:10, 1597-1612
- Quak, CW & Vries, ME de. *Topological and temporal modelling in GML*. In J Drummond (Ed.), 2005, *Proceedings of the topology and spatial databases workshop* (pp. 1-8). Glasgow: University of Glasgow

Siyuan F, Griffiths A, & Paton N, 2007, GML for Representing Data from Spatio-Historical Databases: A Case Study, *Transactions in GIS*, 11(2): 233–253

# Objectively scrutinising the impact of the obesogenic environment on obesity in Yorkshire, England: a multi-level cross-sectional study

Hobbs, M<sup>1</sup>., Green, M<sup>2</sup>., McKenna, J<sup>1</sup>., Jordan, H<sup>2</sup> and Griffiths, C<sup>1</sup>

<sup>1</sup>The Centre for Active Lifestyles, Leeds Beckett University

<sup>2</sup>School of Health and Related Research, Sheffield University

January 14<sup>th</sup> 2015

**KEYWORDS:** Obesogenic environment, physical activity facilities, green space, food outlets, obesity

## Summary

Policy makers are beginning to engage with the idea that the built environment may be a contributing factor to obesity. Despite an increasing policy focus, identifying associations between exposure to an obesogenic environment and increased adiposity has proved challenging. The evidence base remains equivocal. By influencing the ‘default’ option in the obesogenic environment such as the proximity of fast food outlets or green space, there may be some potential to affect dietary intake, physical activity and obesity. This study aims to examine the spatial distribution of obesogenic environment across Yorkshire and the impact on adult adiposity.

## 1. Introduction

Obesity rates in UK adults continue to be some of the highest in Europe with 24% and 25% of males and females reported to be obese respectively. Of concern is that governments and local authorities have repeatedly attempted to address the issue of obesity. Despite some attenuation, their approaches on the whole have been ineffective. Policy makers are subsequently, beginning to engage with the idea that the built environment may be a contributing factor to the obesity epidemic. Indeed, public health professionals in the UK are now encouraged to address the prevalence of fast food outlets in their area to support healthier lifestyles (Cavill and Rutter 2013).

Neighborhood built environments; both food and physical activity, have been labelled obesogenic. They are said to facilitate an over-consumption of energy-dense, nutrient poor foods at the expense of minimal energy expenditure; thus increasing obesity. Despite an increasing policy focus, identifying associations between exposure to an obesogenic environment and increased body weight has proved challenging and the evidence base remains equivocal. For instance, a recent systematic review (Fleischhacker et al. 2011) found that of those studies examining these exposures in relation to increasing body weight, fewer than half reported positive associations. Further, even fewer of the reviewed studies that were included were conducted in England. The evidence base is therefore not strongly placed at present to support interventions into politically difficult modifications of alleged obesogenic environments.

Despite a lack of scientific support, modification of the obesogenic environment currently represents a key focus of local authority health policy. It is an attractive population level intervention to limit risk factors conducive to obesity. Briefly, genetic evolution has been wholly unable to match the rapidity of the environmental and societal transitions made within the 21st century. Therefore, individuals currently pay an overwhelming amount of attention to override their natural habits due to the environmental cues of contemporary life. For example, the high energy densities of many fast foods challenge human appetite control systems with conditions for which they were never designed (Prentice and Jebb 2003). By influencing the ‘default’ option in the obesogenic environment such as the proximity of fast food outlets or green space, there may be some potential to affect dietary intake, physical activity and ultimately obesity.

The ‘default option’ for an individual may be influenced somewhat by the obesogenic environment however; some population groups for example low socioeconomic status (SES) may be more vulnerable than others to the effects of the obesogenic environment. Individual- and area-level SES measures are independently related to obesity. It is increasingly important to consider both individual- and area-level measures of SES using multi-level modelling. Lower area level SES with a greater density of fast food outlets may amplify individual risk factors for obesity such as low income, education or absence of transport; a phenomena known as deprivation amplification. Despite the apparent consequences of individuals operating within obesogenic environments the research to date has insufficiently and inconsistently addressed the issue.

It is interesting that despite lobbying from policy, contemporary evidence on the obesogenic environment is essentially in its infancy. It may be unsurprising that those associations linking exposure to food and physical activity environments to weight status are at present equivocal. Besides, comparisons between studies are made difficult by a US-centric evidence base and should be interpreted with care. England in any case has a very different environmental temperament to the US yet, this important distinction has rarely explicitly been raised. Lastly, few studies have data on both the food and physical activity environment. Therefore, there is little understanding of which aspect of the neighborhood is comparatively stronger in predicting obesity and thus offering greatest policy impact. In addition to these considerations, the lack of associations identified in studies could be explained through other limitations.

Due to the infancy of the evidence, much is cross-sectional limiting the ability to draw causal inference. Furthermore, an individual’s neighbourhood (which in turn defines exposure) is often arbitrary defined for instance, a 400m circular buffer around a participants home. This definition rarely has any theoretical underpinning as to the size or shape of the neighbourhood. Neighbourhoods used within obesogenic environment research are inconsistently defined and rarely represent the locations used to actually buy food. Encouraging a range of measures at present is the best estimate. Importantly, neighbourhoods and exposure to the obesogenic environment often extend beyond home neighbourhood to the work or commute neighbourhood. However, research predominantly focuses on just the residential environment. Further research is needed to inform evidence based policy to tackle the obesity epidemic.

This study therefore aims to examine:

- i) The spatial distribution of the food and physical activity (obesogenic) environment across Yorkshire by gender, deprivation and ethnicity.
- ii) The impact of exposure to the food and physical activity (obesogenic) environment on an individual’s adiposity.

## **2. Methods**

### Study Design & Data Sources

The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) recommendations were used to add methodological rigour. This large (n=25,000) multi-level cross-sectional study represents collaboration with Sheffield University’s Yorkshire Health Study (YHS). Following ethical approval participants were contacted through consenting general practitioner surgeries (response rate; 15.9%). Relevant exposure metrics of the physical activity and nutrition environment were then computed in ArcGIS (*ESRI, version 10.2*) using the UK Ordnance Survey Points of Interest (PoI) dataset.

### Exposure Variables

Participant’s home postcodes were geocoded and mapped using ArcGIS. Food outlet PoI’s were categorised into supermarkets, takeaways and other retail. PA facility PoI’s were included as a whole category. The density of PoI’s (both food and PA) was then calculated by placing circular buffers (100m, 400m, 800m, 1000m, 1600m and 2000m) on home postcode locations. Proximity was

represented by calculating straight line distance from home postcode to the nearest PoI (food and PA). Further, the quantity of green space was represented at the Middle Super Output Area (MSOA) and Lower Super Output Area (LSOA) the individual lives in. Briefly, a LSOA and MSOA are UK Census geographies designed for small-area statistical analysis. Finally, rural or urban classification for each MSOA and LSOA will be obtained from the Commission for Rural Communities Framework 2004.

#### Outcome Variables

Originally, this study represents weight status (dependent) through multiple outcomes. Body mass index (BMI) was calculated by dividing weight (kg) by height (cm) squared. Participants were then split into categories of underweight (<18.50), healthy weight (18.50 – 24.99), overweight (25.00 – 29.99) and obese (>30.00) respectively. Waist Circumference (cm) was then categorised into no, high and very high increased risk of health complications based on gender specific cut-points. The National Institute of Clinical Excellence (NICE 2014) suggests the assessment of health risks associated with increased adiposity should now be based on BMI and WC (Table 3) to determine if the adult is at any increased risk of associated health issues.

#### Covariate Variables

Age, gender, ethnicity and SES were collected from the participants and the highest level of education was used as a measure of individual level SES.

#### Bias & Missing Data

Data validation was an extremely thorough process. Data was checked by all parties involved in the collaboration. The missing data was then explored (by the first author) dependent upon the type, amount and distribution of missing data. Based on the large sample size, data was included only if postcode, gender, age, ethnicity and either BMI or WC are present, age was greater than 18 years of age and the postcode lay within the Yorkshire boundary. This resulted in 25 706 and 25 294 cases for BMI and WC respectively.

#### Statistical Methods

The dataset has a multi-level hierarchical structure which consists of adults nested within neighbourhoods. Based on BMI two binary variables of i) overweight and obese or not and ii) obese or not were created for logistical analysis. A further logistical analysis for WC will be carried out for i) increased risk or not and ii) substantially increased risk or not. Subsequently, multi-level modelling using two level hierarchical logistical models will be used determine variance at different levels using the dependent variables of BMI and WC. Predictors will be sequentially added to models and possible interactions between explanatory variables will be explored.

#### Strategic Alignment

Relating back to the strategic aims of GISRUK, this project represents interdisciplinary collaboration between The School of Public Health and Related Research (SchHARR) at Sheffield University and The Centre for Active Lifestyles at Leeds Beckett University. Further, the project is led by a PhD student supported by Senior Lecturers and Research Associates from both institutions. The goal is to respond to the urgent need to identify evidence based policy to tackle the obesity epidemic. Further, with the help of GIS the project aims to guide long-term town planning, policy change and redesign of existing urban environments to maximise physical activity, nutrition and minimise sedentary behaviour and obesity; all determinants of human health.

### **3. References**

1. Cavill, N., and Rutter, H. (2013) **Healthy people, healthy places briefing: obesity and the environment: regulating the growth of fast food outlets**. London, Public Health England.
2. Fleischhacker, S., Evenson, K., Rodriguez, D., and Ammerman, A. (2011) A systematic review of fast food access studies. **Obesity Reviews**, 12 (501), pp.460-471.
3. Prentice, A., and Jebb, S. (2003) Fast foods, energy density and obesity: a possible mechanistic link. **Obesity Reviews**, 4, pp.187-194.



# Evaluating the *Spraycan*: understanding participant interaction with a PPGIS

J. J. Huck<sup>1</sup>, D. Whyatt<sup>2</sup> and P. Coulton<sup>1</sup>

<sup>1</sup>Imagination Lancaster, LICA, Lancaster University

<sup>2</sup>Lancaster Environment Centre, Lancaster University

March 9<sup>th</sup> 2015

## Summary

Whilst widely accepted as an important facet of software design, the evaluation of PPGIS usability is often overlooked in research. This work comprises a novel approach to the evaluation of the *Spraycan* PPGIS, whereby rich insights into participant behaviour are drawn from data that are natively collected by the platform as opposed to through additional questionnaires, log files or similar. The approach will be validated against a 'traditional' questionnaire, before conclusions are drawn relating to the usability of the *Spraycan* as a platform for the collection of vague spatial data, in the hope of developing a greater understanding into the way in which people interact with geographic problems.

**KEYWORDS:** PPGIS, Fuzzy Geography, Place, HCI, Usability.

## 1. Introduction

Huck et al. (2013, 2014) introduced the *Spraycan* PPGIS platform for capturing imprecise notions of place from the public. Utilising an airbrush interface and the multi-point-and-attribute relational data structure, this PPGIS is designed to collect geographic data from participants without restricting those data to the primitive point, line and polygon structures upon which traditional systems rely. Rather, a participant may use the airbrush interface to define 'fuzzy' geographic regions, without defined boundaries, and with variations in intensity of the spray reflecting perceived spatial variations in membership or meaning. Whilst this platform has successfully been used for a range of applications, there has thus far been no investigation into *how* participants use and interact with the map and interface.

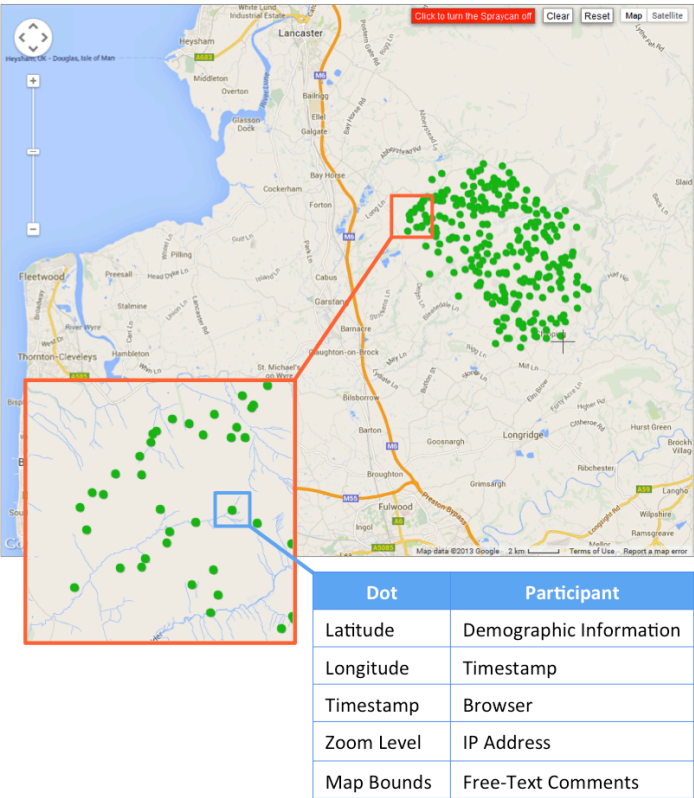
Public Participatory GIS (PPGIS) typically refers to a system for the collection of geospatial data from the public, and is a concept that may be traced back to Carver (1991). PPGIS aims to produce maps and spatial stories that help to characterise the local space (Bugs, 2010), a goal which can be limited by the difficulties associated with the collection of non-Cartesian, contradictory, or shifting forms of knowledge (Elwood, 2006; Montello et. al. 2003; Huck et. al. 2014). It is the collection of 'vague' or 'fuzzy' data such as these that the *Spraycan* aims to facilitate. Sieber (2006) and Rinner (2009) lament the lack of studies designed to measure the effectiveness of PPGIS', with emphasis traditionally placed upon the development of the systems themselves, rather than the study of whether the system is useful or how it is used (Zhao and Coleman, 2006). Some projects have addressed this issue using techniques drawn from the discipline of Human-Computer Interaction (HCI) (Haklay and Tobon, 2002; Haklay and Tobon, 2003; Zhao and Coleman, 2006; Bugs, 2010). HCI examines the interaction between humans and computers and the extent to which a computer system supports users to achieve specific goals (Zhao and Coleman, 2006).

Haklay and Tobon (2003) suggest that HCI studies are vital to the success of PPGIS because they aim to understand how people interact with computer applications, and therefore help researchers understand users' expectations as well as the ways in which they use, understand and value the

system (Zhao and Coleman, 2006). Attempts to apply HCI techniques to PPGIS thus far have relied upon techniques such as audio and screen recording, ‘thinking aloud’ protocols, the examination of system log files, interviews, and questionnaires (Haklay and Toban, 2002; Demsar, 2007). The *Spraycan* platform, however, provides the facility to undertake some investigation into participant interactions ‘natively’, using the data that are already collected as part of the analysis, thus reducing the requirement for expensive and invasive evaluation techniques as well as allowing for the retrospective analysis of data collected in previous surveys. This ‘native’ approach is, however, of limited use without validation, and so this work will also employ a traditional survey questionnaire for comparison, in order to assess the extent to which the findings from these ‘native’ techniques match those from more ‘traditional’ HCI approaches. As well as the validation of the ‘native’ approach, this work will also comprise an investigation into *how* participants interact with the *Spraycan*, in order to learn more about the suitability of this platform for the collection of vague spatial data from participants.

## 2. Methodology

The airbrush interface of the *Spraycan* allows participants to construct vaguely defined regions by continuously adding random dots within a specified distance of the mouse location. Each individual ‘dot’ of paint created by the user is stored in the *multi-point-and-attribute* data format along with a number of attributes relating to its own spatial and aspatial properties, including properties relating to the user that created it (illustrated in Figure 1). These attributes include: a latitude-longitude location; a millisecond timestamp; the zoom level of the map at the point of creation; the bounds of the map at the point of creation; and free text comments provided by the participant, amongst others.



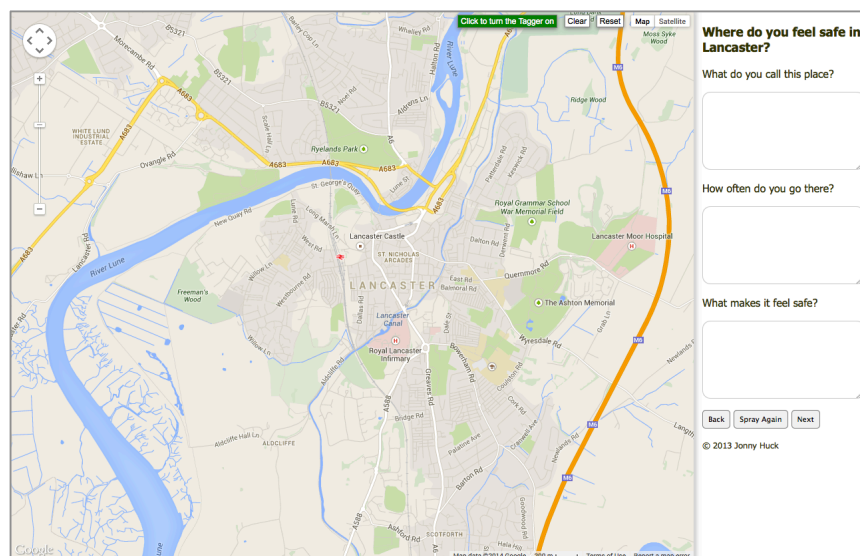
**Figure 1:** Illustration of the *multi-point-and-attribute* data format, with each ‘dot’ of paint related to a number of attributes. *Reproduced from Huck et. al. (2014).*



A group of 100 undergraduate geography students at Lancaster University have participated in this analysis, which comprised a short spatial survey using the *Spraycan* (via <http://map-me.org>), followed by a questionnaire (via <http://www.surveymonkey.com>) allowing them to reflect upon their experience. The questions from the Map-Me survey are given below, with one spatial question (in bold, answered by spraying onto the map) accompanied by three aspatial ‘contextual’ questions (bullet-points, answered with free-text):

- **Where do you feel safe in Lancaster?**
  - What do you call this area?
  - How often do you go there?
  - What makes it feel safe?
- **Where do you feel unsafe in Lancaster?**
  - What do you call this area?
  - How often do you go there?
  - What makes it feel unsafe?

A screenshot of one of the above questions is shown in Figure 2. Automated analysis of the collected spray patterns and their associated attributes will permit investigation into *how* the *Spraycan* is used, *how effective* it is in the collection of spatially-vague thoughts and feelings from participants, and whether or not any bias may be introduced into surveys by the interface or platform itself. Findings relating to these topics will then be compared to supporting questions asked via the ‘traditional’ questionnaire, in order to evaluate the quality of such analysis in comparison to more traditional techniques.



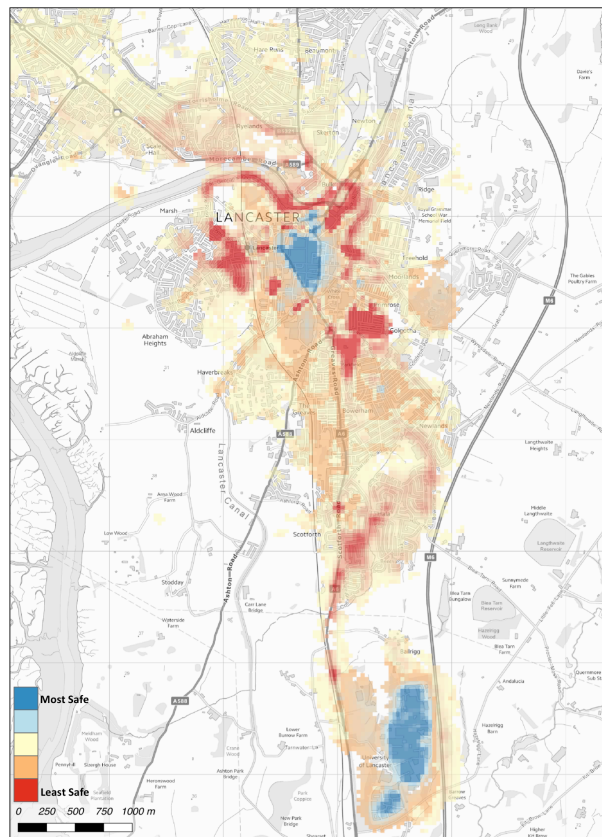
**Figure 2:** Screenshot of one of the questions from the *Map-Me* survey (via <http://map-me.org>). Participants use the *Spraycan* interface to answer the spatial question (in bold), and the free-text boxes to answer the aspatial questions, which are intended to add context to the spray patterns.

In order to explore these themes, each participant was directed to a web address, which redirected them to one of five different *Map-Me* surveys using a pseudorandom number generator. Each of these

surveys comprised exactly the same questions, but there were differences in the survey itself such as: the order of the questions, initial map location, initial zoom level and initial base map (road map, aerial photography, terrain map etc). In this way, analyses may be undertaken in order to identify the impact of these changes upon the data created by the participants, as well as an overall assessment of the way in which participants use the *Spraycan* platform. Results of these analyses may then be compared with the results of the questionnaire in order to validate these native, automated approaches to the investigation of HCI with a PPGIS.

### 3. Preliminary Results

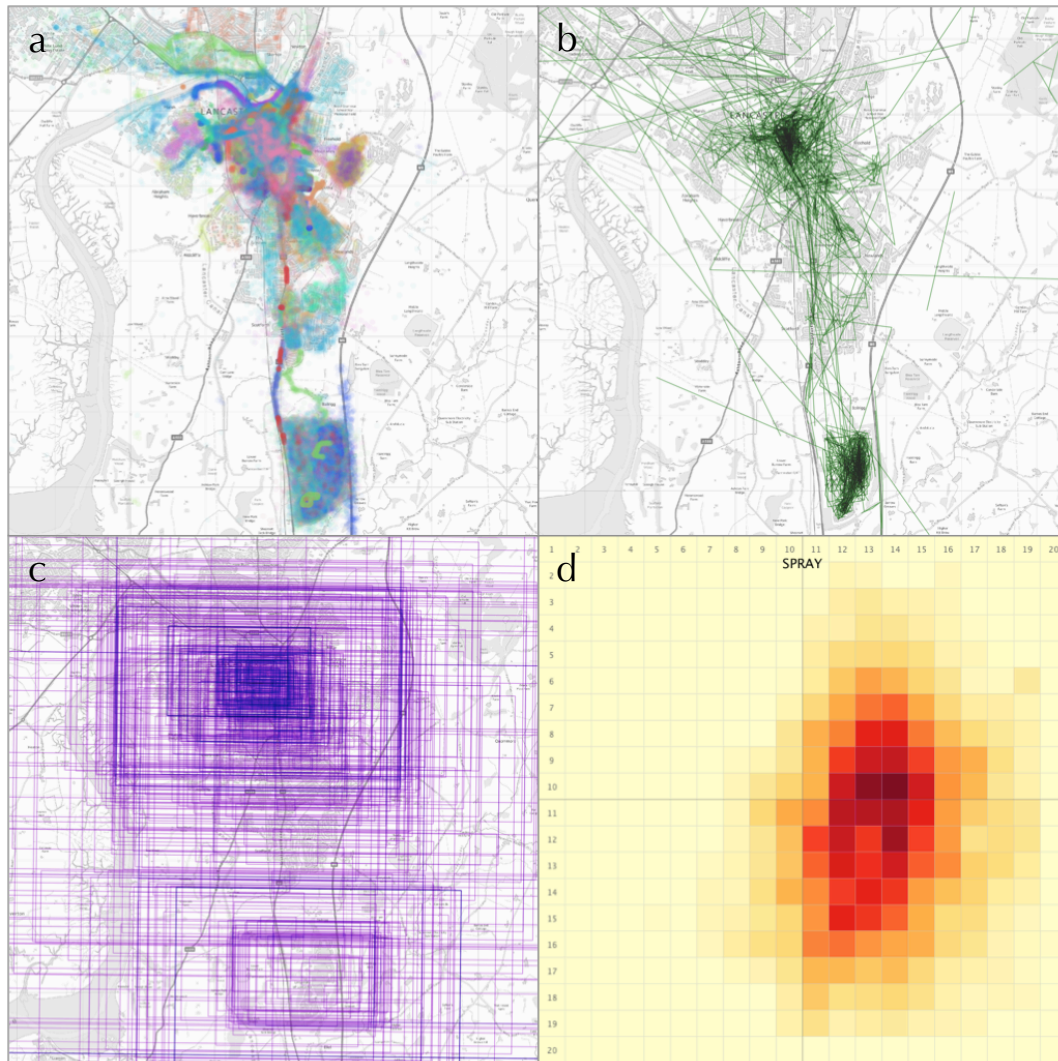
Figure 3 demonstrates the overall ‘consensus’ of where participants identified as ‘safe’ and ‘unsafe’ around Lancaster, calculated as described by Huck et al. (2014). The patterns here are very much ‘as expected’, with Lancaster town centre, the University campus, the Infirmary (hospital) and areas of parkland generally considered to be ‘safe’ places; whilst outlying areas including some housing estates and main roads are generally considered to be more ‘unsafe’.



**Figure 3:** A map displaying the ‘consensus’ of all participants of whether or not an area is considered to be ‘safe’ or ‘unsafe’ around Lancaster (calculated as per Huck et al. 2014).

Some examples taken from the preliminary stages of the analysis described in this abstract are given in Figure 4. Figure 4a displays all of the different ‘places’ (each in a different random colour) that were identified by participants within the study area. Each of these *places* has been named and reasoned as ‘safe’ or ‘unsafe’ by each participant for use in further analysis. Figure 4b shows the track taken by the *Spraycan* tool (green) all of the participants, allowing interrogation of how participants moved around the map. The expected focus upon the town centre (upper hotspot) and University campus (lower hotspot) are clearly visible. Figure 4c illustrates the position and bounds of

participants' viewpoints (map windows) whilst they were spraying, demonstrating the tendency for participants to zoom in on the town centre (the upper 'hotspot') and the University campus (the lower 'hotspot'), whilst taking a broader view of the areas in between (the larger squares). Finally, Figure 4d demonstrates the relative position of all participants' spray within the viewport, irrespective of bounds or location, with darker colours representing more spray. This is used to identify any potential sources of bias within the PPGIS, and in this case shows a general preference for spraying to the right of centre. Potential causes of this could include, for example, the position of the text boxes and controls on the right of the screen (Figure 2), though further work would be required in order to confirm this.



**Figure 4:** A selection of sample images taken from the preliminary stages of this analysis (described above).

#### 4. Conclusion

Most previous approaches to the evaluation of PPGIS utilise 'traditional' HCI methodologies such as questionnaires, audio recordings, screen capture, interviews or log-file analysis. The *Spraycan* platform, however, provides the facility to undertake some investigation into participant interactions

‘natively’ using the data that are already collected as part of the analysis, thus reducing or even removing the requirement for additional evaluation techniques, as well as permitting the retrospective analysis of data collected in previous surveys. This work will aim to validate these ‘native’ approaches by comparison of findings with those of a traditional questionnaire, and to learn more about *how* participants use the *Spraycan*; gaining deeper insights into the factors that influence the way that people interact with vague geographic problems.

## 5. Biography

Jonny Huck is a 4th year part-time PhD student researching Geographical Information Science jointly with Imagination Lancaster and the Lancaster Environment Centre at Lancaster University. His interests include web mapping, the representation of ‘place’, geospatial visualisation, and the application of new technologies to spatial analysis.

Dr Duncan Whyatt is a Senior Lecturer in GIS within the Lancaster Environment Centre, Lancaster University. His research interests span social and environmental applications of GIS, with specialisms in air pollution.

Dr Paul Coulton is a Senior Lecturer in Design within Imagination Lancaster. His research interests are primarily around experience design, interaction design, and design fictions. His research often encompasses an ‘in the wild’ evaluation methodology, utilising ‘app stores’ and social networks as experimental platforms.

## References

- Bugs, G., Granell, C., Fonts, O., Huerta, J., & Painho, M. (2010). An assessment of Public Participation GIS and Web 2.0 technologies in urban planning practice in Canela, Brazil. *Cities*, 27(3), 172-181.
- Carver, S. J. (1991). Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information System*, 5(3), 321-339.
- Demšar, U. (2007). Combining formal and exploratory methods for evaluation of an exploratory geovisualization application in a low-cost usability experiment. *Cartography and Geographic Information Science*, 34(1), 29-45.
- Elwood, S. (2006). Critical issues in participatory GIS: Deconstructions, reconstructions, and new research directions. *Transactions in GIS*, 10(5), 693-708.
- Haklay, M., & Tobón, C. (2002). Usability Engineering and PPGIS-Towards a Learning-improving Cycle. <http://discovery.ucl.ac.uk/16784/1/16784.pdf>. Retrieved 24th October 2014.
- Haklay, M., & Tobón, C. (2003). Usability evaluation and PPGIS: towards a user-centred design approach. *International Journal of Geographical Information Science*, 17(6), 577-592.
- Huck, J., Whyatt, D., & Coulton, P. (2013). Development and application of a "spray-can" tool for fuzzy geographical analysis. In *Proceedings of the GIS Research UK 21st Annual Conference*.
- Huck, J., Whyatt, D., & Coulton, P. (2014). *Spraycan: A PPGIS for capturing imprecise notions of place*, *Applied Geography*, 55, 229-237.
- Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-

3), 185-204.

Rinner, C., & Bird, M. (2009). Evaluating community engagement through argumentation mapsöa public participation GIS case study. *Environment and Planning B: Planning and Design*, 36, 588-601.

Sieber, R. (2006) Public Participation Geographic Information Systems: A Literature Review and Framework. *Annals of the Association of American Geographers*. 96 (3), 491-507.

Zhao, J., & Coleman, D. J. (2007). An empirical assessment of a web-based PPGIS Prototype. In Submitted to the 2007 Annual Conference of the Urban and Regional Information Systems Association.

# Abstract Feature Representation as a Cartographic Device for Mixed-Reality Location-Based Games

J. J. Huck<sup>1</sup>, P. Coulton<sup>1</sup>, A. Gradinar<sup>1</sup> and D. Whyatt<sup>2</sup>.

<sup>1</sup>Imagination Lancaster, LICA, Lancaster University

<sup>2</sup>Lancaster Environment Centre, Lancaster University

March 4, 2014

## Summary

This paper presents a number of cartographic design solutions to the creation of a map for the mixed-reality location-based game ‘*Pac-Lan: Zombie Apocalypse*’. The research-purpose of this game is to explore ways in which players may be encouraged to become less reliant upon the device screen during gameplay, and so more fully engaged with the physical environment in which the game is played. In this paper we specifically consider approaches to designing the game-map in such a way as to discourage players from becoming solely reliant upon it for navigation, and instead interact more with their surroundings during gameplay. This paper therefore considers four maps as potential solutions for ‘*Pac-Lan: Zombie Apocalypse*’, which serve explore the use of abstract feature representation as a cartographic device to encourage player engagement with the landscape. A description of the design rationale for each of the maps is presented here, along with some preliminary findings from user evaluation of the maps against a defined set of design goals.

**KEYWORDS:** Cartography, Location Based Games, Feature Abstraction, Mixed-Reality Spaces, Landscape Legibility.

## 1. Introduction

This paper presents a *research through design* (Gaver, 2012) approach to the creation of a web-map for the mixed-reality Location-Based Game (LBG) ‘*Pac-Lan: Zombie Apocalypse*’ which operates on the Android mobile platform. Based upon the popular ‘*Pac-Man*’ arcade game (Namco, 1980; Figure 1), this sequel to the original ‘*Pac-Lan*’ (Rashid et al., 2006; Coulton et al. 2006a; Coulton et al., 2006b) is played by up to five players: one of which takes the role of *Pac-Lan*, with the remaining four taking the roles of the ‘ghosts’ (*Blinky*, *Pinky*, *Inky* and *Clyde*). Players run around a real-life ‘maze’ defined by the physical landscape (buildings, woodland, water-bodies etc.) and earn points by ‘tagging’ physical ‘pellets’. Frisbees fitted with Near Field Communication (NFC) tags that are attached to street furniture represent the physical ‘pellets’, and ‘tagging’ them is simply a matter of physically touching them with their NFC-enabled smartphones upon which the LBG application is running. Similarly, players may also ‘capture’ each other by ‘tagging’ them in a similar way, this time using an NFC tag attached to each player’s back. The winner of the game is the player with the greatest number of points when the game ends, which is either when the time runs out, when all of the ‘pellets’ have been ‘tagged’ by *Pac-Lan*, or when *Pac-Lan* is captured by a ghost.

The research purpose of ‘*Pac-Lan: Zombie Apocalypse*’ is to explore the legibility of mixed-reality spaces, specifically through the concept of the ‘*Dichotomy of Immersion*’. Where *immersion* refers to the degree of involvement that a player has with a computer game (Brown, 2004), the *Dichotomy of Immersion* describes the peculiar situation created in LBG’s whereby a player’s attention is constantly divided between the physical world within which the game is being played (the physical world), and the screen of their mobile device (the digital world). The split of attention between the physical and digital game components is usually dominated by interaction with the screen at the expense of interaction with the landscape, which can limit engagement with (and therefore immersion into) the LBG (Zhang et al, 2012), as well as contribute to the low uptake of LBG’s by members of the public (Lund et al, 2012). It is hoped, therefore, that addressing this imbalance in this research will

contribute towards the development of more engaging LBG's, with a higher level of uptake.



**Figure 1:** Screenshot of a classic version of Pac-Man (Namco, 1980). *Reproduced from* <http://en.wikipedia.org/wiki/File:Pac-man.png>.

The game-map is a central and essential component of any LBG and so has been one of the major areas of focus within this research. In order to discourage user interaction with (and reliance upon) the mobile screen, the map must be designed in such a way as to encourage users to glance at the map and navigate in a ‘head-up’ manner, using their surroundings; as opposed to navigating in a ‘head-down’ manner, looking at the map throughout their journey, as is typically the case when users navigate using a mobile phone. Instead of simply using a Google Map or similar as a game-map, a number of bespoke cartographic designs have therefore been developed in order to facilitate greater immersion into the physical game. It is the design of these game-map that will be addressed within this paper, specifically in the context of the use of abstract feature representation as a cartographic device.

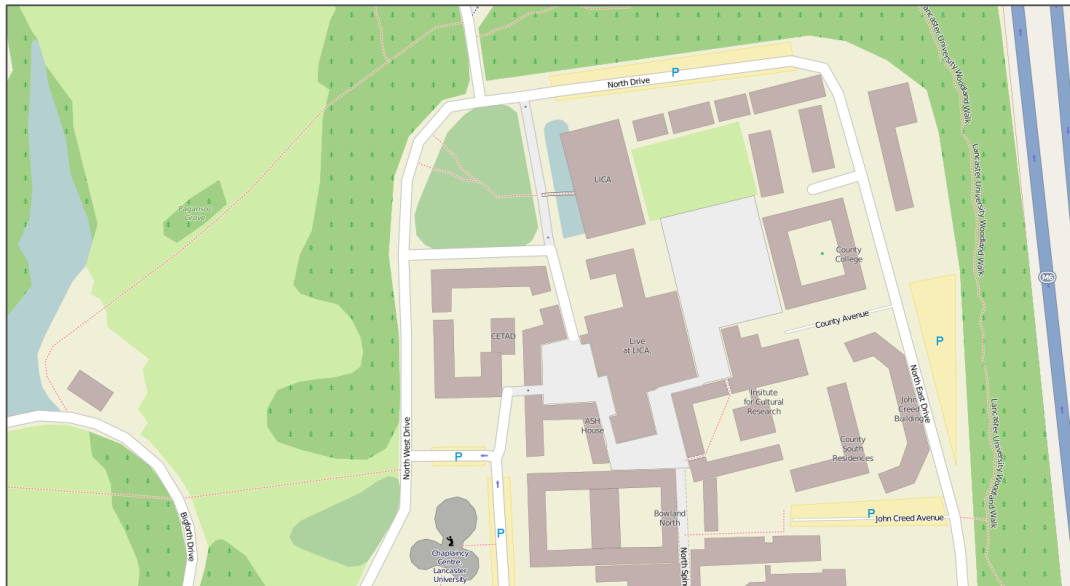
## 2. Design Solutions

The game-map for ‘*Pac-Lan: Zombie Apocalypse*’ has three design goals: to promote immersion into the game through the use of a suitable aesthetic; to perform well within the context of a mixed-reality LBG (i.e. outdoors); and to encourage players to navigate ‘head-up’ rather than ‘head-down’ during gameplay. The latter of these design goals is the most important given the context of the wider ‘*Pac-Lan*’ research aims, and in this paper it will be approached by the use of abstract feature representation as a cartographic device. It is hypothesised that a small amount of abstraction in map features may encourage players to ‘look up’ more and verify what they see on the map against their physical surroundings, thus increasing their engagement with their physical surroundings. This is in contrast with the use of a more traditional (precise) map, which will not require validation against the landscape; or a map that is ‘too abstract’, which may be too difficult to read quickly whilst playing, thus increasing interaction with the screen at the expense of the landscape.

As already discussed, this paper will consider four potential solutions to the above design goals, along with some preliminary findings relating to user evaluation of the maps. The maps have been all created using data from OpenStreetMap (<http://www.openstreetmap.org/>), or derived from it within PostGIS (<http://postgis.net/>), and have been rendered using Mapnik (<http://mapnik.org/>). For the purpose of comparison, all maps are shown at the same standard orientation (north at the top), zoom level (17, equivalent to a geographic scale of 1:4514) and extent (showing part of the Lancaster University campus). This view is also shown as drawn in OpenStreetMap in Figure 2 for the purposes

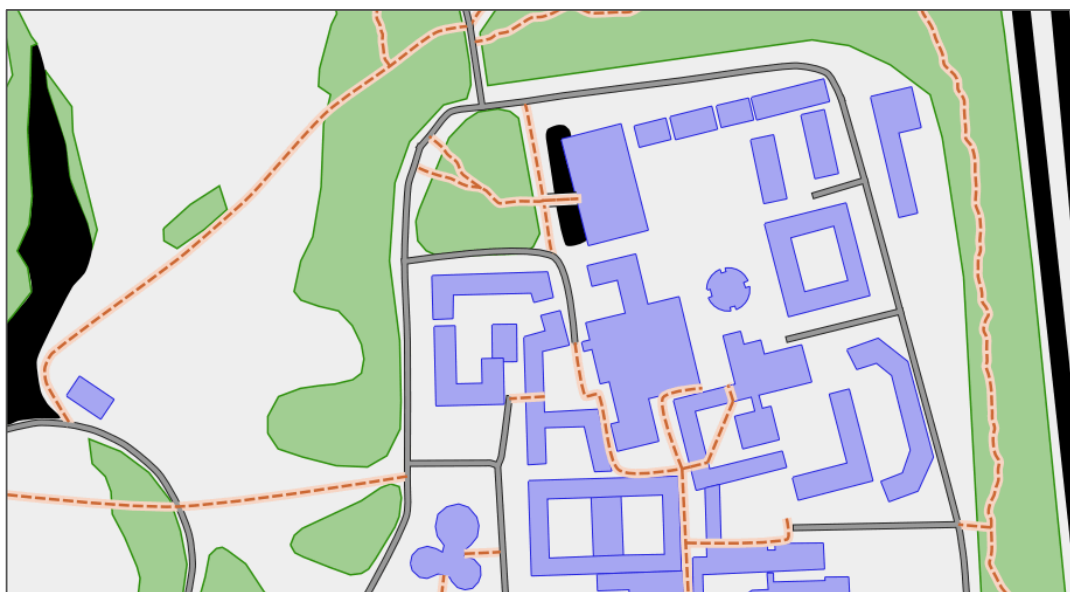


of comparison. None of the figures include a legend, north arrow, scale bar or similar, as such features are not typically provided in LBG's or other mobile applications. In-keeping with the above design goals, all of the maps are very limited in the number of features that they contain, with only 4 or 5 feature classes typically included to facilitate easy reading. For the same reason, augmentations such as labels and points of interest have also been omitted from all of the map designs.



**Figure 2:** The 'standard' Mapnik-rendered *OpenStreetMap* map style.

## 2.1 The 'Anti-Glare Map'



**Figure 4:** The *Anti-Glare Map*.

The first map to be presented is the *Anti-Glare Map* is intended primarily to perform well outdoors, and investigate the alternate hypothesis that a clear and precise map may be more successful than an



abstract map in encouraging ‘head-up’ play as players will be able to digest spatial information more quickly. As such, the *Anti-Glare Map* does not exhibit any level of abstract feature representation, and so will act as a ‘control’ in this investigation with regard to the effectiveness of this technique. The *Anti-Glare Map* utilises a triadic colour scheme in order to gain a high degree of contrast between features whilst maintaining colour harmony. Features are divided into five classes: ‘building’, ‘road’, ‘footpath’, ‘trees’ and ‘hazard’, and a light-grey background was chosen because lighter background colours are typically less susceptible to screen glare. Hazards are filled with black, accenting them in comparison to the background and other features, whereas the other features (those using the triadic colour scheme) all include an accent using a darker shade of the same colour. This accent is used to outline all of the features except footpaths in order to make them ‘pop’ out from the light background, and is used as a dashed centreline for the footpaths, in order to create a contrast between the footpaths and the roads.

## 2.2 The ‘Pac-Map’

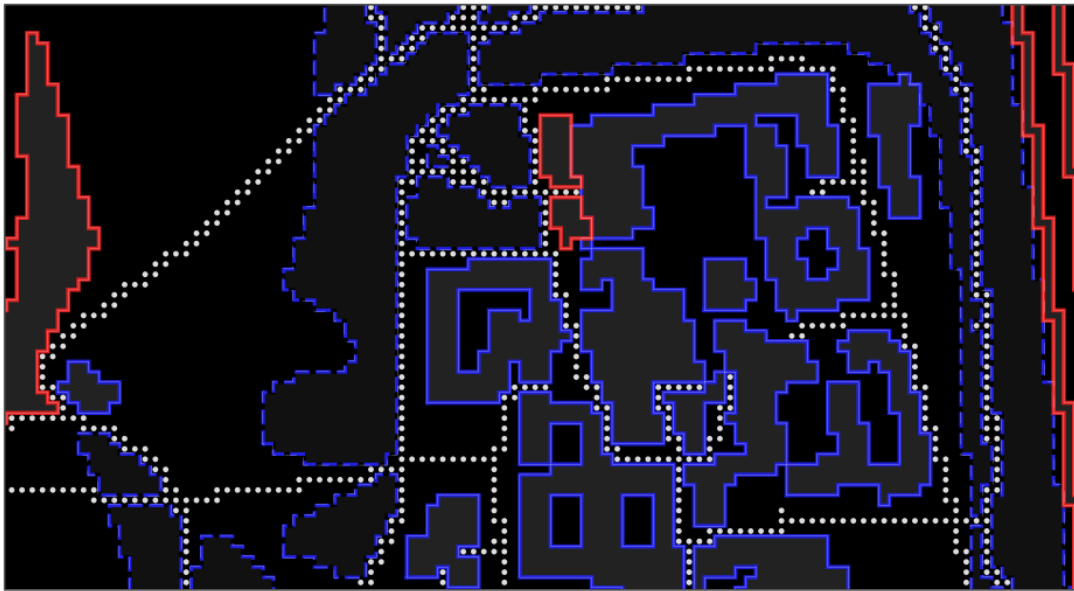


Figure 3: The *Pac-Map*.

The *Pac-Map* (Figure 3) is designed primarily to match the aesthetic of the classic Pac-Man arcade game (Namco, 1980; Figure 1). Abstract feature representation has been achieved in this map by generalising all nodes to the nearest 10m, and forcing vertices to be oriented either north-south or east-west: resulting in abstract features that also contribute to the game aesthetic. There are only 4 feature classes included in the map: ‘building’, ‘path’, ‘trees’ and ‘hazard’, all of which are rendered using styles directly inspired by the Pac-Man game (Figure 1). Trees and buildings are drawn in the same blue as the Pac-Man maze, and a complementary red has been used to mark out hazards. In order to ensure that the dark palette employed by this map performs well outdoors, the lines in the map are thick, with very fine white lines drawn into the blue and red in order to increase their contrast with the black background. Pathways (including roads and footpaths) have been marked out with white dots, once again to gain contrast with the dark background, whilst also reflecting the ‘pellets’ that Pac-Man collects from within the maze in the original game.

### 2.3 The ‘RPG Map’

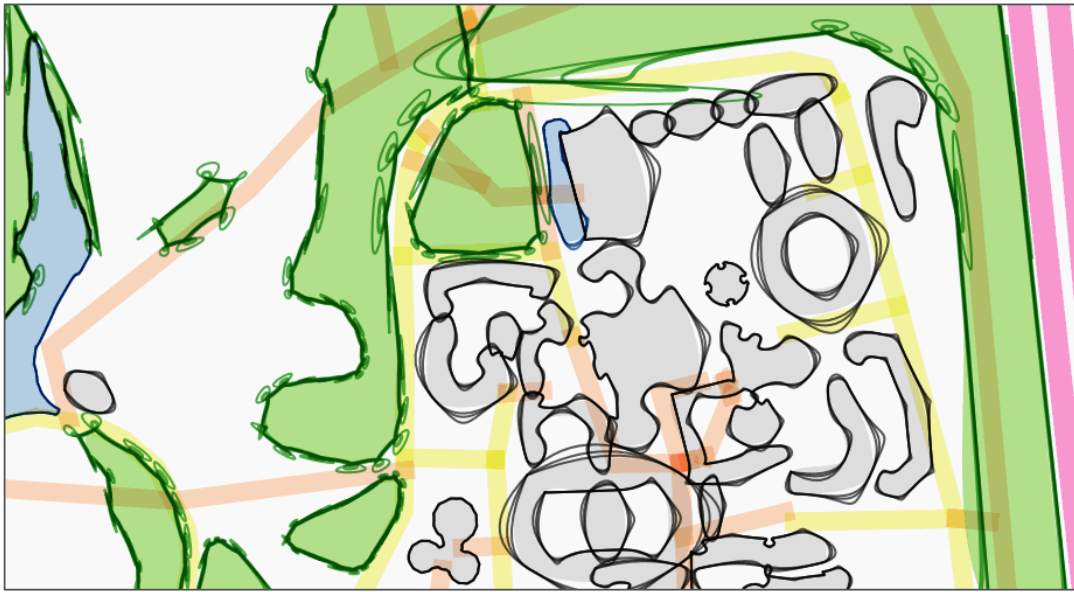


**Figure 5:** The *RPG Map*

The *RPG Map* is inspired by the ‘classic’ Role Playing Games (RPG’s) of the 1980’s and 1990’s. The data has been abstracted into a grid of 20m cells, each of which can only contain one of five feature classes: ‘building’, ‘road’, ‘water’, ‘trees’ or ‘hazard’. Cells were then dissolved into contiguous areas of each data type, and coloured using tiled textures collected from freely available online sources. The use of a coarse 20m grid gives this map a greater level of abstract feature representation than the *Pac-Map*, therefore making it more difficult to rely upon for navigation, in order to investigate the effect that this has upon the players’ interactions during gameplay. The coarse grid, RPG-style textures and playful features (e.g. the use of a ‘lava’ texture to denote hazards) lend a definite ‘game aesthetic’ to the map, but in less-specific manner to the *Pac-Map*, permitting exploration as to the effect of this upon players’ perceived level of immersion.

### 2.4 The ‘Sketchy Map’

‘Sketchiness’ as a device for enhancing the aesthetic or narrative qualities of cartographic outputs has been explored previously by Wood et al. (2012), and has also been employed by Griffen et al. (2014) as a visual variable in maps. In this case, however, ‘sketchiness’ is used as an alternate approach to abstract feature representation, acting to obscure the precise position and shape of geographic features. The ‘hand-drawn’ or ‘sketchy’ effect on the polygons has been achieved by a combination of polygon smoothing, line smoothing, multiple-overlay and image composite operations in order to give the impression that they have been drawn using felt-tip pens (akin to the approach first suggested by Ashton, 2012). Conversely, the line features were simplified using the Visvalingam-Whyatt line generalisation algorithm (Visvalingam and Whyatt, 1993), and overlaid using transparency and image composite operations in order to give the appearance of having been drawn using highlighter pens. This approach will allow the comparison of abstract feature representation arising from ‘sketchiness’ against the grid-based approaches used in the *Pac-Map* and the *RPG-Map* as a device for the encouragement of ‘head-up’ gameplay. The main difference with this approach is that the level of abstraction varies from feature to feature as opposed to being uniform across the dataset as is the case in the grid-based approaches, which may prove more disorientating for users. The ‘hand-drawn’ aesthetic promotes a ‘playful’ feel to the map, but without specifically evoking a ‘game’, permitting further investigation into the effect of the map aesthetic upon game immersion.



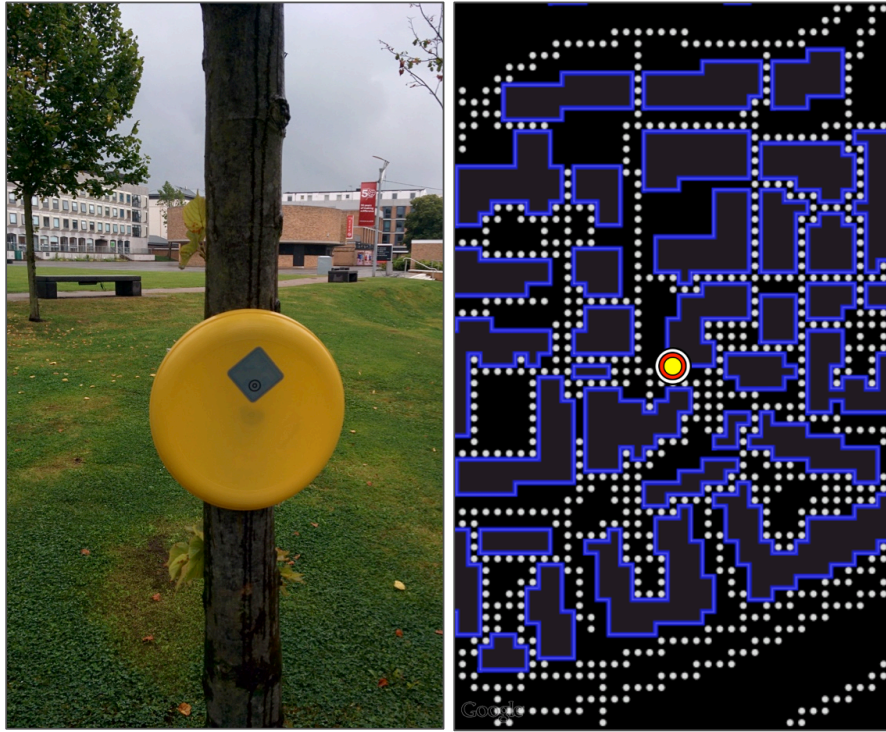
**Figure 6:** The *Sketchy Map*

### 3. Methodology

For the purpose of assessing the four proposed design solutions, a for-purpose Android-based LBG has been developed that is less complex than the full *'Pac-Lan: Zombie Apocalypse'* game, and so allows a more detailed exploration of the maps without distractions arising from other game elements. This LBG requires users to navigate between (and physically 'tag') 20 NFC-enabled 'pellets' (Fig 7a) that have been placed around an area of student accommodation at Lancaster University. Only one 'pellet' is displayed at once on the on-screen map, which disappears when tagged as the map re-centres onto the next 'pellet' (screenshot in Figure 7b). The map style changes after every fifth 'pellet' and a pseudorandom number generator determines both the order of the maps, and the order in which the 'pellets' must be 'tagged'. In order to view the map, users must hold down the volume button on the mobile device, and the map is hidden from view as soon as this button is released. In this way, it is possible to keep a log of the amount of time during which each map is viewed, in addition to the location and movements of each player using the on-board GPS receiver in the mobile device. Following gameplay, each player was given a semi-structured questionnaire and is interviewed in order to gain feedback relating to the quality of each map against the above design goals.

### 4. Preliminary Results

Whilst still in the process of data collection, we are able to demonstrate results from the first 8 players as preliminary findings. The quantitative data collected in the log files confirms that, as expected, players spent the least amount of time looking at the (least abstract) *Anti-Glare Map* (c. 31% of the time), and the most amount of time spent looking at the (most abstract) *RPG Map* (c. 47% of the time). Of the remaining two maps, more time was spent looking at the *Sketchy Map* (c. 44% of the time), with its variations in level of abstraction from feature to feature, in comparison with the uniformly abstract *Pac-Map* (c. 38% of the time). Whilst these findings are interesting (albeit expected), the amount of time spent looking at the map is unlikely to be inversely correlated to the level of engagement with the landscape. As such it is the qualitative data relating to the players' perceptions of the impact that the maps had upon their engagement with their surroundings that is therefore of more interest to this research.



**Figure 7:** **a:** A physical ‘pellet’ for players to ‘tag’, complete with NFC tag. **b:** A screenshot of the LBG used to evaluate the maps, displaying the *Pac-Map* (covering the actual test area) and location of the next ‘pellet’ to tag.

Through a simple vote within the questionnaire, players identified the *Pac-Map* and *Sketchy Map* as being equally the “*most suitable map for use in an LBG*”, whereas the *RPG-Map* was considered to be the best for generating engagement with the environment. The reasons for the latter, however, were very clear in the associated comments, with the *RPG Map* being unanimously considered to be “*very difficult*”, and “*frustrating*” to use, with one player even suggesting that it was “*totally unusable*”. This, along with a complaint that the map suffered from screen glare, is a clear suggestion that the *RPG-Map* is ‘too abstract’, and therefore not well suited to an RPG. These comments were interestingly contrasted with those relating to the *Anti-Glare Map*, which was described as “*too easy*” by two users, and caused one user to feel they “*spent too much time looking at the map because it was easy to [navigate with]*”. These preliminary findings lend support to the hypothesis that a map exhibiting abstract feature representation can lead users to engage more with their surroundings, and that too great a level of abstraction can become counter-productive in this regard.

Of the remaining maps, the *Pac-Map* seemed to be considered as more well balanced: “*I could tell what things were represented but still looked up*”; and as an attractive or well-suited map design: “*Nice feel*”, “*It’s like the original Pac-Man*”. Similarly, the *Sketchy Map* was considered as “*pleasing on the eye*” and “*more fun*”, as well as “*showed just enough to navigate but required you to look around*” and “*challenging enough to keep it interesting*”. These comments suggest that both were well received by users and fulfilled their desired purpose well, again lending support to the above hypothesis.

When the quantitative findings are also considered, however, the *Pac-Map* appears to have performed best across the three principal design goals: to promote immersion into the game through the use of a suitable aesthetic; to perform well within the context of a mixed-reality LBG (i.e. outdoors); and to encourage players to navigate ‘head-up’ rather than ‘head-down’ when playing a LBG.



## 5. Conclusion

The consensus from the preliminary user feedback suggests that the *Pac-Map* was the most effective with regard to the intended design goals, and was even described as “*the best suited to a game*” by one user. Overall the data from this preliminary analysis seem to suggest that using abstract feature representation may indeed be a suitable approach to the creation of maps for the promotion of immersion within LBG’s, and that the application of ‘too much’ abstraction will start to degrade the quality of the map for this purpose. Whilst it is clear that further user testing and analysis is required before any firm conclusions may be drawn, it would appear that maps produced in this manner can discourage players from navigating ‘head-down’; forcing them to engage with the landscape in order to find their next target.

This *research through design* project has begun to explore the design of a map that will encourage immersion into a mixed-reality LBG through the promotion of ‘head-up’ navigation. Each of the design options presented in this abstract will continue to be evaluated by users in order to assess them against the above design goals so that more firm conclusions may be drawn from this work. It is intended that the resulting design knowledge will aid other cartographers in the design of maps for mixed reality LBG’s, and contribute towards increased uptake of LBG’s within the gaming ecosystem.

## 6. Biography

Jonny Huck is a 4<sup>th</sup> year part-time PhD student researching Geographical Information Science jointly with Imagination Lancaster and the Lancaster Environment Centre at Lancaster University. His interests include web mapping, the representation of ‘place’ in GIS, cartography, and the application of new technologies to spatial analysis.

Dr Paul Coulton is a Senior Lecturer in Design within Imagination Lancaster. His research interests are primarily around experience design, game design, and design fictions. His research often encompasses an ‘in the wild’ evaluation methodology, utilising ‘app stores’ and social networks as experimental platforms

Adrian Gradinar is a 2<sup>nd</sup> year PhD student at Lancaster University, researching around The Internet of Things within the Digital Public Space, especially how digital information can be integrated with familiar objects. He is also interested in how digital games could be interconnected with the physicality of the surrounding world.

Dr Duncan Whyatt is a Senior Lecturer in GIS within the Lancaster Environment Centre, Lancaster University. His research interests span social and environmental applications of GIS, with specialisms in air pollution.

## 7. References

- Ashton, A. (2012) Sketchy Maps with Geometry Smoothing. <https://www.mapbox.com/blog/sketchy-maps/>. Accessed 03/03/2015.
- Brown, E. and Cairns, P. (2004) A grounded investigation of game immersion. ACM Press.
- Coulton, P., Rashid, O., and Bamford, W. (2006a) Experiencing ‘touch’ in mobile mixed reality games. In proceedings of the International Conference in Computer Game Design and Technology.
- Coulton, P., Rashid, O., Bamford, W., & Edwards, R. (2006b). Running with the PAC. In Proceedings of the 1st World Conference for Fun'n Games.

- Gaver, W. (2012). What should we expect from research through design?. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 937-946). ACM.
- Griffin, A. L., Spielman, S. E., Jurjevich, J., Merrick, M., Nagle, N. N., & Folch, D. C. (2014) Supporting Planners' Work with Uncertain Demographic Data. In Proceedings of VIS 2014, 9-14 November 2014, Paris, France.
- Lund, K, Lochrie, M & Coulton, P (2012), 'Designing Scalable Location Based Games that Encourage Emergent Behavior: Special issue on Ambient and Social Media Business and Application (Part I)' International Journal of Ambient Computing and Intelligence, vol 4, no. 4, pp. 1-20., 10.4018/jaci.2012100101
- Rashid, O., Bamford, W., Coulton, P., Edwards, R., and Scheible, J. (2006). PAC-LAN: mixed-reality gaming with RFID-enabled mobile phones. *Comput. Entertain.* 4, 4, Article 4.
- Visvalingam, M., & Whyatt, J. D. (1993). Line generalisation by repeated elimination of points. *The Cartographic Journal*, 30(1), 46-51.
- Zhang, L., & Coulton, P. (2011). Using deliberate ambiguity of the information economy in the design of a mobile location based games. In *MindTrek '11 Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments.* (pp. 33-36). New York: ACM. 10.1145/2181037.2181044

# Development of public transport accessibility in the Czech Republic

Ivan I<sup>\*1</sup>

<sup>1</sup>Institute of Geoinformatics, VSB – Technical University of Ostrava, Czech Republic

November 7, 2014

## Summary

The convenient system of public transport services belongs to key factors influencing the final decision of public transport use. This paper analyses the development of the level of public transport services, utilising data from Database of Public Transport Connections developed and maintained by the authors. These databases contain regularly updated data about inter-municipality public transport connections since 2007 which are suitable for commuting. The development of public transport accessibility in the Czech Republic is analysed by applying spatial analysis methods. Results indicate more complicated relationships between public transport accessibility and local socioeconomic changes.

**KEYWORDS:** public transport, accessibility, Czech Republic.

## 1. Introduction

Public transport has a long history in the Czech Republic and compared to another similar in the central Europe countries (e.g. Poland or Slovenia), high offer of public transport connections still remains also due to financial subsidies by government. This fact caused sufficient level of public transport accessibility to all municipalities what belongs to basic idea of law 194/2010 about public passenger transport services that generally ensures basic transport serviceability of a region but without specific number of bus or train connections. Anyway due to rationalization and short cuts to public transport offer, the level is getting lower. Even so development of modal split in the Czech Republic copies similar development in European countries. The main aspect is stagnant or decreasing share of public transport use (decreased from 17.5% in 1995 to 10.9% in 2013; without city public transport) and increasing number of journeys by cars.

This paper analyses the development of level of public transport services, utilising data from Database of Public Transport Connections developed and maintained by the authors. These databases contain regularly updated data about inter-municipality public transport connections since 2007 which are suitable for daily commuting. The public transport accessibility is interconnected with local socioeconomic situation – public transport system reflects actual human needs and demands and reversely the society is influenced by the public transport accessibility. Such iteration process may quickly increase local disparities if it is not regulated. The development of public transport accessibility in the Czech Republic is analysed by applying methods of spatial analysis of current situation and change since 2007 to 2014. Results indicate more complicated relationships between public transport accessibility and local socioeconomic changes.

## 2. Database of Public Transport Connections

Czech Republic has the advantage that all time tables are centralized in central information system

---

\* igor.ivan@vsb.cz

maintained by CHAPS Ltd. Valid time tables together with developed application TRAM are able to search valid public transport connections between all municipalities within 100 kilometres (Euclidean distance) and what creates the database with more than 12.5 million of municipality combinations. Several variables are searched in time tables for each combination of municipalities (e. g. travel time, number of changes, price, and existence of return connection) for five times (to 6, 7, 8, 14 and 22 o'clock). These times define the beginnings of three work shifts. Valid public transport connection must meet defined criterias. Travel time is smaller than 90 minutes, number of changes is smaller than 5, arrival time cannot be earlier than 60 minutes before, and departure time from origin cannot be earlier than 120 minutes before arrival to destination (more in Ivan et al., 2013 or Horák et al., 2014). For purposes of this paper, timetables for trains and buses (no urban transport) valid in March 2007, 2011 and 2014 have been utilized.

### 3. Development of public transport accessibility

The transport accessibility was assessed mainly using rate of accessible municipalities  $RA$ , which is defined as the number of accessible municipalities per number of all municipalities tested for existing transport connection (in %) on certain time in case of one-way travelling.

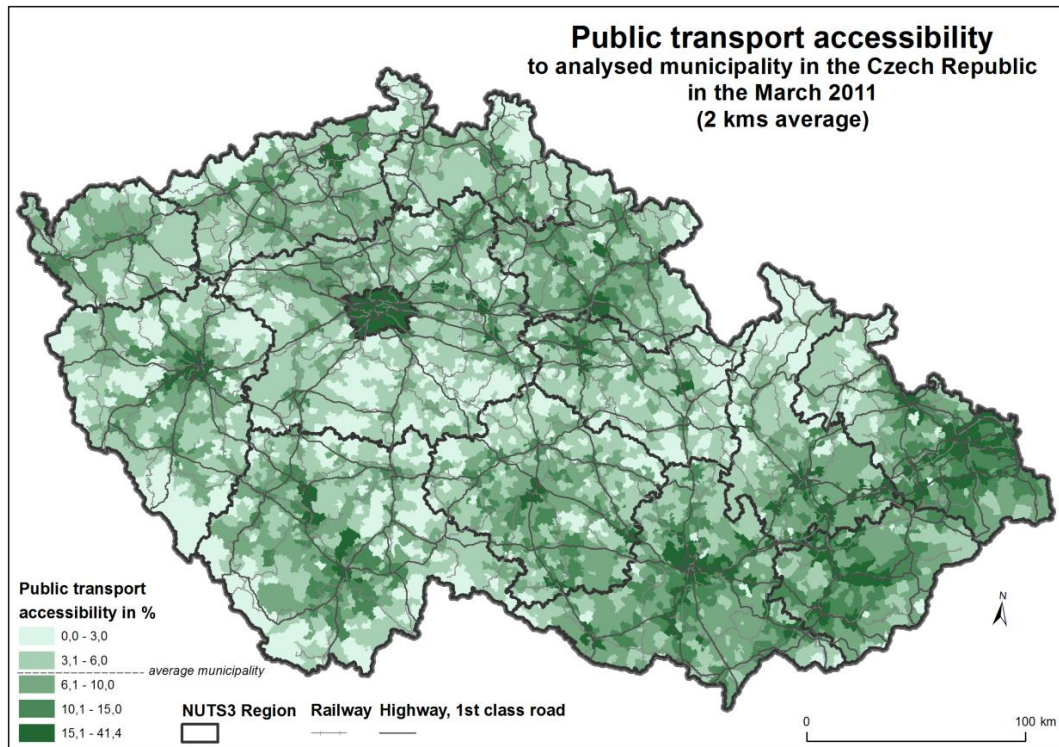
$$RA_{h,i} = \frac{NMA_{h,i}}{NMT_{h,i}} * 100 \quad (1)$$

where  $NMA$  is the number of accessible municipalities,  $NMT$  is the total number of municipalities within a given Euclidean distance (i.e. 100 km),  $h$  is hour and  $i$  is the index of municipality.

The map (Fig. 2) shows spatial distribution of public transport accessibility indicator. Potential differs from 0 to 100, where 100 means that there is a possibility to travel to analysed municipality from all municipalities within 100 kilometres, so the ideal destination for commuters from surrounding municipalities. However, many times commuters are using a public transport stop in nearby municipality within walking distance. Therefore spatial filter has been used and final value is equal to the average of potentials of municipalities within 2 kilometres. This distance is considered as maximal walking distance to a public transport stop.

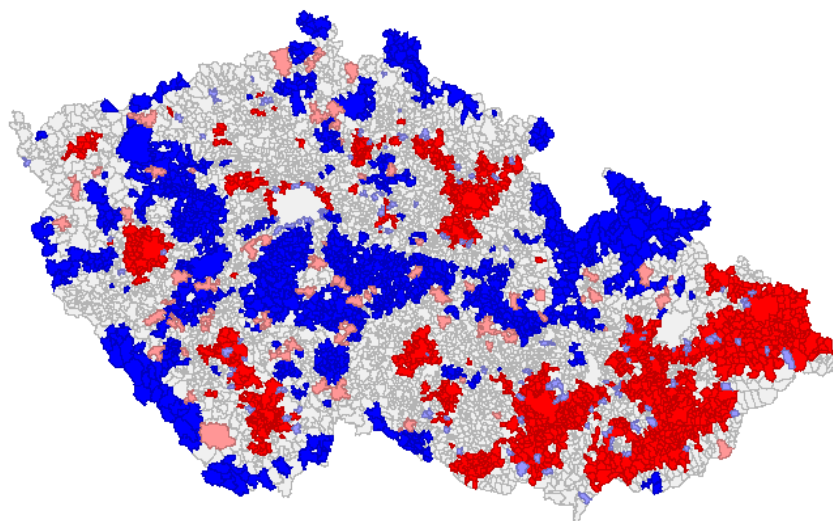
Average municipality in Czechia is accessible for any of analysed five arrival times from 6% of surrounding municipalities within 100 kilometres. Spatial distribution is depicted in the map below (figure 2) and spatial clustering (effect of second order) as well as spatial trend (effect of first order) from west to east are evident in this map. Spatial trend is confirmed by positive and statistically significant ( $p = 0.01$ ) correlation between public transport potential and x coordinate ( $R = 0.284$ ). The more east is the municipality the better is its accessibility. Roads (highways, motorways and first class roads) and railways have influence on public transport potential too. Average of theoretical transport potential of municipalities within 2 kilometres from railway or road network is 6.9%. This is higher than national average. Difference in public transport potential between municipalities within and farther than 2 kilometres from roads and railways (4.9%) has been proved by ANOVA. Bigger influence on average public transport potential is caused by roads with an average equal to 7.5%. Average for railways is smaller 7%.





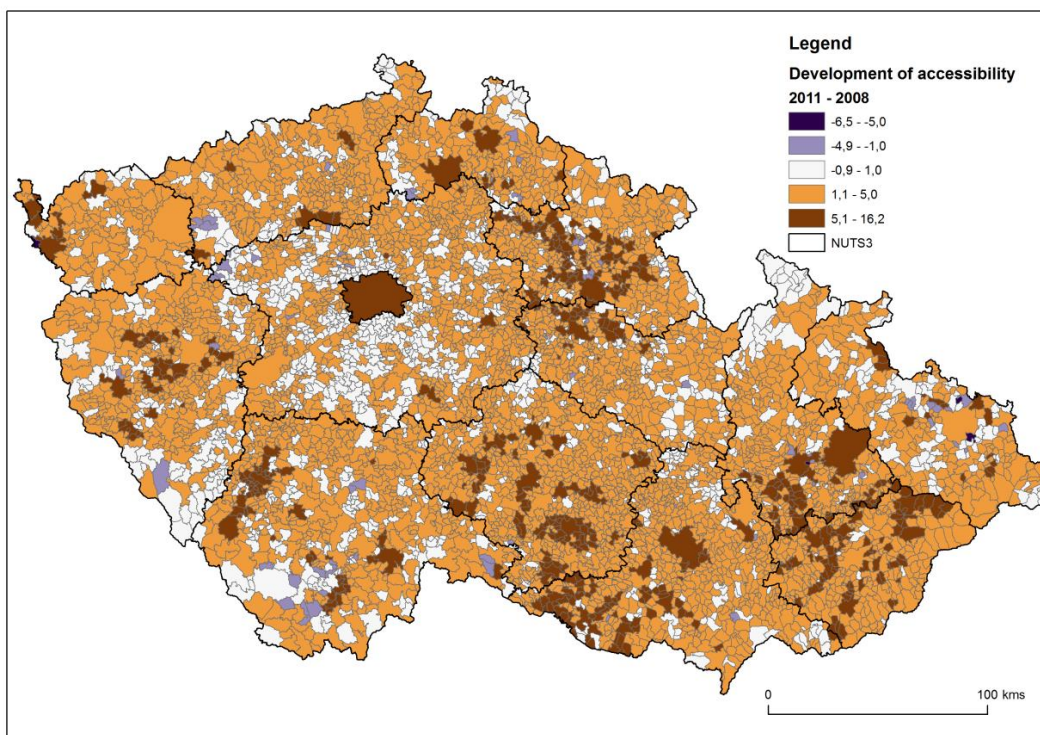
**Figure 1** Rate of accessible municipalities in March 2011

Obviously if a municipality has a good public transport accessibility, surrounding municipalities should have higher accessibility too. These municipalities could share the same public transport link. Spatial clustering is analysed using local autocorrelation of public transport accessibility. Moran's  $I$  is 0.575 what proves clustering of municipalities with higher or lower accessibility. The map in figure 2 depicts these clusters of municipalities. Red colour characterizes municipalities with higher public transport accessibility surrounded also by municipalities with higher accessibility. These municipalities are concentrated mainly in the eastern parts of Czechia what confirms previous hypothesis about western-eastern trend. On contrary blue colour describes municipalities with low accessibility surrounded by municipalities also with low level. These municipalities create several clusters situated mainly in the western part of Czechia and in the northern part of Moravia (eastern part of Czechia).



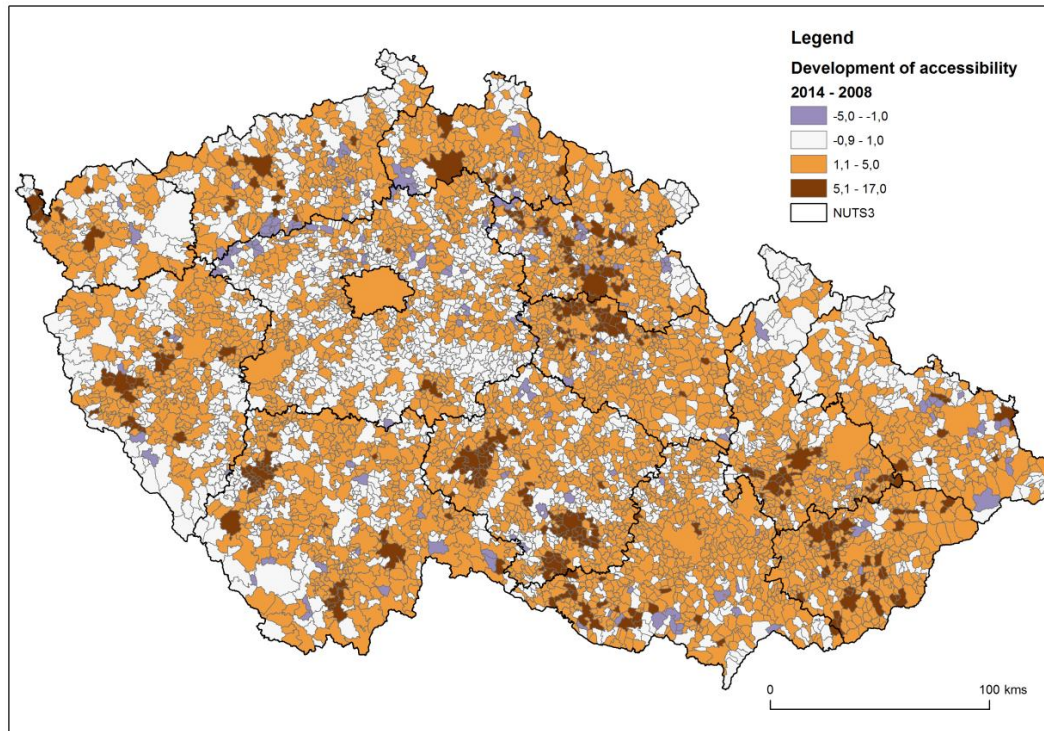
**Figure 2** Cluster map of public transport potential

Two maps below describe the development of public transport accessibility on the municipality level represented by rate of accessible municipalities. The time series covers development since 2008 to 2014 and was divided into two periods – 2008–2011 and 2011–2014. While during the first period there is more significant increase of RA with mean value +2.27%, with maximal increase above 10%, with maximal decrease higher than 6%, in the second period the development is more stagnant with small decrease (mean value -0.54).



**Figure 3** Change of RA (2008 – 2011)

Spatial trends are evident also in case of development. In the first period, three main areas are typical due to higher increases and two by significant decreases. Some of these drops can be explained by changes in financial subsidies in public transport by regional authorities (even they officially announced something different). Situation in the second period is partly reacting on the previous development. Thus while the RA has increased in some of these areas, the RA has decreased in areas with increases for 2008–2011 period. Finally some regions exist that experienced an increase for both periods.



**Figure 4** Change of RA (2011 – 2014)

#### 4. Acknowledgements

The research is supported by the Czech Science Foundation, project Spatial simulation modelling of accessibility, No. 14-26831S.

#### 5. Biography

Igor Ivan – He defended his PhD thesis focused on public transport accessibility and door-to-door accessibility in 2010. His main research activities deal with spatial analysis at micro-geographical scale, analysis of individual data, and transport analysis related mainly to public transport.

#### References

- Horák J, Ivan I, Fojtík D and Burian J (2014). Large scale monitoring of public transport accessibility in the Czech Republic. In *Proceedings of ICC 2014*, Velké Karlovice, 28-30.5.2014, 1-7.
- Ivan I, Horák J, Fojtík D and Inspektor T (2013). Evaluation of Public Transport Accessibility at Municipality Level in the Czech Republic. *13th International Multidisciplinary Scientific GeoConference SGEM 2013*, vol. 1, 1088 p.



# SAFEVolcano: Spatial Information Framework for Volcanic Eruption Evacuation Site Selection-allocation

Jumadi<sup>1</sup>, Steve Carver<sup>2</sup>, and Duncan Quincey<sup>3</sup>  
School of Geography, University of Leeds

April 15, 2015

## Summary

Volcanic disasters are commonly difficult to predict accurately in terms of when the events come, how big the magnitude, where the spatial extent of the impact, and who will be exposed. In the worst condition, people at risk confuse where they should evacuate themselves, although they already know that they are in danger. Similarly, stakeholder who responsible for evacuating people may have difficulties to manage evacuation site during critical times. To solve this problem, we propose SAFEVolcano, a GIS-based framework for managing evacuation camp selection-allocation considering the dynamic of the volcanic disaster extent. As an implementation example of the framework, we developed and demonstrated ArcGIS Python Plugin, which is available at <http://goo.gl/zdTRxG>.

**KEYWORDS:** GIS; emergency response; volcanic eruption; evacuation management, evacuation shelter selection-allocation.

## 1. Introduction

Volcanic eruption occurrences are commonly difficult to predict accurately in term of when the events come, how big the magnitude, where is the spatial extent of the impact, and who will be exposed. Such situation occurred at Merapi in 2010 confused people during the evacuation process (Jenkins et al., 2013; Mei et al., 2013; Suroño et al., 2012), caused the death of about 2,000 people at El Chichón Volcano Mexico in 1982 (Tilling, 2009), and caused the misunderstanding between the authorities and the population during an emergency situation at Kelut Volcano in 2007 (De Bélizal et al., 2012). Effective management in evacuating people at risk is one of the successful keys to deal with those situation (Mei et al., 2013). In fact, people in the active volcano commonly have been prepared evacuation planning whenever the disaster come. Similarly at Merapi in 2010, but the plan was unworkable due to the failure of the prediction (Mei et al., 2013). Managing the unpredicted situation need adaptable spatial information to various situation (Marrero et al., 2013). The excellences of spatial data to support disaster management are highlighted by some works (Cole et al., 2005; Cutter, 2003; Donohue, 2002; Laituri and Kodrich, 2008; Leonard et al., 2008; Marrero et al., 2013; Mehta et al., 2013; Mei et al., 2013; Mei and Lavigne, 2013; Rivera et al., 2010; Tsai and Yau, 2013).

Particular works on evacuation simulation using spatial data has been presented for several purposes namely destroyed building evacuation (Lo et al., 2006; Zheng et al., 2009), volcanic eruption evacuation transport routing (Marrero et al., 2013, 2010), evacuation shelter site selection purposes (Chu and Su, 2012; Kar and Hodgson, 2008; Kilci, 2012; Liu et al., 2011). They present useful method for building evacuation shelter inventory database using some physical and social parameters. Meanwhile, evacuation management also aims to allocate people to safe site effectively. The procedure should determine the capacity of each refuge and define who will be in the shelter. Geographically, grouping people from the same place is important, because moving people in this case also means moving social capital that is essential for disaster resilience (Dynes, 2006; Tobin et al., 2007). In this operation, estimating the number of fatalities like Marrero et al. (2012) and distributing refugee into each shelter like Kongsomsaksakul et al. (2005) are needed. The contingency plan in Rivera et al. (2010) provides more integrated flow in making decision to select and allocate evacuation shelter during a volcanic crisis in Ubinas Volcano Peru. They categorized the hazard zone

---

<sup>1</sup> gyji@leeds.ac.uk

<sup>2</sup> s.j.carver@leeds.ac.uk

<sup>3</sup> d.j.quincey@leeds.ac.uk

as small, medium, high explosion that convince the flexibility of the plan. Several possibilities for people at risk allocation also has been developed. However, that is not a GIS-oriented procedure so that further study is needed for generic GIS application.

Based on the explanation, there are several related studies exist on hazard evacuation site suitability selection (Kar and Hodgson, 2008), earthquake evacuation site suitability selection (Chu and Su, 2012; Liu et al., 2011), shelter capacity distribution (Kongsomsaksakul et al., 2005), and tools evacuation time simulation(Marrero et al., 2010). However, the problems in evacuation shelter selection-allocation in integrated way has not been addressed. Therefore, we propose SAFEVolcano i.e., a systematic GIS framework to estimate people at risk, select safe evacuation camp away from harmful area and allocate people at risk for each location based on occurring volcanic hazard extent.

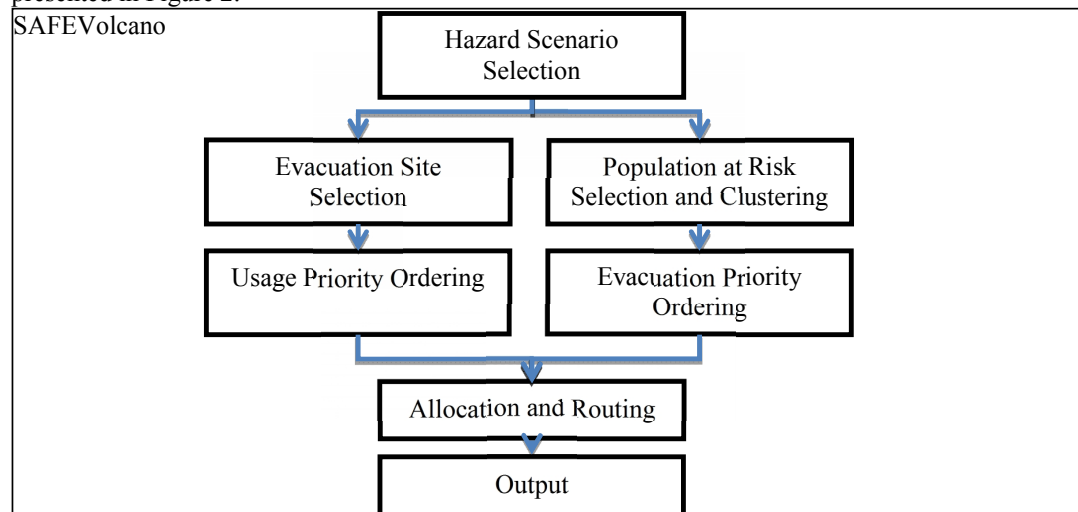
## 2. SAFEVolcano Framework

### 2.1. Input Data Requirements

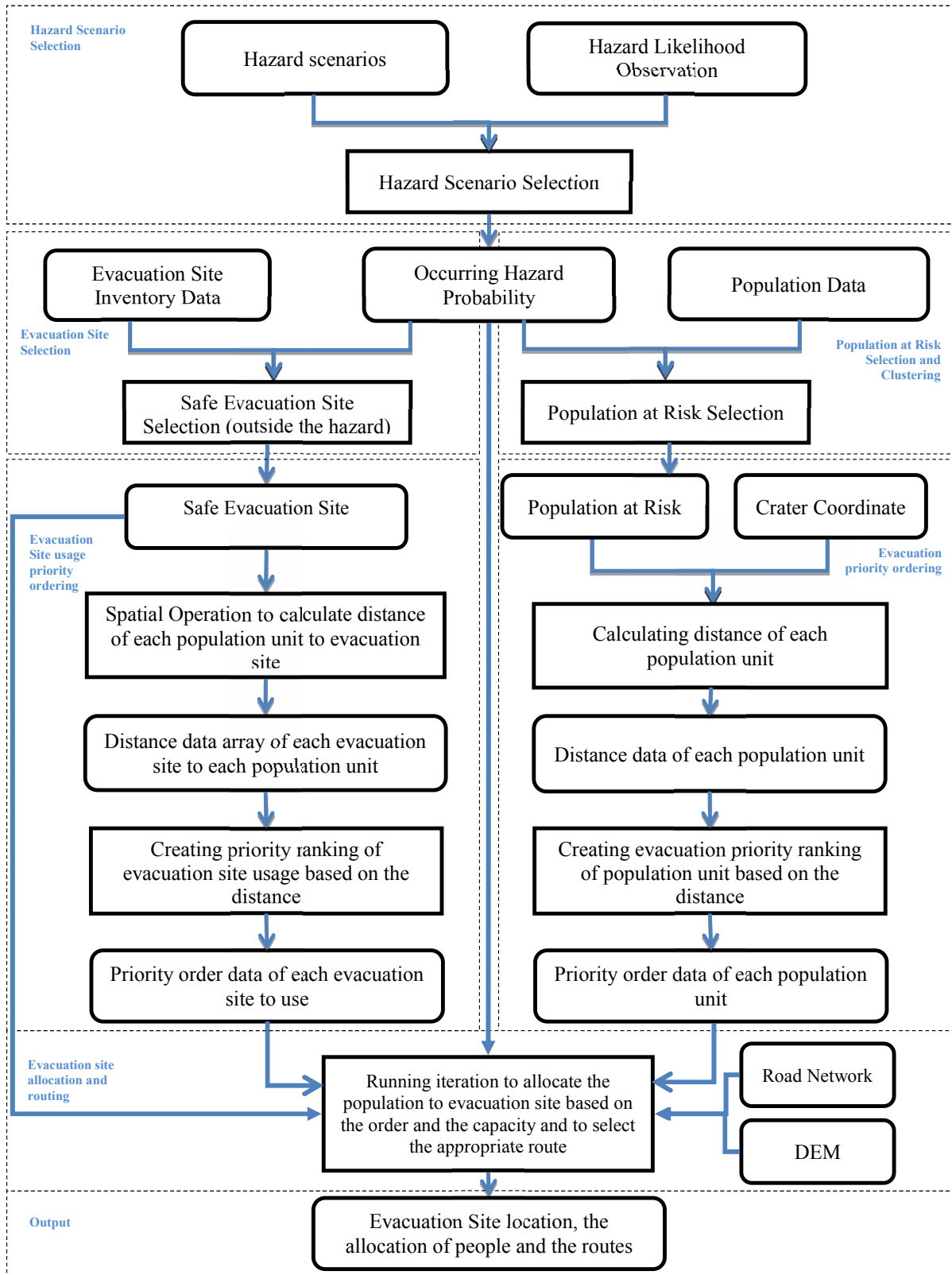
1. Center coordinate of the crater (volcanic vent) of the volcano, this data is used to calculate proximity of population to the center of hazard. The source can be from GPS or imagery.
2. Hazard scenarios, multiple scenarios of hazard are needed to anticipate the probabilities of hazard occurrences. This data can be obtained from modeling result like VORIS (Scaini et al., 2014), TITAN2D (Charbonnier and Gertisser, 2009), LAHARZ (Darmawan et al., 2014), VOLCALPUFF (Barsotti et al., 2010), FLOWGO (Harris and Rowland, 2001) or Q-LavHA (Mossoux et al., 2014).
3. Population, this data can be provided from official statistical data. The data can be based on administrative boundaries or postal address for better aggregation.
4. Road networks which can be provided by GPS survey (Marrero et al., 2010; Mei et al., 2013), scanning traditional map (Marrero et al., 2010) or image interpretation.
5. Evacuation site and its capacity, this data can be provided from image interpretation, public facilities like schools, churches/mosques are commonly used as temporary evacuation camp during crisis. Suitability assessment can be applied (Kar and Hodgson, 2008) for better result in the evacuation camp inventory data building.
6. Hazard characteristics data based on certain pre-eruption occurrence that are about seismic data (Mei et al., 2013) and direction of the vent (Scaini et al., 2014).
7. Digital Elevation Model (DEM), it is used to get slope information in the evacuation route selection process.

### 2.2. GIS Operation Framework

Generally the SAFEVolcano framework is described in the Figure1 and the detail of the operation is presented in Figure 2.



**Figure 1. General Framework**



**Figure 2.**Detailed Framework

### **2.2.1. Hazard Scenario Selection**

The operation aims to select the most applicable hazard model to the occurring/upcoming disaster event based on the likelihood. A wide range of magma compositions and eruption styles are used to develop these scenarios (Scaini et al., 2014). During disaster responses, there are several criteria can be used to observe the upcoming hazard namely seismicity and visual changing around the crater (Mei et al., 2013). Moreover, based on the observed seismic intensity, the characteristic of the upcoming eruption can be forecasted (Chouet, 1996).

### **2.2.2. Safe Evacuation Site Selection**

The purpose of this processing is to select safe evacuation sites. Flexible plan is needed to anticipate many possibilities of hazard occurrence (Marrero et al., 2013). The operation can be simply performed using disjoint topological operation in GIS to select the evacuation site (POINT) outside the occurring hazard zone (POLYGON) from evacuation site database.

### **2.2.3. Population at Risk Selection**

Exact selection of the people at risk is important to minimize evacuation cost (Whitehead, 2003). This processing aims to highlight the population units which are located on the hazard zone and estimate the number of population. The population unit commonly grouped as postcode (Robert Berry et al., 2010) or municipal area (Mei et al., 2013). The operation can be simply performed using overlapping topological operation in GIS to select the population unit (POLYGON) that overlaps with hazard zones (POLYGON).

### **2.2.4. Distance Calculation of Each Population Unit to the Crater and the Priority Ranking of Evacuation Process**

We assumed that the proximity to hazard center (crater) will be higher in the risk. Therefore, the nearest population unit needs to be evacuated firstly, followed by the next range based on the distance. To apply this assumption, shorting the population unit based on the proximity is needed. The distance is measured from the CENTROID of population unit to the given coordinate of the crater (POINT). There are many algorithms to perform this operation, for example Euclidean distance (Danielsson, 1980).

### **2.2.5. Distance Calculation of Each Population Unit to Evacuation Site Based on Road Network Routing and Priority Shorting of Evacuation Site Usage**

Similarly, finding optimum distance from evacuation unit to the evacuation site is needed to enhance transport time effectively (Marrero et al., 2010). Therefore, knowing the distance in the rank is important to short the priority of evacuation site usage. The distance is measured from the CENTROID of population unit to the given safe evacuation site (POINT).

### **2.2.6. Evacuation Site Allocation**

Finally, we design iteration procedure to distribute people at risk to safe evacuation site considering the social grouping based on their origin (population unit), the evacuation site capacity, the population unit priority to evacuate, and the evacuation site priority to use. After the allocation operation is performed, then the following routing selection operation is employed to each pair of population unit and evacuation site.

### **2.2.7. Evacuation Route Selection**

GIS operation such as least-cost path is applicable in the route selection procedure. In this selection process, some physical criteria should be considered namely slope (Yu et al., 2003), accessibility of road network (Liu et al., 2006) and the hazard occurrences (Uno and Kashiya, 2008). It is assumed that the most suitability routes for evacuation are the shortest way to get evacuation site location, with the flattest slope, widest and best road condition, and if possible away from the hazardous area. Using this assumption, we made classification of slope based on the steepness (Yu et al., 2003) and road network based on the road type classification, and the hazard based on the intensity/magnitude.

### 3. Implementation Example

Based on the framework, we use Python to develop geoprocessor plugin in ArcGIS as provided in the Figure 3. The script and the dataset example are available at <http://goo.gl/zdTRxG>. Using the dataset (Figure 4 to 9), we demonstrate the operation of this framework. The result of population at risk, the selected evacuation site, the number of evacuees allocated, and the routes directing from the population origin to the selected evacuation site is provided in the Figure 10.

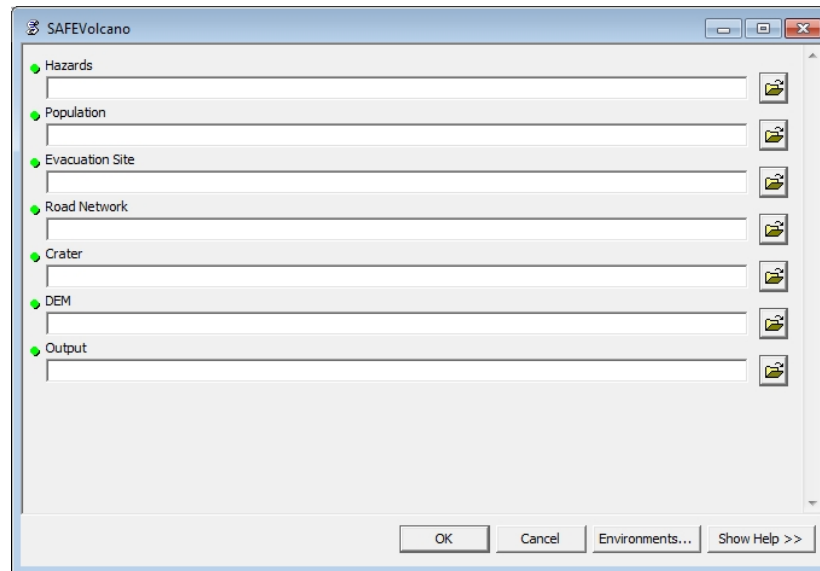


Figure 3. Example of SAFEVolcano Implementation Using ArcGIS Python Plugin

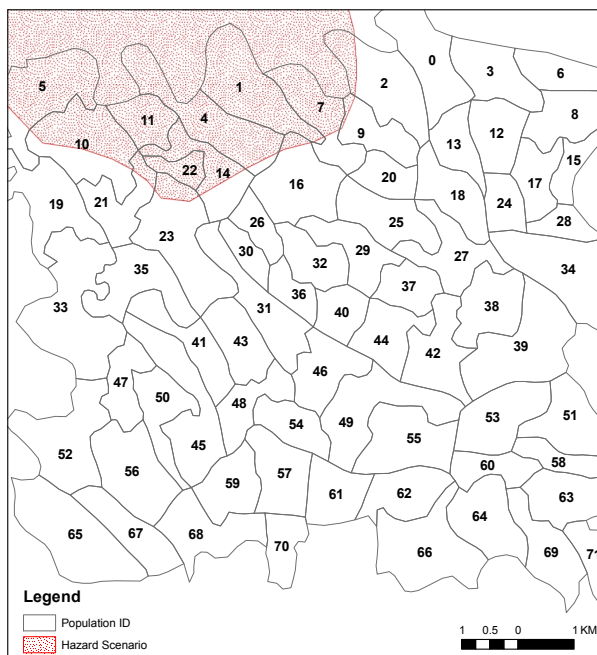


Figure 4. Hazard Scenario 1 (Showing the surrounding population unit with the ID)

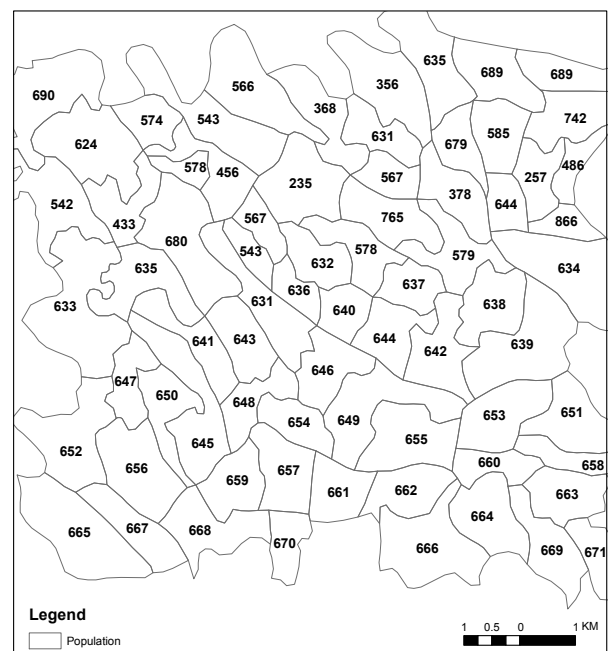


Figure 5. Population Unit (Showing the number of inhabitant)



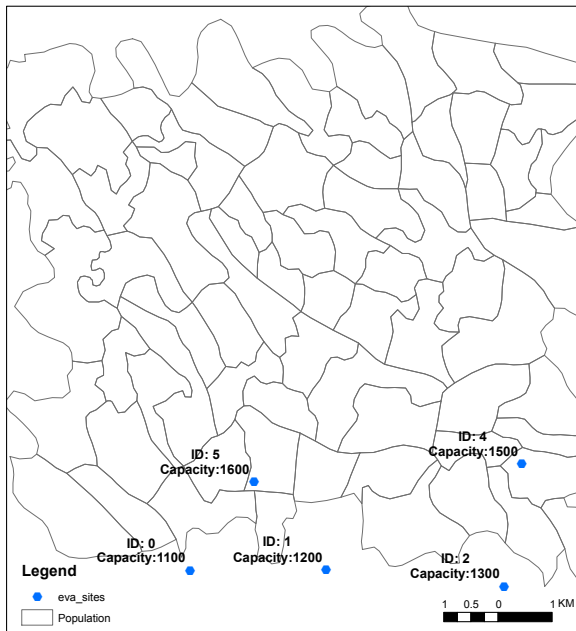


Figure 6. Evacuation Sites

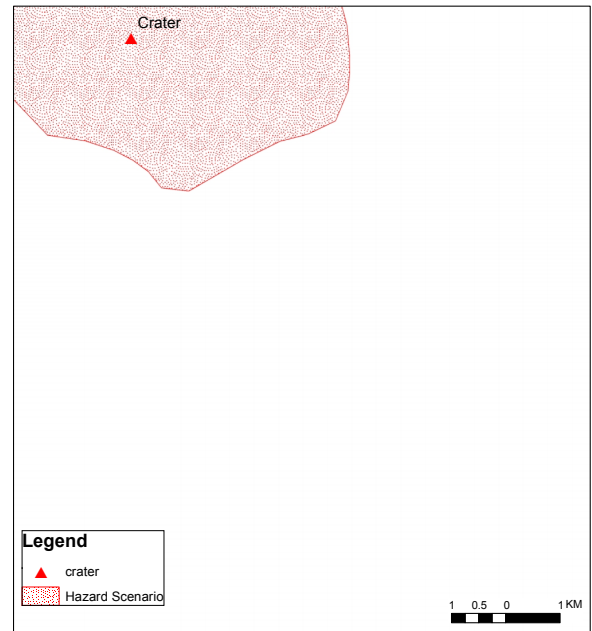


Figure 8. Crater (Volcanic Vent) Location

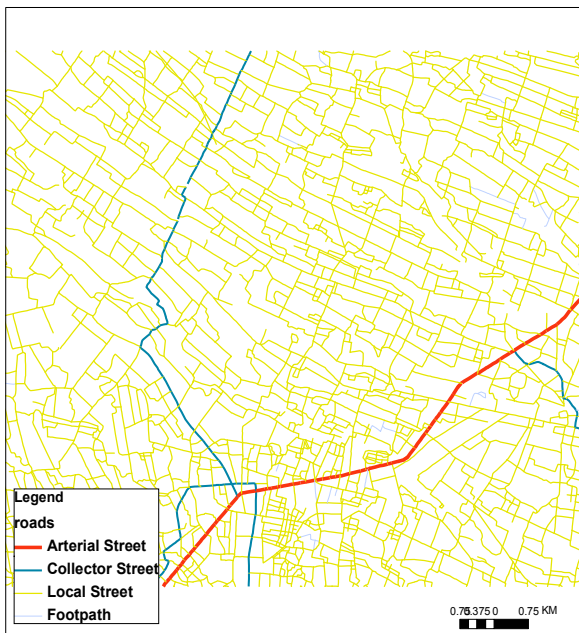


Figure 7. Roads Network

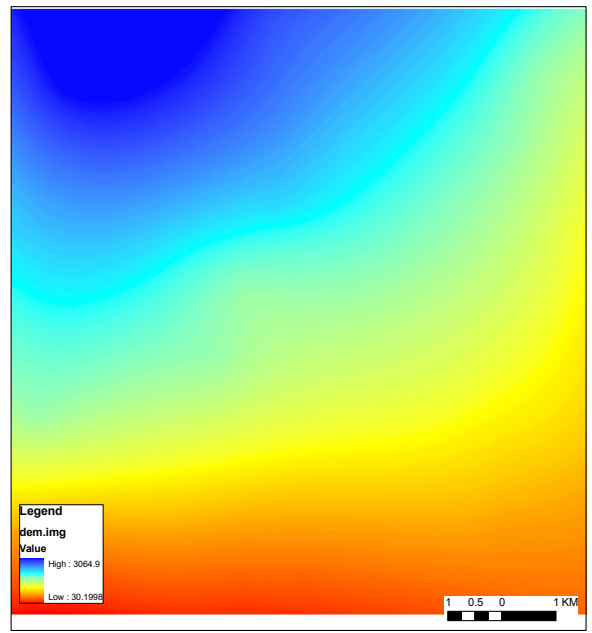


Figure 9. Digital Elevation Model

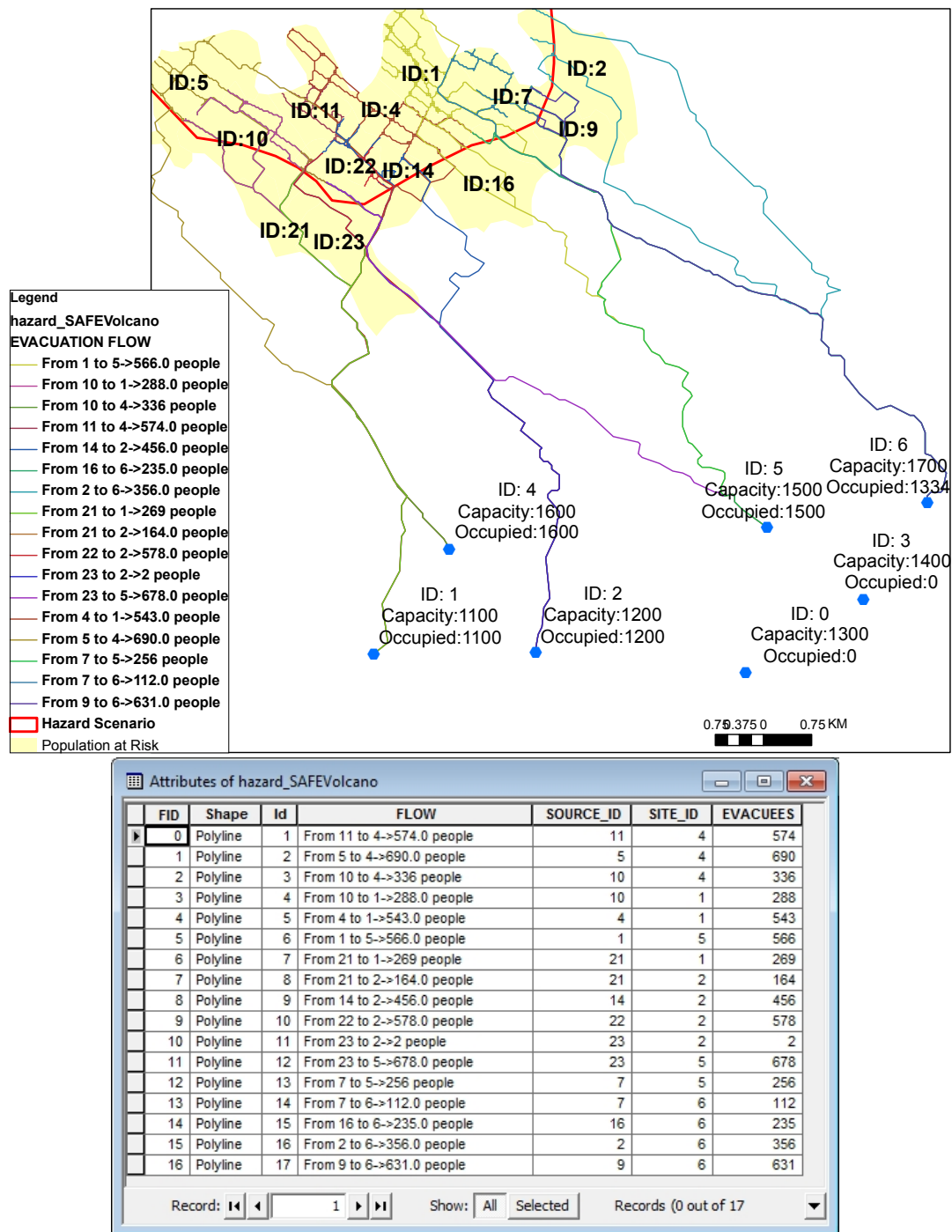


Figure 10. Evacuation Site Selection-Allocation and Routing Result

When the predicted hazard is changed, the evacuation scenario can be generated rapidly. Figure 11 provide an example of different input of the hazard scenario. In this scenario (hazard scenario 2), it is forecasted that the occurring eruption will be bigger than the previous one (hazard scenario 1). Consequently, the impacted population areas are wider, and the number of populations is bigger. As result, almost all of the evacuation sites are fully occupied to allocate people at risk.

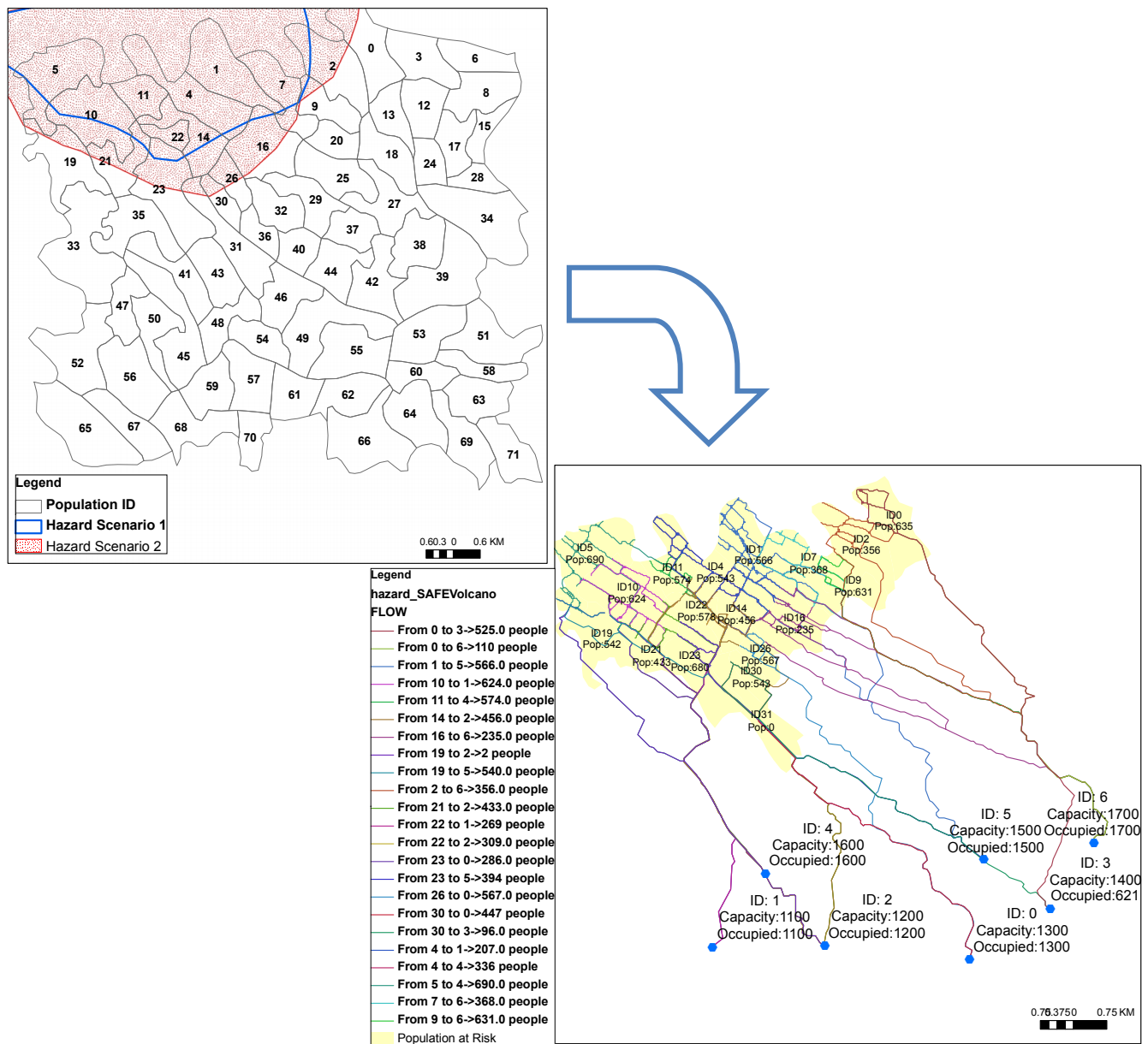


Figure 11. Evacuation Site Selection-Allocation and Routing Result with Bigger Scenario of Eruption (Hazard Scenario 2)

#### 4. Limitation

A problem emerged when we evaluated the results of population at risk estimation. Figure 12 provided an overview of this limitation which is clearly shown the different between the spatial extent of predicted hazard and the extent of population at risk. The selection of the population at risk, in this spatial operation, is based on the overlapped area between population unit and the hazardous area. Population Unit ID 23, for example, the exposed area is about half of the total area, but its entire inhabitant is calculated in the estimation of population at risk. The spatial operation is unable to simply divide the number of population with the proportion of exposed area because the residential areas are commonly not evenly distributed. Therefore, the accuracy of population at risk estimation is depend on the level of population unit detail.

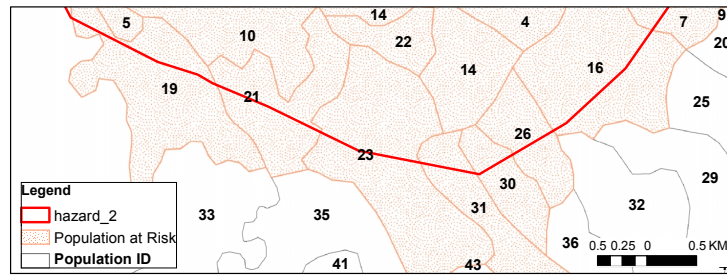


Figure 12. Spatial Processing Limitation in Population at Risk Estimation

## 5. Conclusion and Future Works

Volcanic eruptions are commonly unpredictable so causing the mislead of evacuation process. Supporting data to manage evacuation site during a critical time is needed. GIS-based framework can be used to manage evacuation camp selection-allocation as well as to select the routes considering the dynamic of the volcanic disaster extent. Future research related to this framework is needed primarily for optimize the hazard scenario retrieval, optimize population clustering, optimize the distribution of people at risk.

## 6. Acknowledgements

The authors wish to thanks Directorate General of Higher Education (DIKTI) of Indonesia and Universitas Muhammadiyah Surakarta for provide funding.

## 7. Biography

Jumadi is the first year Ph.D. student at the University of Leeds. He holds MSc in Geoinformation for Spatial Planning and Risk Management (Double Degree MSc Program between UGM, Indonesia and ITC University of Twente, The Netherlands). His research interests are GIS and Disaster Management.

Steve Carver is a Geographer and Senior Lecturer at the University of Leeds. He has over 20 years' experience in the field of GIS and multi-criteria evaluation with special interests in wildland, landscape evaluation, and public participation.

Duncan Quincey is a Lecturer in Geomorphology at the University of Leeds. His research focuses mainly on the dynamics of mountain glaciers, in particular, Hindu-Kush Himalayan glaciers and within the context of climate change.

## References

- Barsotti, S., Andronico, D., Neri, A., Del Carlo, P., Baxter, P.J., Aspinall, W.P., Hincks, T., 2010. Quantitative assessment of volcanic ash hazards for health and infrastructure at Mt. Etna (Italy) by numerical simulation. *J. Volcanol. Geotherm. Res.* 192, 85–96. doi:10.1016/j.jvolgeores.2010.02.011
- Charbonnier, S.J., Gertisser, R., 2009. Numerical simulations of block-and-ash flows using the Titan2D flow model: examples from the 2006 eruption of Merapi Volcano, Java, Indonesia. *Bull. Volcanol.* 71, 953–959. doi:10.1007/s00445-009-0299-1
- Chu, J., Su, Y., 2012. The Application of TOPSIS Method in Selecting Fixed Seismic Shelter for Evacuation in Cities. *Syst. Eng. Procedia, Information Engineering and Complexity Science - Part I 3*, 391–397. doi:10.1016/j.sepro.2011.10.061
- Chouet, B.A., 1996. Long-period volcano seismicity: its source and use in eruption forecasting. *Nature*. 03/1996, Volume 380, Issue 6572.
- Cole, J.W., Sabel, C.E., Blumenthal, E., Finnis, K., Dantas, A., Barnard, S., Johnston, D.M., 2005. GIS-based emergency and evacuation planning for volcanic hazards in New Zealand. <http://www.nzsee.org.nz>.

- Cutter, S.L., 2003. *GI Science, Disasters, and Emergency Management*. *Trans. GIS* 7, 439–446. doi:10.1111/1467-9671.00157
- Danielsson, P.-E., 1980. Euclidean distance mapping. *Comput. Graph. Image Process.* 14, 227–248. doi:10.1016/0146-664X(80)90054-4
- Darmawan, H., Wibowo, T., Suryanto, W., Setiawan, M., 2014. Modeling of pyroclastic flows to predict pyroclastic hazard zone in Merapi volcano after 2010 eruption, in: *EGU General Assembly Conference Abstracts*. Presented at the EGU General Assembly Conference Abstracts, p. 1685.
- De Bélizal, É., Lavigne, F., Gaillard, J.C., Grancher, D., Pratomo, I., Komorowski, J.-C., 2012. The 2007 eruption of Kelut volcano (East Java, Indonesia): Phenomenology, crisis management and social response. *Geomorphology, Volcano Geomorphology: landforms, processes and hazards* 136, 165–175. doi:10.1016/j.geomorph.2011.06.015
- Donohue, K., 2002. Using GIS for all-hazard emergency management.
- Dynes, R.R., 2006. *Social Capital Dealing with Community Emergencies*.
- Harris, A.J., Rowland, S., 2001. FLOWGO: a kinematic thermo-rheological model for lava flowing in a channel. *Bull. Volcanol.* 63, 20–44.
- Jenkins, S., Komorowski, J.-C., Baxter, P.J., Spence, R., Picquout, A., Lavigne, F., Surono, 2013. The Merapi 2010 eruption: An interdisciplinary impact assessment methodology for studying pyroclastic density current dynamics. *J. Volcanol. Geotherm. Res., Merapi eruption* 261, 316–329. doi:10.1016/j.jvolgeores.2013.02.012
- Kar, B., Hodgson, M.E., 2008. A GIS-Based Model to Determine Site Suitability of Emergency Evacuation Shelters. *Trans. GIS* 12, 227–248. doi:10.1111/j.1467-9671.2008.01097.x
- Kılıcı, F., 2012. *A Decision Support System for Shelter Site Selection With GIS Integration: Case for Turkey (MSc Thesis)*. Bilkent University, Turkey.
- Kongsomsaksakul, S., Yang, C., Chen, A., 2005. Shelter Location-Allocation Model for Flood Evacuation Planning. *J. East. Asia Soc. Transp. Stud.* 6, 4237–4252. doi:10.11175/easts.6.4237
- Laituri, M., Kodrich, K., 2008. On Line Disaster Response Community: People as Sensors of High Magnitude Disasters Using Internet GIS. *Sensors* 8, 3037–3055. doi:10.3390/s8053037
- Leonard, G.S., Johnston, D.M., Paton, D., Christianson, A., Becker, J., Keys, H., 2008. Developing effective warning systems: Ongoing research at Ruapehu volcano, New Zealand. *J. Volcanol. Geotherm. Res., Volcanic risk perception and beyond* 172, 199–215. doi:10.1016/j.jvolgeores.2007.12.008
- Liu, Q., Ruan, X., Shi, P., 2011. Selection of emergency shelter sites for seismic disasters in mountainous regions: Lessons from the 2008 Wenchuan Ms 8.0 Earthquake, China. *J. Asian Earth Sci., The 2008 Wenchuan Earthquake, China and Active Tectonics of Asia* 40, 926–934. doi:10.1016/j.jseas.2010.07.014
- Liu, Y., Hatayama, M., Okada, N., 2006. Development of an adaptive evacuation route algorithm under flood disaster. *Annu. Disaster Prev. Res. Inst. Kyoto Univ.* 49, 189–195.
- Lo, S.M., Huang, H.C., Wang, P., Yuen, K.K., 2006. A game theory based exit selection model for evacuation. *Fire Saf. J.* 41, 364–369. doi:10.1016/j.firesaf.2006.02.003
- Marrero, J.M., García, A., Llinares, A., Cruz-Reyna, S.D. la, Ramos, S., Ortiz, R., 2013. Virtual tools for volcanic crisis management, and evacuation decision support: applications to El Chichón volcano (Chiapas, México). *Nat. Hazards* 68, 955–980. doi:10.1007/s11069-013-0672-4
- Marrero, J.M., García, A., Llinares, A., Rodríguez-Losada, J.A., Ortiz, R., 2010. The Variable Scale Evacuation Model (VSEM): a new tool for simulating massive evacuation processes during volcanic crises. *Nat Hazards Earth Syst Sci* 10, 747–760. doi:10.5194/nhess-10-747-2010
- Marrero, J.M., García, A., Llinares, A., Rodríguez-Losada, J.A., Ortiz, R., 2012. A direct approach to estimating the number of potential fatalities from an eruption: Application to the Central Volcanic Complex of Tenerife Island. *J. Volcanol. Geotherm. Res.* 219–220, 33–40. doi:10.1016/j.jvolgeores.2012.01.008
- Mehta, P., Müller, S., Voisard, A., 2013. MoveSafe: A Framework for Transportation Mode-based Targeted Alerting in Disaster Response, in: *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, GEOCROWD '13*. ACM, New York, NY, USA, pp. 15–22. doi:10.1145/2534732.2534735

- Mei, E.T.W., Lavigne, F., 2013. Mass evacuation of the 2010 Merapi eruption. *Int. J. Emerg. Manag.* 9, 298–311. doi:10.1504/IJEM.2013.059871
- Mei, E.T.W., Lavigne, F., Picquout, A., de Bélizal, E., Brunstein, D., Grancher, D., Sartohadi, J., Cholik, N., Vidal, C., 2013. Lessons learned from the 2010 evacuations at Merapi volcano. *J. Volcanol. Geotherm. Res.*, Merapi eruption 261, 348–365. doi:10.1016/j.jvolgeores.2013.03.010
- Mossoux, S., Feltz, A., Poppe, S., Canters, F., Kervyn, M., 2014. Calibration of Q-LavHA a Quantum GIS plugin for lava flow simulation. Presented at the Cities on Volcanoes 8, Yogyakarta.
- Rivera, M., Thouret, J.-C., Mariño, J., Berolatti, R., Fuentes, J., 2010. Characteristics and management of the 2006–2008 volcanic crisis at the Ubinas volcano (Peru). *J. Volcanol. Geotherm. Res.* 198, 19–34. doi:10.1016/j.jvolgeores.2010.07.020
- Robert Berry, Richard Fry, Gary Higgs, Scott Orford, 2010. Building a geo-portal for enhancing collaborative socio-economic research in Wales using open-source technology. *J. Appl. Res. High. Educ.* 2, 78–92. doi:10.1108/17581184201000007
- Scaini, C., Felpeto, A., Martí, J., Carniel, R., 2014. A GIS-based methodology for the estimation of potential volcanic damage and its application to Tenerife Island, Spain. *J. Volcanol. Geotherm. Res.* 278–279, 40–58. doi:10.1016/j.jvolgeores.2014.04.005
- Surono, Jousset, P., Pallister, J., Boichu, M., Buongiorno, M.F., Budisantoso, A., Costa, F., Andreastuti, S., Prata, F., Schneider, D., Clarisse, L., Humaida, H., Sumarti, S., Bignami, C., Griswold, J., Carn, S., Oppenheimer, C., Lavigne, F., 2012. The 2010 explosive eruption of Java's Merapi volcano—A “100-year” event. *J. Volcanol. Geotherm. Res.* 241–242, 121–135. doi:10.1016/j.jvolgeores.2012.06.018
- Tilling, R., 2009. El Chichón's “surprise” eruption in 1982: Lessons for reducing volcano risk. *Geofísica Int.* 48, 3–19.
- Tobin, G.A., Whiteford, L.M., Jones, E.C., Murphy, A.D., 2007. Chronic Hazard: Weighing Risk against the Effects of Emergency Evacuation from Popocatépetl, México, in: *Proceedings of the Applied Geography Conference*. Presented at the Applied Geography Conference.
- Tsai, M.-K., Yau, N.-J., 2013. Improving information access for emergency response in disasters. *Nat. Hazards* 66, 343–354. doi:10.1007/s11069-012-0485-x
- Uno, K., Kashiya, K., 2008. Development of simulation system for the disaster evacuation based on multi-agent model using GIS. *Tsinghua Sci. Technol.* 13, 348–353.
- Whitehead, J.C., 2003. One million dollars per mile? The opportunity costs of Hurricane evacuation. *Ocean Coast. Manag.* 46, 1069–1083. doi:10.1016/j.ocecoaman.2003.11.001
- Yu, C., LEE, J., MUNRO-STASIUK, M.J., 2003. Research Article: Extensions to least-cost path algorithms for roadway planning. *Int. J. Geogr. Inf. Sci.* 17, 361–376. doi:10.1080/1365881031000072645
- Zheng, X., Zhong, T., Liu, M., 2009. Modeling crowd evacuation of a building based on seven methodological approaches. *Build. Environ.* 44, 437–445. doi:10.1016/j.buildenv.2008.04.002

# Geodemographics and spatial microsimulation: using survey data to infer health milieu geographies

Jens Kandt<sup>\*1</sup>

<sup>1</sup>Department of Geography, University College London

9 January 2015

## Summary

This paper presents an approach to infer lifestyle geographies from survey microdata as building block of purpose-built geodemographics. 33,000 England and Wales residents have been clustered into nine lifestyle milieus based on a range of behavioural and attitudinal variables. The milieus strongly differ by individual social and demographic circumstances. Spatial microsimulation can be used to estimate probabilistically the geographical distribution of milieus. Preliminary results for London are presented in this abstract, demonstrating how extensive behavioural information of social surveys can be combined with the nearly complete coverage of spatial census data within a geodemographics framework to inform policy interventions.

**KEYWORDS:** geodemographics, spatial microsimulation, health behaviours, urban lifestyles, milieus

## 1 Introduction

So-called health behaviours and their spatial manifestation have long been of interest to social epidemiologists and health geographers, because while they appear to significantly affect population health, they are potentially modifiable [Blaxter (1990, 2010)]. But as much of conventional health geography and social epidemiology focusses on the role of objective measures of social similarity in shaping health, theoretical social science suggests that individual subjective orientations and experiences play an important part in shaping health behaviours [Veenstra and Burnett (2014); Baum and Fisher (2014); Williams (1995), calling for a stronger focus on lifestyles in health research.

Geodemographics - the arts of classifying local areas by the characteristics of their residents - has long been discussed as tool to infer lifestyle milieus at the ecological level and in so doing inform strategic public health interventions [Abbas et al. (2009); Openshaw and Blake (1995)]. But traditionally, geodemographic classifications have been rather generic, little conceptually targeted

---

<sup>\*</sup>j.kandt.12@ucl.ac.uk



and strongly rely on objective population characteristics [Longley (2005); Singleton and Longley (2009); Voas and Williamson (2001)].

This paper summarises work that is underway to incorporate subjective orientations and lifestyle aspects into health geodemographics. The work involves the integration of extensive, individual-level social survey data with nearly complete-coverage census neighbourhood statistics through a combination of sample segmentation and spatial microsimulation. The output will be discussed with reference to uncertainties arising in this undertaking and scope and limits of informing public health and social policy interventions.

## 2 Data and methods

Lifestyle research originates with Bourdieu's work on social practice, in particular his detailed investigations of French middle class taste and cultural consumption [Bourdieu (1984, 1977, 1990)]. It is beyond the scope of this paper to discuss his theory in detail; but two conclusions are of particular relevance here. First, health behaviours do not occur in isolation but are socially situated in an often unconsciously adopted array of social practices. Second, social practices result from interactions between subjective orientations and individual (social) circumstances. These subjective orientations are expressed in actions such as leisure activities, cultural consumption, taste, social, civic and political participation, media use as well as stated values and beliefs. Health behaviours occur within these lifestyle dimensions, and it has been argued that addressing health behaviours requires an understanding of the subjective, behavioural context [Williams (1995); Veenstra and Burnett (2014); Nettleton and Green (2014)].

The UK Understanding Society longitudinal survey collects this information on a sample of more than 40,000 individuals. Waves 2 and 3, collected between 2010 and 2012, are available and provide a range of relevant information including some health behaviours themselves [Knies (2014)]. 56 relevant questions have been identified and have been combined to 31 scales after Principal Component Analysis [see Appendix for a list variables]. 33,000 respondents living in England and Wales have been clustered based on these variables through a two stage clustering procedure involving Ward's hierarchical and k means clustering. The resulting milieus were investigated with respect to behavioural patterns as well as their socio-demographic and economic profiles by means of  $\chi^2$ -based tests and one-way ANOVAs.

These milieus are then geographically projected into small areas using deterministic spatial microsimulation, specifically Iterative Proportional Fitting (IPF) [more on these methods, see Harland et al. (2012); Lovelace and Ballas (2013)]. In short, IPF weighs a survey respondent's representativeness of a given spatial zone (e.g. ward) based on matching socio-demographic variables which are available for the respondent and for each zone in form of aggregate statistics. Subsequently, weighted statistics can be created for any outcome of interest that is included in the survey. In this case, IPF is being used to estimate the spatial distribution of lifestyle milieus and their associated health behaviours. This stage of the work is on-going, and some preliminary output is included in



Table 1: Summary of health milieus 1 to 4. All aspects reflect statistically significant cluster differences.

#	cluster label and frequency	key characteristics	socio-demographic profile
1	enduring isolation (11%)	unhealthy behaviours (smoking, low physical activity, diet), low leisure participation, low social and political participation, low news consumption	middle-aged, low income, low qualification, single or in couple with one child, in public housing
2	unconcerned starters (8%)	lower physical activity, unhealthy diet, low local attachment, low social and political participation, low news consumption, higher internet use	younger (16-34), low income, basic qualifications, early career, majority single, in private accommodation, urban, London
3	retiring generation (12%)	low physical activity, lower levels of smoking, mixed diet, low leisure participation, lower social integration, basic political participation, very low news consumption, no internet use, high TV consumption	majority over pensionable age, low income, low qualifications, married or widowed, in owned home, often providing care for other person
4	locally anchored (10%)	average health behaviours, lower levels of smoking, average leisure participation, very high local attachment, higher social integration, average political participation, average lower news consumption	often women, approaching pensionable age, lower-medium income, basic qualifications (majority GCSE or below), couples often with children, in owned home/on mortgage

this extended abstract for illustration.<sup>1</sup>

### 3 The health milieus

Nine distinct milieus were identified and investigated with respect to socio-demographic and economic characteristics as well as measure of self-rated health. Tables 1 and 2 summarise key behavioural characteristics and the socio-demographic profiles of each milieu. All reported characteristics refer to statistically significant differences between clusters. The labels are provisional and still subject to refinements.

The size of the milieus ranges between 8 and 15 per cent. Three clusters, currently called *enduring isolation*, *unconcerned starters* and *retiring generation* represent three groups whose health behaviours would be considered unhealthy, in particular with respect to physical activity and diet. Although these clusters are of similarly low social status and incomes, they provide different demographic and behaviour contexts which are suggestive of different types of social pathways at work with differential impacts and implications for public health responses. The three wealthier groups *involved cultural consumers*, *rising extroverts* and *committed citizens* show minor differences

<sup>1</sup>The statistical software used has been R ?.

Table 2: Summary of health milieus 1 to 4. All aspects reflect statistically significant cluster differences.

#	label	key characteristics	socio-demographic profile
5	established cultural consumers (13%)	high level of sports, healthy diet, lower levels of smoking, very high leisure participation, above-average local attachment, higher political participation, high news consumption, higher internet use	middle-aged, very high income, high qualification, in advanced careers, families with children, in owned home, in London and South East
6	rising extroverts (10%)	high levels of sports, lower levels of smoking, high leisure participation, very low local attachment, higher political participation, high news consumption, high internet use	younger, very high income, high qualification, young couples sometimes with children, in transition to home ownership, live in London and South, urban
7	committed citizens (8%)	medium to higher levels of sports, healthy diet, lower levels of smoking, higher leisure participation (arts and sights), higher local attachment, very high civic participation (organisation member and volunteering), higher political participation, high news consumption, higher internet use	approaching retirement, very high income, high qualification, married with children, in own home, almost half live in London and South, often providing care for other person
8	laid-back detachment (13%)	very low levels of physical activity, mixed diet, lower levels of smoking, lower leisure participation, moderate local attachment, higher political participation, lower news consumption, higher internet use	younger to middle aged, lower-medium income, basic qualification (majority GCSE or lower), mixed ethnic background (10% Asian), families with children, in owned property/on mortgage
9	digital age autonomy (15%)	higher levels of sports, mixed diet, lower levels of smoking, specifically arts-related leisure activities, low local attachment, low civic and political participation, low news consumption, high internet and social media use	younger, lower-medium income, basic qualification (majority GCSE or lower), single or young couples often with children, on mortgage

in overall healthy behaviours. Yet they, too, reveal different behavioural tendencies with respect to leisure activities, local orientations, social, civic and political participation. Finally, another set of three clusters, *locally anchored*, *laid-back detachment* and *digital age autonomy*, with medium levels of incomes and basic qualifications differ with respect to levels of exercising and a range of leisure and social orientations.

Figure 1 shows how the clusters distribute across chronic disease risk and mean income. While each income level is broadly associated with a particular range of disease risks, the distribution of milieus suggests further differentiation of health-related pathways. This becomes particularly clear

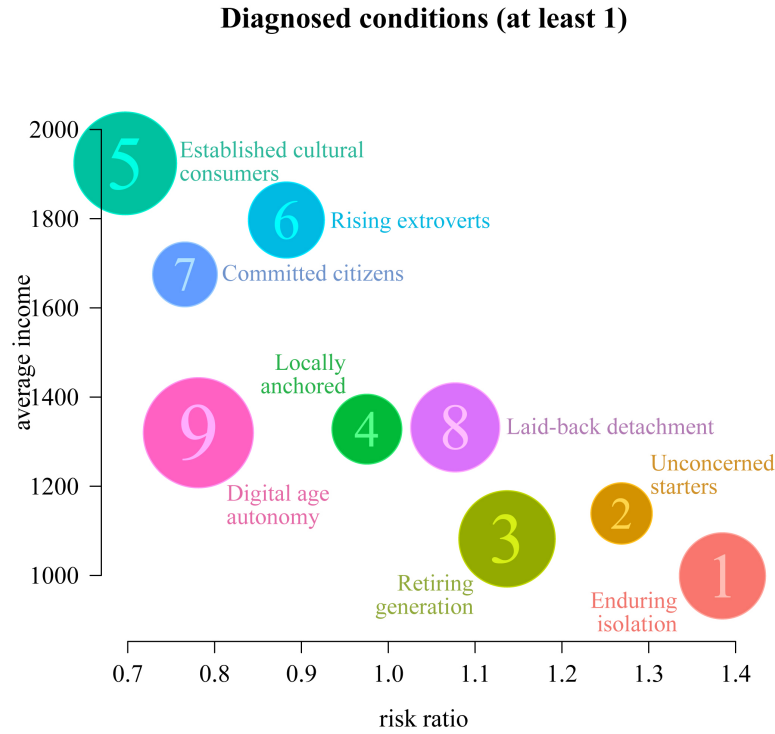


Figure 1: Milieu differentiation by income and age-and sex-standardised chronic disease risk. The size of the circles reflects the size of the clusters.

when viewing milieus 4, 8 and 9: their economic and demographic profiles are similar; yet their difference in disease risk is significant. Vice versa, larger income differences between clusters 5, 6 and 7 do not translate into corresponding proportional risk differences. In fact, contrary to our expectations, cluster 6 is worse of than clusters 7 and 9. Conventional social epidemiological studies typically assert a social gradient across status groups in health; but this uni-dimensional view of social status is likely to mask milieu-specific pathways that do not strictly reproduce the social gradient in health.

#### 4 Adding the spatial perspective: probable prevalence of health milieus (preliminary results)

The strong socio-demographic distinctiveness of milieus offers particular opportunities to project the milieus geographically. Figure 2 shows the first experimental run of spatial microsimulation for 2011 wards in London, matching UK 2011 census neighbourhood statistics with respondent characteristics sex, age and socio-economic status (NSSEC-5). Even this very limited range of matching variables on this coarse geographical scale produces distinct spatial distributions for each

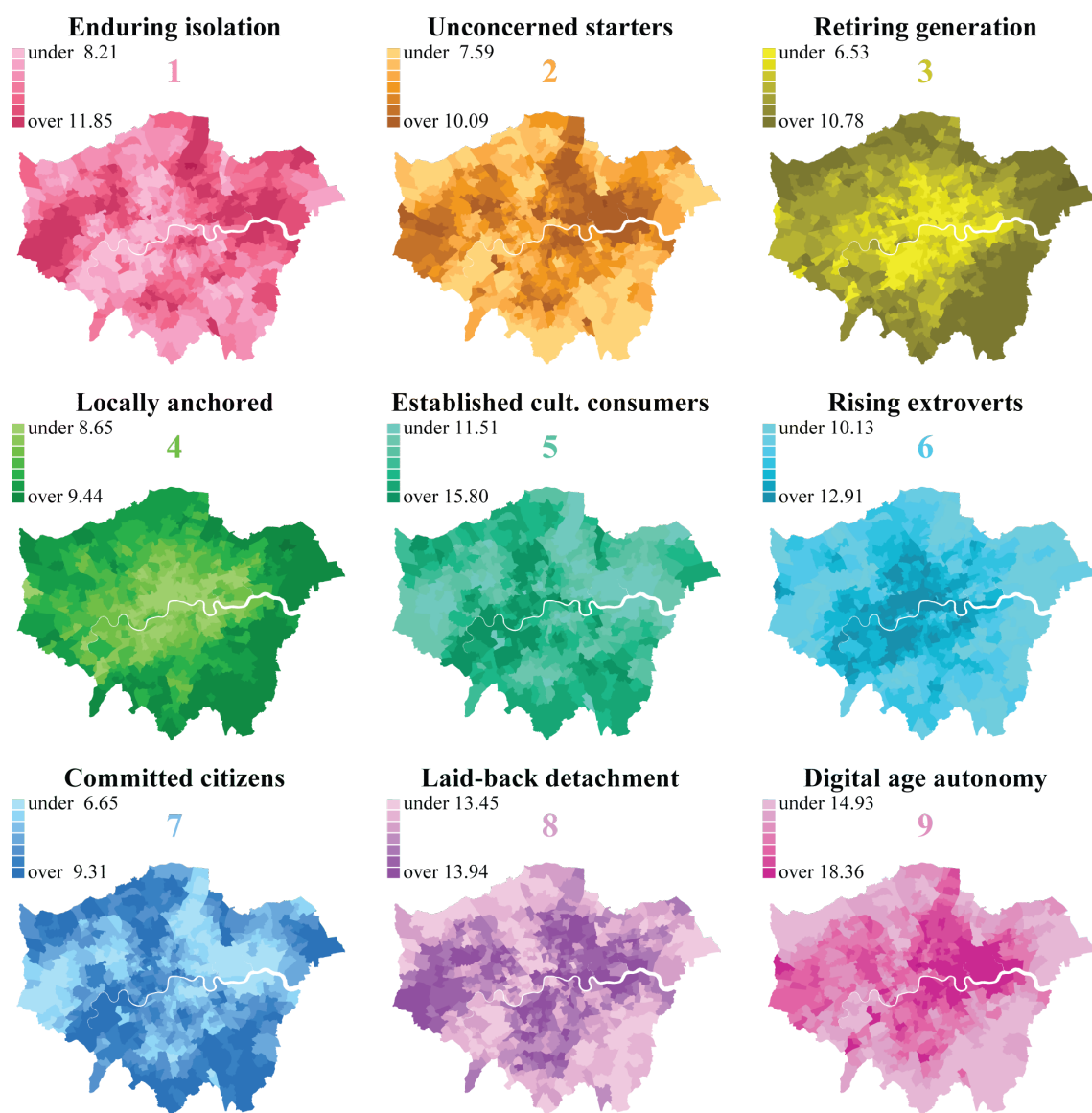


Figure 2: Milieu probabilities in London wards (2011) derived from deterministic spatial microsimulation

milieu. The maps show different relative frequencies of milieus and, given the IPF-derived weighting, can be interpreted as probabilities of the prevalence of the respective health milieu. Overall, the emerging milieu-specific geographies seem plausible in the London context. Nevertheless, it should

be remembered that this is an illustration of on-going work rather than a definitive result. It remains to be seen how the geographies change as both the spatial resolution and the number of matching variables are increased.

## **5 Conclusions**

This paper presents a social theory-grounded approach within a geodemographic framework to combine the power of extensive social surveys with the wide coverage of an administrative data. The work will continue with refining the spatial microsimulation, specifically by extending the matching variables to those that also prove milieu-discriminant and by increasing the spatial granularity to better reflect local variations of population characteristics. Subsequently, some validation will be carried out by comparing the resulting geographies with the detailed geocoded information that is available in Understanding Society as well as other data, notably the 2011 Output Area Classification.

In summary, the findings suggests that behavioural orientations and their co-varying health behaviours vary by multiple social and demographic characteristics; they may therefore be geographically simulated under close observation of the uncertainties associated with synthetic estimates. Thus, adding a behavioural building block to health geodemographics may be a promising way forward in making the tool more relevant for strategic public health and social policy interventions.

## **6 Acknowledgements**

This work is part of an ongoing, ESRC-funded PhD, supervised by Prof. Paul Longley and Prof. Jenny Robinson, Department of Geography, University College London, and is benefitting from their comments and feedback. I would like to extend further thanks to James Cheshire and Nicola Shelton for their reviews.

## **7 Biography**

Jens Kandt is a PhD candidate at the Department of Geography, University College London, and a researcher at LSE Cities, London School of Economics and Political Science. His work focusses on linking spatial statistics and social theory to understand dynamics of urban environments and their implications for transport, mobility and people's health. He holds an engineering degree in planning from the German University of Dortmund and has research and work experience in the UK, India, Germany, Ghana and Hong Kong.

## References

- Abbas, J., Ojo, a., and Orange, S. (2009). Geodemographics—a tool for health intelligence? *Public health*, 123(1):e35–9.
- Baum, F. and Fisher, M. (2014). Why behavioural health promotion endures despite its failure to reduce health inequities. *Sociology of health & illness*, 36(2):213–25.
- Blaxter, M. (1990). *Health and lifestyles*. Routledge, London.
- Blaxter, M. (2010). *Health*. Cambridge: Polity.
- Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge University Press, New York.
- Bourdieu, P. (1984). *Distinction: a Social Critique of the Judgment of Taste*. Harvard University Press, Cambridge.
- Bourdieu, P. (1990). *In Other Words*. Stanford University Press, Stanford.
- Harland, K., Heppenstall, A., Smith, D., and Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, 15:1–24.
- Knies, G. (2014). *Understanding Society UK Household Longitudinal Study: Wave 1-4, 2009-2013 User Manual*. ISER Institute for Social and Economic Research, Colchester.
- Longley, P. A. (2005). Geographical Information Systems: a renaissance of geodemographics for public service delivery. *Progress in Human Geography*, 29(1):57–63.
- Lovelace, R. and Ballas, D. (2013). Truncate, replicate, sample: A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41:1–11.
- Nettleton, S. and Green, J. (2014). Thinking about changing mobility practices: how a social practice approach can help. *Sociology of health & illness*, 36(2):239–51.
- Openshaw, S. and Blake, M. (1995). Geodemographic segmentation systems for screening health data. *Journal of epidemiology and community health*, 49 Suppl 2(Suppl 2):S34–8.
- Singleton, A. D. and Longley, P. A. (2009). Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29(3):289–298.
- Veenstra, G. and Burnett, P. J. (2014). A relational approach to health practices: towards transcending the agency-structure divide. *Sociology of health & illness*, 36(2):187–98.
- Voas, D. and Williamson, P. (2001). The diversity of diversity: a critique of geodemographic classification. *Area*, 33(1):63–76.
- Williams, S. J. (1995). Theorising class, health and lifestyles: can Bourdieu help us? *Sociology of Health and Illness*, 17(5):577–604.

## Appendix - List of scales and variables

scale	variable name	description	survey module
nutrition	Usdairy	Usual type of dairy consumption	nutrition
nutrition	Usbread	Type of bread eats most frequently	nutrition
nutrition	Wkfruit	Days each week eat fruit	nutrition
nutrition	Wkvege	Days each week eat vegetables	nutrition
smoke	Ncigs	Usual no. of cigarettes smoked per day	smoking
smoke	Smcigs	Ever smoked cigarettes regularly	smoking
smoke	Smncigs	Number of cigarettes smoked in past	smoking
walk	Wlk30min	Number of days walked at least 30 minutes	physical activity
sports	Sportsfreq	Moderate intensity sports frequency	leisure, culture and sport
sports	Sports3freq	Mild intensity sports frequency	leisure, culture and sport
advice	Scopngbhc	Advice obtainable locally	neighbourhood (self-compl)
belong	Scopngbha	Belong to neighbourhood	neighbourhood (self-compl)
borrow	Scopngbhd	Can borrow things from neighbours	neighbourhood (self-compl)
dark	Crdark	Feel safe walking alone at night	local neighbourhood
friends	Scopngbhb	Local friends mean a lot	neighbourhood (self-compl)
improve	Scopngbhe	Willing to improve neighbourhood	neighbourhood (self-compl)
stay	Scopngbhf	Plan to stay in neighbourhood	neighbourhood (self-compl)
talk	Scopngbhh	Talk regularly to neighbours	neighbourhood (self-compl)
close	Closum	How many close friends	social network
family	Simfam	Proportion of friends who are also family members	social network
local	Simarea	Proportion of friends living in local area	social network
network	Simage	Proportion of friends with similar age	social network
network	Simrace	Proportion of friends of same race	social network
network	Simateduc	Proportion of friends with similar level of education	social network
network	Simjob	Proportion of friends who have a job	social network
socnet	Netcht	Hours spent interacting with friends through social websites	social network
civic	Civicduty	Sense of civic duty	political engagement
civic	Civicduty	Sense of civic duty	political engagement
polcomp	Poleff1	Qualified to participate in politics	political self-efficacy
polcomp	Poleff2	Better informed about politics	political self-efficacy
polcost	Polcost	Cost of political engagement	political engagement
polcost	Polcost	Cost of political engagement	political engagement
polcyn	Poleff3	Public officials don't care	political self-efficacy
polcyn	Poleff4	Don't have a say in what government does	political self-efficacy
polinf	Perpolinf	Perceived political influence	political engagement
polinf	Perpolinf	Perceived political influence	political engagement
polit	Vote6	Level of interest in politics	politics
polit	Vote6	Level of interest in politics	politics
voteben	Perbfts	Personal benefit in voting	political engagement
voteben	Perbfts	Personal benefit in voting	political engagement
voteint	Voteintent	Voting intention	political engagement
voteint	Voteintent	Voting intention	political engagement

scale	variable name	description	survey module
votenorm	Votenorm	Voting as a social norm	political engagement
votenorm	Votenorm	Voting as a social norm	political engagement
org	Orgm	Which organisations member of	groups and organisations
org	Orga	Active in organisations	groups and organisations
org	Orgmt	Member of organisations NSC	groups and organisations
org	Orgat	Active in organisations NSC	groups and organisations
volum	Volfreq	Frequency of volunteering	voluntary work
arts1	Arts1freq	Arts activities frequency	leisure, culture and sport
arts2	Arts2freq	Arts events frequency	leisure, culture and sport
hist	Herfreq	Historical sites frequency	leisure, culture and sport
lib	Libfreq	Library frequency	leisure, culture and sport
musm	Musfreq	Museum frequency	leisure, culture and sport
news	Newsouce	Sources of News	news and media use
tv	Tvhours	Hours of TV per weekday	news and media use



# Designing a location model for face to face and on-line retailing for the UK grocery market

Elena Kirby-Hawkins<sup>1</sup>, Graham Clarke<sup>2</sup> and Mark Birkin<sup>3</sup>

<sup>1</sup>School <sup>1</sup>of Geography, University of Leeds

April 15, 2015

## Summary

The aim of this paper is to explore the patterns of e-commerce sales in Yorkshire and Humberside area and use this analysis to the building and testing of a retail location model which can estimate local and regional sales both for physical stores and for internet sales. This is important for retailers as future geographical growth is likely to be driven by a mixture of new stores and e-commerce. This study creates an unique opportunity for the retailers to develop innovative site location techniques to estimate instore and online sales which will then help drive future regional growth models.

**KEYWORDS:** e-commerce, grocery retail, spatial interaction model

## 1.Introduction

The vast and rapid expansion of internet usage has generated widespread on line sales, making the UK one of the leading countries for e-commerce. However, little is currently known about the geography of consumers using e-commerce. The aim of this paper is two-fold. First, we use data provided by a major UK grocery retailer to explore the patterns of e-commerce sales in the Yorkshire and Humberside area of the UK. Key hypotheses to test will be the importance of location itself (rural versus urban traffic), the importance of usage by different geodemographic groups and the importance of the possible relationship between physical store presence and on-line sales.

Once we can better understand the geography of on-line sales, we can then attempt to use this analysis to inform the building and testing of a retail location model which can estimate local and regional sales both at physical stores and for internet sales. This is important for retailers as future geographical growth is likely to be driven by a mixture of new stores and e-commerce. This study creates an unique opportunity for the retailers to develop innovative site location techniques to estimate instore and online sales which will then help drive future regional growth models.

## 2.Data

The actual sales data at the output level geography has been provided by a UK leading supermarket chain. The data has been derived from the loyalty card scheme for the three months in 2013 with over 800,000 unique customers. The estimated sales data for online and face to face sales has been provided by CACI marketing agency based on the survey of 50,000 participants.

The study area is Yorkshire and Humberside region with 791 postal sectors

---

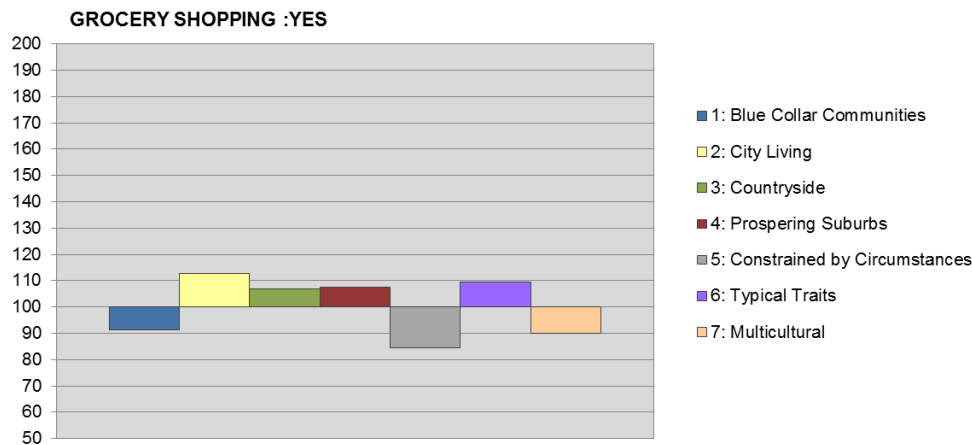
<sup>1</sup>gyekh@leeds.ac.uk

<sup>2</sup>G.P.Clarke@leeds.ac.uk

<sup>3</sup>M.H.Birkin@leeds.ac.uk

### 3.Geodemographics of e-commerce

Many scholars have indicated the variance among different demographic groups towards online shopping. These variations can be largely explained by two theories – efficiency theory and diffusion of innovation. The latter states that new technologies emerge in the cities and are initially adapted by young professional, affluent males (Rogers, 2003). On another hand, consumers living in rural locations with limited access to shops are more likely to shop online (Farg, 2006). The combination of various socio-economic characteristics in relation to online buying is represented in Figure 1.

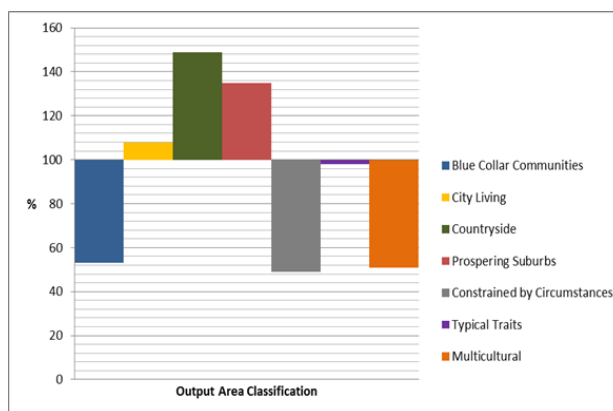


**Figure 1.** Online grocery preferences by OAC groups.

Source: The British Population Survey, 2010

There is a clear indication that Blue Collar Communities, Constrained by Circumstances and Multicultural communities are not enthusiastic online buyers. However, within these groups there are distinctive variations with some of Older Blue Collar and Afro-Caribbean communities more interested in e-commerce (with the scores above average). Despite the fact that all subgroups scored above average in City Living category, their results were not particularly high compared to Accessible Countryside subgroup, where 30% are more likely to buy grocery online than the average UK consumer. This fact supports the efficiency theory stating that the less accessible areas will be the greatest e-commerce users. The other enthusiastic online users are Prospering Suburbs and Typical Traits.

The multivariable analysis of ONS data in comparison to actual sales data from the supermarket chain is presented in Figure 2.



**Figure 2.** Online preferences among OAC Supergroups based on actual grocery sales data

The least variations towards online shopping are among customers belonging to the City Living and Typical Traits groups. However, the large gap between these two data sets within Blue Collar Communities, Constrained by Circumstances and Multicultural, can be partly explained by the fact that this particular supermarket chain is not very popular among customers belonging to these OAC groups. Overall the actual data supports the national statistical data with an index above or below the average data, which equals to 100.

#### 4. Quadrant analysis

To explore the relationship between online buying, store provision and population density the quadrant analysis technique has been applied. For the purpose of this analysis the following indicators have been generated.

C prov - grocery floorspace provision by competitors stores across the study area calculated based on the results of a Spatial Interaction Model

S prov - grocery floorspace provision by the major supermarket chain

Urban – rurality of the area or population density calculated as number of people per square kilometre in individual postal sectors

Share – e-business share in total grocery expenditure based on the loyalty scheme data provided by supermarket chain

These indicators have been ranked from 1 to 791 with the highest value starting with 1 (791 being the number of postal sectors).

The first test is on the relationship between e-business share and client's provision. For the highest client's provision in the 80 postal sectors or top 10% of the total 791 postal sectors, the average rank of "Share" is 520, which is above average (396) for all postal sectors. There is a strong suggestion that there is a substitution between physical and virtual channels taking place.

The second test is a 'quadrant analysis' of client's provision against competitors' provision. The hypothesis here would be that for given levels of client's provision (S prov) then low levels of Competitors' provision will tend to encourage higher levels of online use because there are no alternatives. The evidence seems partly to support this idea with higher than average rank share in the areas with less client's provision with e-shares of 10% and 14% and with lowest e-share of 4% in the areas with higher client's stores provision. Interestingly, in the areas with an extensive physical stores presence the e-share is very high with 14%.

**Table 1** Client's and competitors store provision in relation to e-share

	C prov < 5.2	C prov > 5.2
S prov < 0.47	Average Rank 378 (share 10%) [ N = 667 ]	Average Rank 262 (share 14%) [ N = 15 ]
S prov > 0.47	Average Rank 632 (share 4%) [ N = 61 ]	Average Rank 386 (share 14%) [ N = 48 ]

The third test is between urbanisation or population density and e-share. Based on the efficiency theory the e-grocery business would be greater in rural areas. An average e-share in the top 10% of the most urban areas is at 9% and average e-share in the top 10% of the most rural areas of 12.5% confirms the efficiency theory hypothesis.

A final question concerns the interaction between client's provision and rurality. The Table 2 below demonstrates that with the lower physical channel provision in the more rural areas the online usage uptake is the greatest of 11.2% with 427 observations compared to 8.1% e-share in the areas more urbanised and with greater client's stores presence. Interestingly, e-share is similar in the less urbanised areas with greater clients' presence and vice versa.

**Table 2** Client's store provision and population density in relation to e-share

	S Prov <0.47	S prov >0.47
Urban<0.002	Average share 11.2% [ N=427]	Average share 9.6% [N=44]
Urban>0.002	Average share 9.4% [ N=255]	Average share 8.1% [N=65]

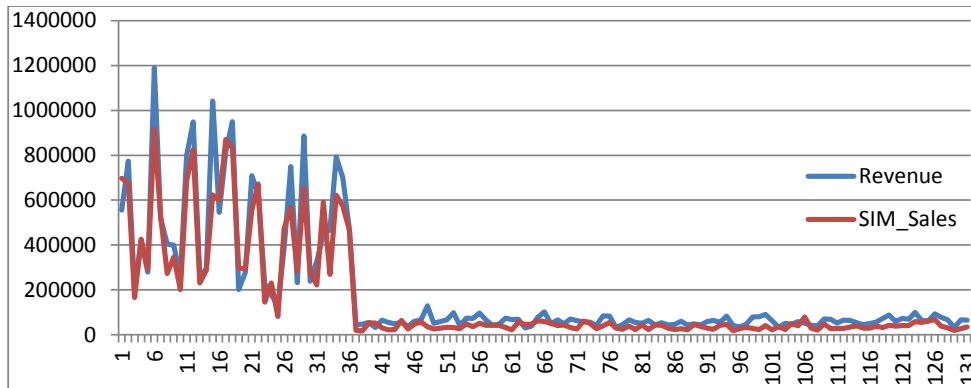
### 5.Face to face Spatial Interaction Model

To estimate physical stores sales the Spatial Interaction Model (SIM) technique has been applied which has been disaggregated by brand and customer types within ACORN classification. The basic formula for SIM is shown in Equation 1.

$$S_{ij}=A_i O_i W_j \exp^{-\beta c_{ij}} \quad (1)$$

Where  $S_{ij}$  is estimated expenditure from postal sector  $i$  and store,  $A_i$  is competition,  $O_i$  is the demand in postal sector  $i$ ,  $W_j$  is store attractiveness,  $\exp^{-\beta c_{ij}}$  is the distance decay parameter with  $\beta$  value disaggregated by rural and urban locations.

The results of SIM is represented in Figure 3 with correlation of 97% between estimated and actual sales across 131 client's stores

**Figure 3** Estimated and actual sales across clients' stores

### 6. Designing online and combined Spatial Interaction Model

The current SIM to estimate online sales provides 60% correlation with actual online sales across postal sectors. Table 3 represents parameters which has been set for online SIM

**Table 3** Online SIM parameters

Demand data	Estimated CACI data for 791 postal sectors
Supply data	Floor spaces for 7 supermarkets providing online service
Alpha value	Disaggregated by Acorn groups
Distance	1

The next stage is to calibrate the online SIM and address the challenges associated with online factor such as beta value, distance deterrence, brand attractiveness and integration of click and collect services. The final model will include face to face and online SIMs and incorporate results of geodemographic and quadrant analyses.

## **7. Acknowledgments**

This project is funded by ESRC in partnership with RIBEN

## **8. Biography**

Elena Kirby-Hawkins is a PhD student at the University of Leeds and has research interests in e-commerce, grocery retailing and spatial analysis of Business Geography

Graham Clarke is Professor at the University of Leeds. His research interests include GIS, urban services, retail and business geography, urban modeling and continuing professional education

Mark Birkin is Professor of Spatial Awareness and Policy at the University of Leeds and Director of the Consumer Data Research Centre. His interests are in simulating social and demographic change within cities and regions using techniques of microsimulation, agent-based modelling and GIS.

## **References**

Farag S (2006). *E-Shopping and its Interactions with In-Store Shopping*. PhD dissertation. Utrecht: Faculty of Geosciences, Utrecht University. Available at <http://igitur-archive.library.uu.nl/dissertations/2008-0603-200316/UUindex.html>, Accessed on 10 July 2013

Rogers E (2003). *Diffusion of Innovations*, 5th Edition. Simon and Schuster.

The British Population Survey (2010). *OAC User Group Grand Index*. Available <http://blogs.casa.ucl.ac.uk/2010/09/07/oac-grand-index/> Accessed on 15 October 2013

# Data-driven modelling of police route choice

Kira Kowalska<sup>\*1</sup>, John Shawe-Taylor<sup>†2</sup> and Paul Longley<sup>‡3</sup>

<sup>1</sup>Department of Security and Crime Science, University College London

<sup>2</sup>Department of Computer Science, University College London

<sup>3</sup>Department of Geography, University College London

November 4, 2014

## Summary

In recent years, increasing digitisation of police patrol activities has enabled new insights into police patrol behaviour. This paper explores digitised traces of police patrol journeys in order to understand police routing preferences and to propose models that enable simulations of patrol behaviour. The models assume that police journeys are undertaken as series of “topics”, which are inferred from vehicle GPS data using a widely used topic modelling technique called *latent Dirichlet allocation*. Initial experiments have shown that they are capable of reproducing police coverage patterns that would not be captured by alternative models assuming optimal behaviour.

**KEYWORDS:** route choice, topic modelling, GPS data, police vehicles

## 1. Introduction

In recent years, increasing digitisation of police patrol activities has brought new opportunities and challenges to policing. Not only has it enabled evaluation of current patrol strategies, but also a shift to more effective, data-driven approaches.

This research project analyses digitised traces of police patrol vehicles to propose data-driven models of police movements around the city. The models give explanations about police behaviour and enable simulations of police movement. Hence, they could aid the development of new patrol strategies and serve as an evaluation tool of police work on patrol.

The project is one of the first attempts to develop data-driven models of police patrol movement. This research gap arises mainly because of non-availability of data capturing detailed police movement around the city. The case study for this research will be the London Borough of Camden, for which GPS traces of police vehicles have been recently released for research purposes.

## 2. Methods

Modelling police vehicle movements is based on the assumption that police plan their journeys in stages, taking their preferred routes through neighbourhoods en route to destination. This assumption is in line with research into vehicle route choice undertaken by Manley (2013) who showed that drivers’ behaviour is rather suboptimal and that their “route selection takes place in phases, linking locations and decision points on route to destination”.

---

<sup>\*</sup> kira.kowalska.13@ucl.ac.uk

<sup>†</sup> j.shawe-taylor@ucl.ac.uk

<sup>‡</sup> p.longley@ucl.ac.uk

The preferred routes are inferred as ‘topics’ from vehicle GPS data using *latent Dirichlet allocation*, a widely used topic modelling technique. Police journeys are then simulated as sequences of the inferred topics.

## 2.1. Topic modelling

Topic modelling techniques were originally developed to discover main themes that pervade a large collection of documents (Blei et al., 2003). At the moment, they are being adopted by other disciplines as well (e.g. population genetics (Pritchard, Stephens, & Donnelly, 2000), image classification (Fei-Fei & Perona, 2005)) with the purpose of finding underlying themes in unstructured collections of items.

In this project, we attempt to develop a novel application of topic modelling to route choice modelling. Although limited previous research used topic models to discover mobile behavioural patterns (Huynh, Fritz, & Schiele, 2008)(Farrahi & Gatica-Perez, 2012), no research has applied topic models to *vehicle GPS data* in particular. On top of that, no research has adopted the discovered patterns for movement modelling. Therefore, the work presented in this report, despite its premature state, could potentially contribute to the research community not only by applying topic modelling to a new type of data, but also through an innovative use of the discovered topics in vehicle route choice modelling.

The most common topic modelling algorithm and the one used in this project is *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003). In the context of document analysis, it defines a topic as a distribution over a fixed vocabulary. It assumes that each document exhibits multiple topics and that words contained in the document are generated by firstly randomly picking a topic from the document’s distribution over topics and then by randomly choosing a word from the topic’s distribution over the vocabulary. Parameters defining the generative process are tuned to the data using a sampling technique called *Gibbs Sampling* (Steyvers & Griffiths, 2007).

Our research project requires novel interpretations of topics, documents and words. In our context of vehicle movement, words become street segments and documents become vehicle journeys (collections of visited street segments). Discovered topics are distributions over street segments. The topics reveal underlying clusters of street segments based on their co-appearance in vehicle journeys and hence enable a simplified representation of vehicle journeys as distributions over the topics.

## 2.2. Route choice modelling using topics

Topics inferred from the data are used to model police vehicle movement. The original street network is reduced to a *topic network*, in which each *topic node* is a collection of street segments that have the highest probability of appearing in that topic. Topic nodes are connected by an edge if their street segments are physically connected. Vehicles that want to travel from one street segment to another start at the topic node where the first segment is assigned and take the shortest path through the topic network from that node to the topic node containing their journey destination. This modelling framework reflects the intuition that journeys are undertaken in stages or as series of topics on route to the destination.

Two variants of the modelling framework are explored. In the first one, the topic network is *unweighted* and vehicles choose paths that minimise the number of topics traversed to their destination. In the second one, topic nodes are *weighted* by the total length (in metres) of street segments assigned to them. Vehicles subsequently choose paths that minimise journey length.

## 2.3. Model validation

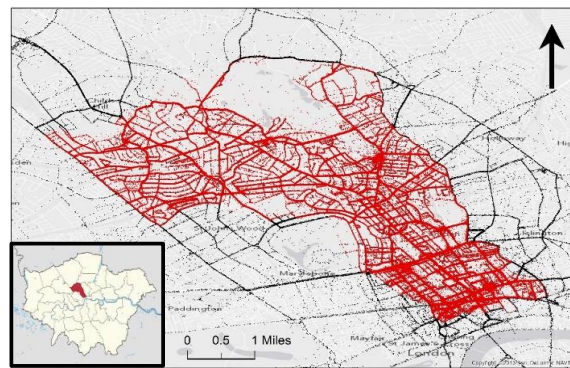
The models are validated against the data by measuring the *Pearson correlation coefficient* (Agresti, 2008) between street coverage generated by the models and the actual street coverage, where coverage is defined as the number of vehicle visits at each street segment in Camden. The generated street coverage contains journeys between all possible origins and destinations in the Camden’s street network, weighted by the probability of observing such an origin-destination pair in the actual data.

Since the models are probabilistic, the generated coverage at a street segment is represented by the sum of probabilities of all possible journeys that would visit that segment. An alternative approach would be to simulate agents (vehicles) based on the probabilistic rules and then use their journeys to calculate the generated coverage. The alternative approach would be prone to sampling bias though, especially if the number of agents was small.

### 3. Results and discussion

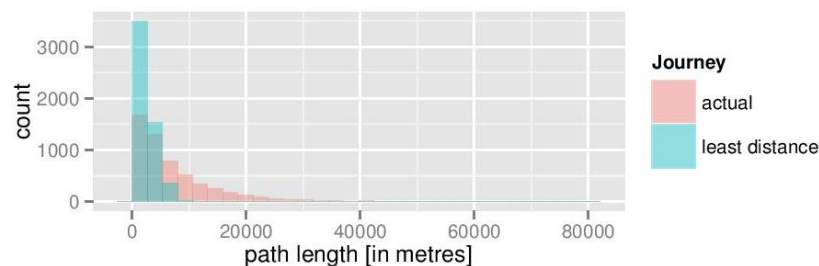
#### 3.1. Police Vehicle Data

Police vehicle data motivating the project were released for research purposes in May 2012 as part of the “Crime, Policing and Citizenship” project<sup>§</sup>. The data include all GPS signals transmitted by police vehicles in the London Borough of Camden in the months of March 2010 and March 2011 (1,188,953 GPS signals in total). The frequency of GPS signal transmissions is roughly every 15 seconds.



**Figure 1.** GPS signals transmitted by police vehicles inside (red) and outside (black) the London Borough of Camden in March 2010.

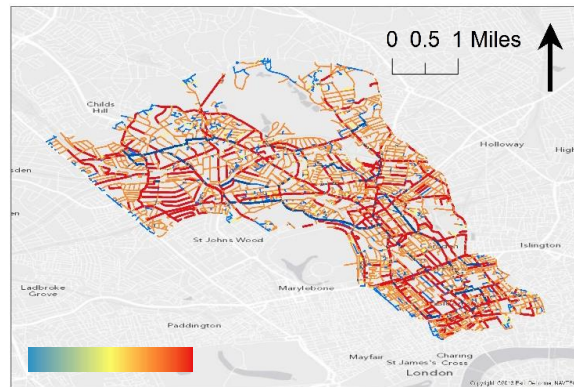
Police route choice preferences observed in the data are sub-optimal in terms of journey length as shown in Figure 2 and strongly biased towards the use of major roads (Figure 3). These patterns are in line with the findings by Manley (2014) that underpin our models.



**Figure 2.** Distributions of lengths of journeys in March 2010 and their least distance alternatives.

<sup>§</sup> UCL Crime Policing and Citizenship: <http://www.ucl.ac.uk/cpc/>.



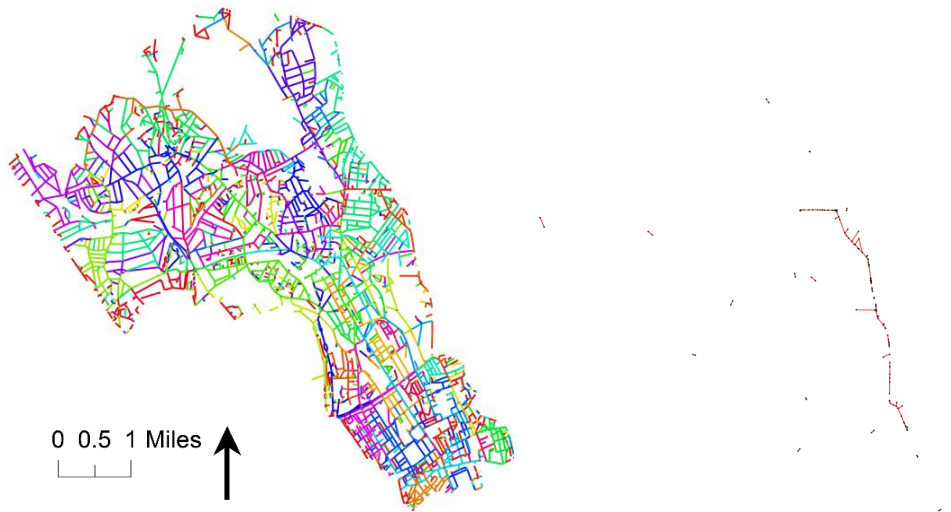


**Figure 3.** Difference between actual and least distance journeys (*actual minus least distance*) in March 2010; yellow corresponds to exact match.

### 3.2. Route choice modelling

Topics inferred from the data are shown in Figure 4. Latent Dirichlet allocation algorithm required specifying the number of topics *a priori* and this number was set to hundred following initial experiments into the influence of the number of topics on topic sizes. However, it is acknowledged that further research is required into optimising the number of topics for modelling purposes.

The assignments of segments to topics in Figure 4 (left) seem to reflect their network proximity, as well as a general road hierarchy. Segments of major roads tend to be clustered together forming ‘stretched’ topics, whereas segments of minor roads seem to be clustered within their neighbourhoods. These observations reflect the intuition that vehicles tend to travel longer distances along major roads but otherwise limit their journeys to nearby locations.

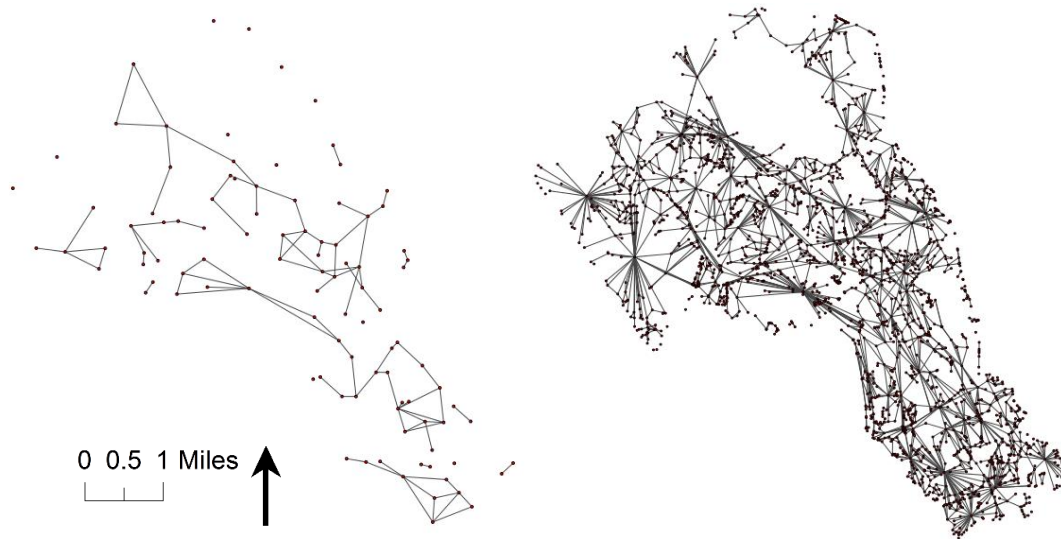


**Figure 4.** (left) Camden’s street network coloured by topic assignments, (right) an exemplary topic, when hundred topics are inferred from police journeys in March 2010 in the London Borough of Camden.

The inferred topics are used to create a topic network. This proves to be problematic as a closer investigation of topics in Figure 4 reveals that topics are often disconnected (see example topic in Figure 4 (right)). The discontinuity might be due to our simplistic approach to assigning segments to topics,

in which a topic is treated as a defined collection of street segments rather than a distribution over all street segments. The issue requires further investigation though, which is outside the scope of this paper.

For modelling purposes, the discontinuity is tackled by creating a topic graph in which each topic is represented by multiple nodes, each representing one of its connected components. The resulting topic graph is shown in Figure 5 (right). An alternative approach could construct the graph based on the largest connected component of each topic only. This would, however, lead to a disconnected graph as shown in Figure 5 (left).

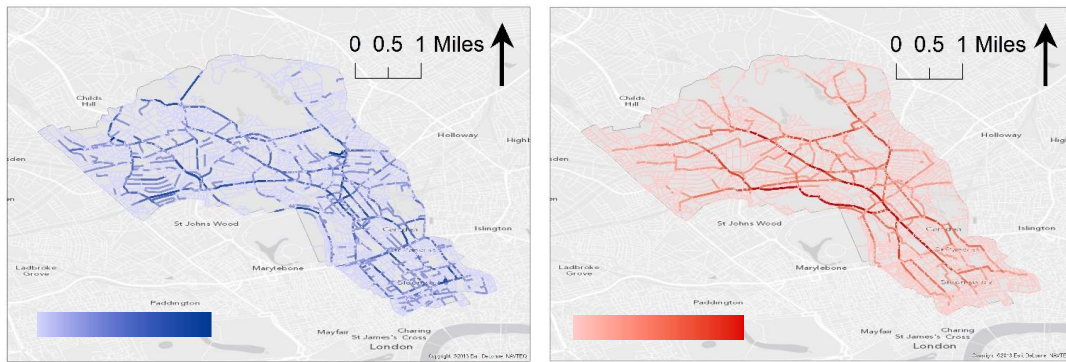


**Figure 5.** Topic graph created from (left) only the largest connected component of each topic, (right) all connected components of each topic, when hundred topics are inferred from police journeys in March 2010.

The topic graph is used to model police vehicle movement according to the procedures introduced in Section 2.2. The generated coverages for the *unweighted* and *weighted* model variants are shown in Figure 6. Their correlations with the actual police coverage (in Figure 7) are **0.204** and **0.349** respectively. Although these correlations are not substantially higher than a correlation of 0.315 between the data and a simplistic model assuming that vehicles always follow least distance paths (also in Figure 7), coverage patterns that they generate reflect subtle route choice preferences visible in the actual police coverage that cannot be captured under the unrealistic least distance assumption.



**Figure 6.** Police coverage generated by (left) unweighted and (right) weighted, topic graph from Figure 5 (right).



**Figure 7.** (left) Actual police coverage in March 2010 and (right) police coverage generated under the least distance assumption.

Further work is required to uncover the full potential of using topics in movement modelling. Possible extensions of the work presented in this scenario include:

- automated inference of the number of topics that would maximise modelling accuracy,
- higher order correlation metrics to measure model accuracy (e.g. correlation between counts on adjacent segments could better reflect the accuracy of a model in reconstructing movement patterns),
- addition of network connectivity information to the topic modelling algorithm in order to increase connectedness of discovered topics.

#### 4. Biography

Kira Kowalska is a first year PhD student in the Jill Dando Institute of Crime and Security Sciences at University College London. Her main research interests lie in the area of machine learning and network analysis, particularly in application to crime and security issues.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include 14 books and more than 125 refereed journal articles and book chapters. He is a former co-editor of the journal *Environment and Planning B* and a member of four other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.

John Shawe-Taylor is a professor at University College London (UK) where he is the Head of the Department of Computer Science. His main research area is Statistical Learning Theory, but his contributions range from Neural Networks, to Machine Learning, to Graph Theory.

#### References

- Agresti, A. (2008). *Statistical Methods for the Social Sciences* (4th ed.). Upper Saddle River, N.J: Pearson.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Farrahi, K., & Gatica-Perez, D. (2012). Extracting mobile behavioral patterns with the distant n-gram topic model. In *Proceedings of the 16th International Symposium on Wearable Computers* (pp. 1–8). IEEE.

- Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 524–531).
- Huynh, T., Fritz, M., & Schiele, B. (2008). Discovery of activity patterns using topic models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*.
- Manley, E. J. (2013). *Modelling Driver Behaviour to Predict Urban Road Traffic*. University College London.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7), 424–440.

# Optimising sentiment analysis in commercial context

Radoslaw Kowalski<sup>1</sup>

Department for Civil, Environmental and Geomatic Engineering

13 December 2014

## Summary

The proposed paper is about data analysis of consumer reviews of Argos products. The research objective is to produce new insights from the data already available. Argos managers will use research outcomes to reduce the percentage of products returned from customers. Methodological innovation for the purpose of this research assignment is to improve on the sentiment analysis model used currently, including an attempt to detect and interpret irony. Furthermore, this study aims at an identification of best solutions for presenting the resultant data summaries.

**KEYWORDS:** sentiment analysis; topic modelling; Argos; product reviews

## 1. Introduction

This study is to help Argos improve the quality of analysis for the product reviews submitted by customers. Product reviews are a critical source of information for Argos product category managers to know whether and how Argos products meet customer expectations. This study is expected to make it easy for the managers to access insight from product reviews, and in consequence help them cut operational costs by reducing the percentage of product returns on the total sales volume. Sentiment analysis, an analysis of attitudes expressed in text, is the most appropriate general approach for the task of automatically identifying what customers think about products from the textual data of product reviews.

## 2. Possible research designs for sentiment analysis

There is a range of possible approaches for carrying out sentiment analysis to learn about opinions of authors of texts. Broadly speaking, sentiment analysis requires a training set of data – a list of words or

---

<sup>1</sup> radoslaw.kowalski.14@ucl.ac.uk

documents that are processed with statistical methods to automatically identify which sentences from the text data being analysed are to be associated with negative comments, which ones with positive comments and which ones with neutral comments in relation to a given subject (Balahur et al. 2014). Methodological innovation can involve changes to how the training sets of data are built (e.g. Balahur et al. 2012; Duric & Song 2012) and used (e.g. Kennington & Schlangen 2014; Tang et al. 2009). Sentiment analysis can be carried out on individual words or symbols such as emoticons (Ptaszynski et al. 2014) and on collocations of words (Balahur et al. 2014) within documents or within sentences.

Product reviews used in this study are about 6 sentences in length and go together with star ratings of a product and its select attributes such as quality or value for money. The current data analysis method used in Argos makes sentiment assumptions about consumer reviews based on the star ratings, but resulting data summaries still do not make it clear which product features were key to determining the overall sentiment score. The failure to represent the data precisely could be a reason why the insight produced through the current analytical method is not widely used and acted upon within the company. The current method will be compared with several alternative approaches (see Table 1) so as to identify the best solution that makes insights from the data summaries more actionable. Two evaluation criteria are to be used to compare the models:

- the highest percentage of success in assigning sentiments to product attributes mentioned in text
- a high level of acceptance among the end users of the analyses, who would judge the methods based on the usefulness of the produced insight

**Table 1** Possible research approaches and their descriptions

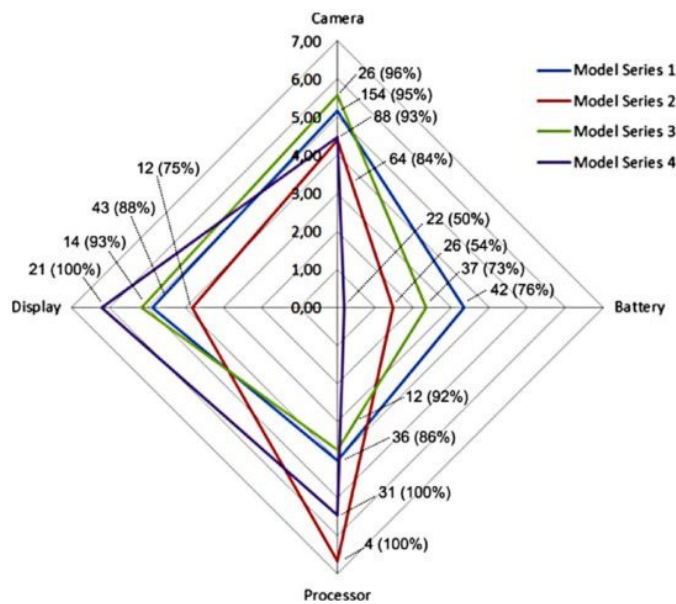
Approach	Description
Support Vector Machines and/or Naïve Bayes theorem	Both methods are frequently used for text analysis (Barhan & Shakhomirov 2012). Bayesian algorithm is simpler to use with short messages, but at the cost of efficiency (Barhan & Shakhomirov 2012). They can be used to analyse data through classifying a set of textual data into sentiment categories automatically, or by relying on a previously provided lexicon with coded sentiment readings (Barhan & Shakhomirov 2012). Other researchers, e.g. Carstens (Carstens 2011), have also modified and improved on those commonly used techniques.
Markov Logic Networks	A statistical method with high potential to analyse data within its thematic context. This statistical method can automatically adapt sentiment analysis to its particular context in which it takes place (Kennington & Schlangen 2014).
Topic modelling	The text can be analysed to find out the dominant topics that pervade through the comments about certain products. It is possible to get a quick glimpse of customers' thoughts about the product this way (Blei 2012). Potentially, the content ascribed to each identified topic could be further analysed with sentiment analysis.
Combining product review data with other data types	Amazon, a competitor of Argos, uses both formal product reviews from the media and individual customers' reviews from the Amazon website to predict sales. The two types of reviews influence each other as well as sales levels (Bao & Chang 2014). It is not certain yet, however, if correlations can be found between product returns and the sentiment readings from data types other than product reviews.
Different methods combined	It is possible to triangulate the results of different methodological approaches in sentiment analysis to cut out misleading bits of text from



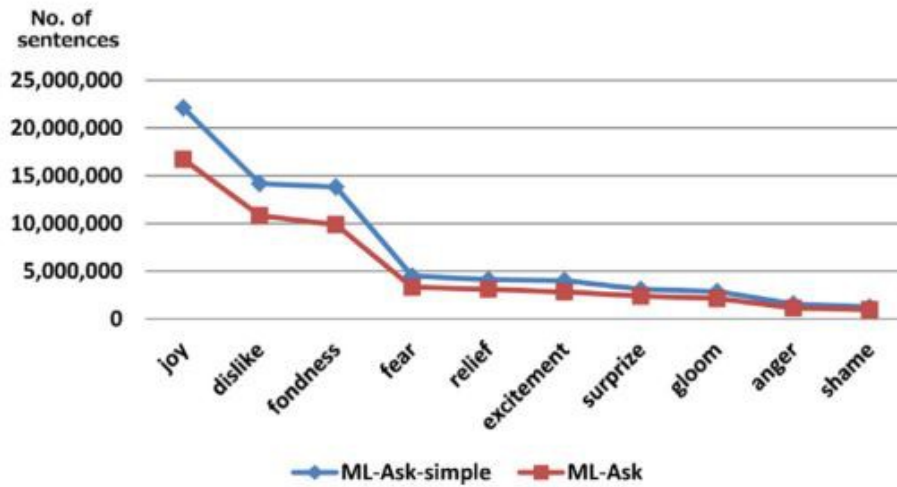
Irony detection	<p>analysis. The analysis may become more accurate through exclusion of misleading fragments of text (Tang et al. 2009).</p> <p>Irony detection also has potential to be used in analysing customer reviews. Ironical comments are known for their potential to have both negative and positive consequences for sales (Reyes &amp; Rosso 2012). So far, however, the most widely used methods have tended to exclude or misclassify ironical comments in the analysis (ibid.).</p>
-----------------	---

### 3. Data representation of sentiment analysis

Representation of sentiment analysis output has not been specifically discussed in any published works reviewed on sentiment analysis so far. There are notable examples of good practice, however (see figures 1 and 2). Interesting examples can also be found in other sciences, such as spatial data analysis, where projects like DataShine convert hard-to-process census data into easy-to-understand interactive maps (Cheshire & O'Brien n.d.). Furthermore, there are also very good visualisation tools available, such as Plotly ([www.plotly.ly](http://www.plotly.ly)) – an online tool that makes data visualisations interactive. Part of this research is to verify which data visualisation methods would be most effective in an authentic commercial context. The visualisation tools must adapt to the challenges of time scarcity and valuable insight generation that is easy to access.



**Figure 1** Sentiment scores for product attributes of similar products, in this case smartphones (Kontopoulos et al. 2013).



**Figure 2** Graphical visualisation of sentence level emotion class annotations done using two techniques, ML-ask and ML-ask-simple (Ptaszynski et al. 2014).

#### 4. Conclusion

This study tackles the problem of improving sentiment analysis through an innovative use of available methodological approaches in an applied context. Furthermore, the study attempts to systematically identify the most valuable forms of expression to the users of insight. After all, a statistical robustness of a statistical model or a seemingly effective form of representation may not necessarily always coincide with user preferences. Sentiment analysis, when applied in a business context, needs to strike a balance between the need for accuracy and simplicity.

#### 3. Acknowledgements

This research is possible thanks to the scholarship of Economic and Social Research Council. I would also thank dr Slava Mikhaylov and dr Helena Titheridge, my academic supervisors at University College London that support me throughout my PhD studies. Furthermore, I owe special thanks to prof Paul Longley, Guy Lansley from UCL and to the Home Retail Group for setting up my project and supporting me with good advice.

#### 4. Biography

Radoslaw Kowalski joined UCL in September 2014 to contribute to research into how unstructured data can be analysed automatically. Radoslaw's research interests lie in the analysis of social media and other spontaneously created texts to learn about personal characteristics of writers without reference to other data types.



## References

- Balahur, A., Hermida, J.M. & Montoyo, A., 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4), pp.742–753. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0167923612001352> [Accessed November 30, 2014].
- Balahur, A., Mihalcea, R. & Montoyo, A., 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28(1), pp.1–6. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0885230813000697> [Accessed November 18, 2014].
- Bao, T. & Chang, T.S., 2014. Why Amazon uses both the New York Times Best Seller List and customer reviews: An empirical study of multiplier effects on product sales from multiple earned media. *Decision Support Systems*, 67, pp.1–8. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0167923614002012> [Accessed November 6, 2014].
- Barhan, A. & Shakhomirov, A., 2012. Methods for Sentiment Analysis of Twitter Messages. In *12th Conference of FRUCT Association*.
- Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), pp.77–84. Available at: <http://dl.acm.org/citation.cfm?doid=2107736.2107741>.
- Carstens, L., 2011. *A multimodal approach by*. Imperial College London.
- Cheshire, J. & O'Brien, O., DataShine. Available at: <http://datashine.org.uk/#zoom=12&lat=51.52&lon=-0.15&layers=BTTT&table=QS411EW&col=QS411EW0007&ramp=Y1OrRd> [Accessed December 13, 2014].
- Duric, A. & Song, F., 2012. Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems*, 53(4), pp.704–711. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0167923612001340> [Accessed November 30, 2014].
- Kennington, C. & Schlangen, D., 2014. Situated incremental natural language understanding using Markov Logic Networks. *Computer Speech & Language*, 28(1), pp.240–255. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0885230813000491> [Accessed November 22, 2014].
- Kontopoulos, E. et al., 2013. Expert Systems with Applications Ontology-based sentiment analysis of twitter posts. *Expert Systems With Applications*, 40(10), pp.4065–4074. Available at: <http://dx.doi.org/10.1016/j.eswa.2013.01.001>.
- Ptaszynski, M. et al., 2014. Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. *Computer Speech & Language*, 28(1), pp.38–55. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0885230813000375> [Accessed November 30, 2014].
- Reyes, A. & Rosso, P., 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4), pp.754–760. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0167923612001388> [Accessed November 23, 2014].
- Tang, H., Tan, S. & Cheng, X., 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7), pp.10760–10773. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0957417409001626> [Accessed November 7, 2014].

# Spatio-Temporal Patterns of Passengers' Interests at London Tube Stations

Juntao Lai<sup>\*1</sup>, Tao Cheng<sup>†1</sup>, Guy Lansley<sup>‡2</sup>

<sup>1</sup> SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental & Geomatic Engineering, University College London

<sup>2</sup> Department of Geography, University College London

April 15, 2015

## Summary

With as many as 3.5 million passengers using the London underground system every day, it is desirable to examine and understand their interests and opinions, and to harness this information to improve the services of Transport for London (TfL). This research aims to achieve a better understanding of passengers' interests by harvesting text from geo-tagged Tweets sourced over a four week period in 2014 from the vicinity of the stations. An unsupervised topic modelling method Latent Dirichlet Allocation (LDA) is used to generate topics, and k-means is used to cluster stations in order to understand the overall patterns of topics.

**KEYWORDS:** social media data, spatial-temporal computation, K-means, topic modelling

## 1. Introduction

More than 3.5 million passengers use the London underground system every day, and more than 1 billion journeys are made every year (TfL, 2014). It is desirable to examine and understand the interests of passengers, and harness this information to improve the services of Transport for London (TfL), or make fuller use of the commercial advertisement potential inside the underground stations. However, it is difficult to collate such information directly from the millions of underground users. With the recent advance of smartphones and internet coverage, microblogging applications such as Twitter have been widely used. Twitter is real-time and is more widely spread compared with other blogging systems and traditional media (Java et al., 2007; Zhao et al., 2011). This creates opportunities to understand the interests of populations in space and time.

Based on the assumption that Tweets around underground stations are likely to be posted by underground users, this research aims to achieve a better understanding of the interests of London Underground passengers. This is achieved by harvesting text from geo-tagged Tweets sourced over a four week period in 2014 from the vicinity of the stations. An unsupervised topic modelling method Latent Dirichlet Allocation (LDA) is used to generate topics from Tweets, and k-means is used to cluster stations in order to understand the overall patterns of topics.

## 2. Spatio-Temporal patterns of Twitter topics at London tube stations

Here we describe the major steps to detect the topics of tweets and present their spatial-temporal distribution. The results are further clustered in order to understand the overall patterns of topics.

---

<sup>\*</sup> Juntao.lai.13@ucl.ac.uk

<sup>†</sup> Tao.cheng@ucl.ac.uk

<sup>‡</sup> G.lansley@ucl.ac.uk

## **2.1. Data Description**

The Twitter data is downloaded from the Twitter Streaming Application Programming Interface (API) service (Twitter Developers, 2012). The temporal range of the data is 4 weeks, from Feb 25 to Mar 24 in 2014. The data has been divided into two groups: the weekday group (Tuesdays, Wednesdays and Thursdays) which has 887,600 Tweets; and the weekend group (Saturdays and Sundays) which has 687,700 Tweets.

## **2.2. Assign Tweets to unique tube stations**

Given the density of London tube stations, a strategy is required to assign individual Tweets to unique tube stations. This is achieved by defining a unique coverage for each tube station, which is generated by using Thiessen (Voronoi) Polygons (Franz, 1991). Tweets are assigned to the station corresponding to the Thiessen polygon they fall within. Since we are mainly interested in topics discussed near or at the tube station, the Tweets within a 10-minute walking distance are extracted as the Tweets for further analysis.

## **2.3. Generate hot topics from Twitter data using LDA**

A script using the R package named “tm” (Feinerer, 2014) was used to remove the “noise” from the Twitter data, which includes the process of removing the whitespaces, numbers, punctuations and stopwords, also converting all the upper case to lower case. The process of “stemming”, which required an R package named “SnowBallC” (Bouchet-Valat, 2014), was also used to reduce inflected words to their stem form, by removing suffixes of the words.

After text cleaning, a topic modelling method named Latent Dirichlet allocation (LDA) is used to generate topics. LDA is an unsupervised generative model which can be used to classify text documents (Blei et al., 2003). A document is made up of groups of words which may belong to different topics. A topic is a bag of words, with corresponding probabilities of each word belonging to this group. LDA is a process that tries to backtrack from the documents to find a set of topics that are likely to have been generated by the collection. LDA represents documents as mixtures of topics that spit out words with certain probabilities (Chen, 2011). The main idea of LDA topic modelling is that the words that appear together many times in the documents are assumed to be related or present similar meaning, and are therefore more likely to be assigned to the same topic. It is more efficient and objective than manually classifying the text.

LDA has been used widely for news and text analysis, but its application on analysing short and informal documents like Tweets has only been implemented recently for detecting and analysing big events, such as earthquakes (Caragea, et al., 2011), the outbreak of flu (Chew et al., 2010), or special events (Cheng and Wicks, 2014). Here we utilise techniques of semantic analysis of Tweets to investigate the daily interests (topics) of underground users. An R package named “lda” (Chang, 2013), written by Jonathan Chang, was used to generate the topics from Tweets in this study.

## **2.4. Clustering stations using K-means**

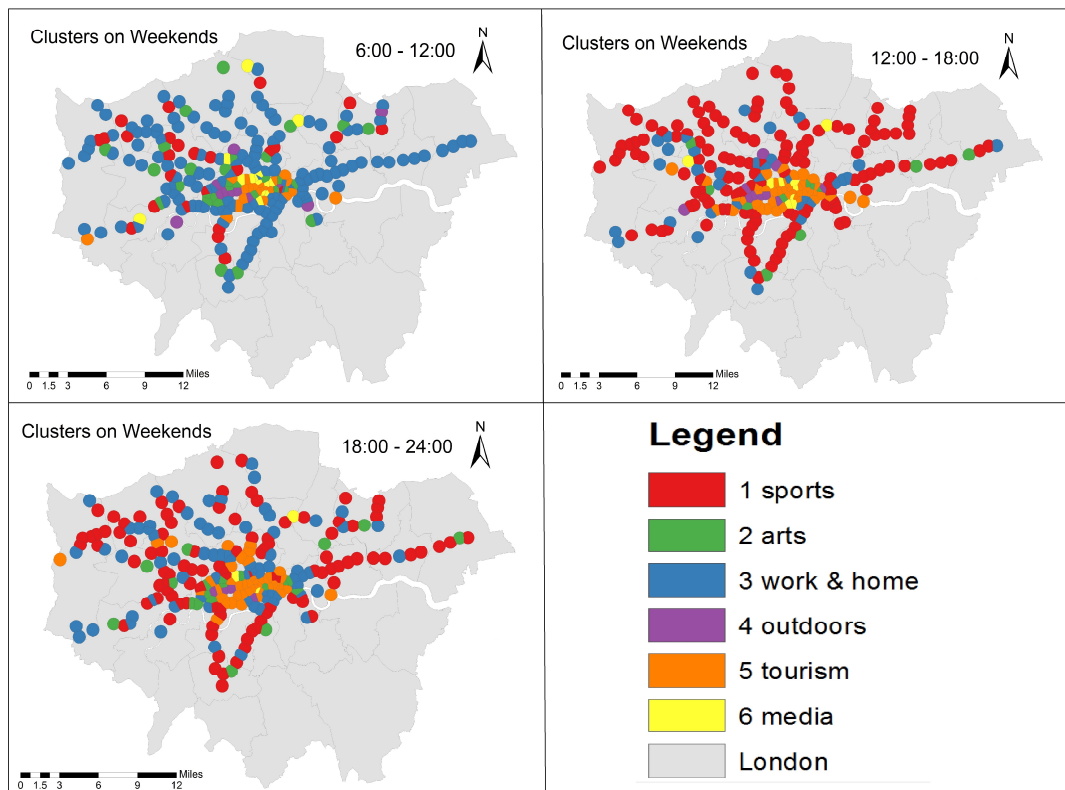
A general spatial temporal pattern of the topics can be revealed by focusing on the most frequent topics, but this only represents one variable among all informative ones (topics). In order to investigate the interest of the users more comprehensively, the influence of all topics that have been generated should be included. Therefore, the stations having similar topic distributions need to be classified into groups.

The classification of the stations based on their topic distributions was carried out using K-means clustering method. K-means clustering (MacQueen, 1967) enables the data to be clustered into a pre-specified number of groups. Based on the results generated by LDA, a table was created to store the topic distributions in both spatial and temporal dimensions. The percentage of the Tweets belonging to each topic were input as variables.

### 3. Results

The Twitter data was clipped by station assigned polygons and each Tweet was joined with one unique station name as an attribute. After data cleaning and formatting, the Tweets were fitted into the LDA topic model, then the words were assigned into groups by the model. After interpreting the groups using their top words, 10 meaningful topics were selected and labeled manually.

Using the topic distributions calculated from LDA, the stations are classified into groups. The categorization of the stations after clustering are displayed on the maps (Figure 1), it can be seen that the major groups of the users on weekends are cluster 1 and cluster 3. According to the cluster center information, cluster 1 has great value in “sports”, hence it could be interpreted as sports fans. Cluster 3 could be represented as working population due to its high proportions in “traffic” and “work-home lifestyle”. The other big group is cluster 5. Based on its big numbers in “tourism & travel” and “food & drinks”, and relatively low values in other topics, this group is more likely to represent tourists. The distributions of clusters on Figure 1 reflect reasonable patterns of those groups. For example, the central areas are mainly occupied by tourist groups, and people are more likely to talk about sports in the afternoon than in the morning.



**Figure 1:** Clusters in Different Time Periods on Weekends

### 4. Conclusion

This study demonstrated the effectiveness of LDA topic modeling in extracting topics from Twitter data, as previously hypothesized. However, many extensions to LDA were suggested to improve the quality of extracting and labelling the topics, such as supervised LDA (McAuliffe, 2007) and Labeled-LDA (Ramage, 2009) which may fit Twitter data better.

This paper also presented as a case study that successfully generated the main topics that underground

passengers discuss using Twitter data, and then explored the distributions of these topics spatially and temporally. Furthermore, these topics are reasonable and distinguishable in terms of representing daily activities of the public. Finally, the stations were classified into several groups in each time periods according to their topic distributions, which could be helpful to better understand the interest of underground passengers from a statistical perspective.

## 5. Biography

Juntao Lai is a PhD student in department of Civil, Environmental and Geomatic Engineering at University College London. His research interest includes semantic and sentiment analysis of social media data and spatial-temporal analytics. His current work is investigating the impact of media to the public satisfaction on policing using Twitter data.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimelab>), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

Guy Lansley is a Research Associate at the Consumer Data Research Centre, UCL, an ESRC Data Investment. His previous research has included exploring the temporal geo-demographics derived from social media data, and identifying socio-spatial patterns in car model ownership in conjunction with the Department for Transport. Whilst, his current work entails exploring population data derived from large consumer datasets.

## References

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003) "Latent dirichlet allocation." *the Journal of Machine Learning Research* 3: 993-1022.
- Bouchet-Valat, Milan. (2014) "Package 'SnowballC'. R package version 0.5.1. [ONLINE] Available at: <http://cran.r-project.org/web/packages/SnowballC/> [Accessed 10 August].
- Caragea, Cornelia, et al. (2011) "Classifying text messages for the Haiti earthquake." *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*.
- Chang, Jonathan. (2013) "Package 'lda'. R package version 1.3.2. [ONLINE] Available at: <http://cran.r-project.org/web/packages/lda/lda.pdf> [Accessed 16 July 14].
- Chen, Edwin. (2011) *Introduction to Latent Dirichlet Allocation*. [ONLINE] Available at: <http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/> [Accessed 15 July 14]
- Cheng, T. & Wicks, T. (2014). Event Detection using Twitter: A Spatio-Temporal Approach. *Plos One*, 9(6), e97807
- Chew, Cynthia, and Gunther Eysenbach. (2010) "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak." *PloS one* 5.11: e14118.
- Feinerer, Ingo. (2014) "Introduction to the tm Package Text Mining in R." *Comprehensive R Archive Network*. [ONLINE] Available at: <http://cran.r-project.org/web/packages/tm/index.html> [Accessed 16 July].
- Franz Aurenhammer (1991). Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys*, 23(3):345–405
- Java, Akshay, et al. (2007) "Why we twitter: understanding microblogging usage and communities." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM.

- MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).*
- Mcauliffe, Jon D., and David M. Blei. (2008) "Supervised topic models." *Advances in neural information processing systems (pp. 121-128).*
- Ramage, Daniel, et al. (2009) "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.* Association for Computational Linguistics.
- Transport for London. (2014). *London Underground – Factsheet.* [ONLINE] Available at: <http://www.tfl.gov.uk/cdn/static/cms/documents/lu-factsheet-jan2012.pdf> [Accessed 29 June 14].
- Twitter Developers. 2012. *The Streaming APIs.* [ONLINE] Available at: <https://dev.twitter.com/docs/streaming-apis> [Access 03 July 14]
- Zhao, Wayne Xin, et al. (2011) "Comparing Twitter and traditional media using topic models." *Advances in Information Retrieval. Springer Berlin Heidelberg.* 338-349.

# Mapping to Disrupt unjust urban trajectories

Lambert Rita, Allen Adriana

<sup>1</sup>Development Planning Unit, University College London

December 20, 2014

## Summary

This paper shares the experience of the research project 'Mapping Beyond the Palimpsest' which adopts grounded applications and cutting edge technologies for community-led mapping and visualization, to reframe the understanding of, and action upon, two highly contested territories in Lima; the Historic Centre - Barrios Altos, and Jose Carlos Mariategui, in the periphery. Adopting a participatory action-learning approach, the research seeks to disrupt the exclusionary trajectory of urban change, develop the writing of more inclusive representations and open up spaces for collectively negotiated outcomes between marginalised citizens, planners and policy makers.

Keywords: planning, mapping, visualisation, justice, Lima

## Title: Mapping to disrupt unjust urban trajectories

This paper shares the experience from the research project 'Mapping Beyond the Palimpsest' also known as 'ReMapLima', which brings together three departments within UCL: The Bartlett Development Planning Unit (DPU) [led by Adriana Allen and Rita Lambert], the Centre for Advanced Spatial Analysis (CASA) [Andrew Hudson-Smith and Flora Roumpani], and The Bartlett School of Architecture/UCL Urban Laboratory [Ben Campkin]. It is undertaken in close collaboration with CENCA, CIDAP and Foro Ciudades Para la Vida - a network of 57 organisations from 20 Peruvian cities, ranging from local government, academics and civil society groups - as well as local communities from two contested settlements in Lima, Perú.

Building upon the DPU research platforms 'The Heuristics of Mapping Urban Environmental Change' (<http://www.bartlett.ucl.ac.uk/dpu/mapping-environmental-change>), CASA's world-leading methodological innovations in spatial analysis, and the Urban Laboratory's 'Picturing Place' methodology, this research seeks to develop innovative and critical strategies for the reading, writing and audiencing of maps. It adopts a participatory action-learning approach, enabling local community mappers to explore innovative pathways for reframing their territory, disrupting the exclusionary trajectory of urban change and developing the writing of more inclusive representations. The research interrogates the role that hegemonic representations can play in bringing forth exclusionary socio-environmental processes and seeks to open up spaces for collectively negotiated outcomes between marginalised citizens, planners and policy makers, ultimately contributing towards the planning of more democratic and sustainable cities.

The two contested settlements under study in this research are Barrios Altos and Jose Carlos Mariategui.

**Barrios Altos (BA)** is located in the historic centre of Lima in an area declared UNESCO World Heritage Site (Figure 1). It is characterised by overcrowded conditions and lack of basic services. Due to its strategic location, it has high land values, however its buildings are left to deteriorate (Figure 2). The lack of public investment for its rehabilitation, together with the illegal change of use into storage, which is occurring at a fast pace, can be understood as a means to evict many of vulnerable inhabitants living on high value land. The process that converts many of the buildings into storage occurs through the retention of facades while gutting the interiors, in order to store containers with merchandise originating mainly from China.



Figure 1- Barrios Altos- The Historic centre



Figure 2- Collapsing buildings of Cultural value in Barrios Altos

The Municipality of Lima has created a special body, PROLIMA, which is in charge of the strategic vision for the renovation of the Historic Centre and its masterplan for 2025. Many of the maps brought together to justify this masterplan are produced by various governmental agencies, and include risk maps, socio-vulnerability maps, crime maps ect. In essence, these justify the renovation of the centre by depicting it as poor, crime ridden, and with high physical risk. Moreover, the renovation is promoted through private investment which would lead to gentrification and the expulsion of the inhabitants. Here, the research seeks to fundamentally challenge what is to be considered cultural heritage, moving beyond the attention solely on the architecturally valuable buildings, to also include the people that have been living there for generations.

The other case study is **Jose Carlos Mariategui**, is in the outskirts of Lima and has developed through a complex history of grassroots invasions and informal land trafficking. Like many other informal settlements in Lima, it is expanding on the steep slopes of the city's margins. In the absence of affordable housing and national housing policies, the occupation of the slopes is the only viable option for the vast masses of the urban poor, which are exposed to high levels of physical risk and water injustices (Figure 3 and 4). The occupation of such 'vertical' areas threaten the sustainability of the city as they coincide with the 'Lomas Costeras', an essential ecological infrastructure for recharging the aquifers that guarantee water for Lima and regulate the effects of climate variability.





Figure 3-The occupation of the slopes in Jose Carlos Mariategui



Figure 4-Exposure to high levels of physical risk

This paper will explain the process of mapping, the technology adopted (including drones, 3D printing, augmented reality and visualization), the spatial analysis undertaken together with local communities, and the learning acquired throughout this process. Two simultaneous modes of mapping were used: from the sky and from the ground. These were conceived with the purpose of articulating grounded applications and cutting edge technologies for community-led mapping and visualization and to explore the political agency and capacity of mapping to reframe the understanding of, and action upon, these highly contested territories.

**Mapping from the sky** used drones to capture 2D and 3D outputs (Figure 5 and 6). Moving beyond militaristic and surveillance applications, their value for planning was explored in the case studies undergoing otherwise 'invisible' change. In Barrios Altos, the bird's eye view captured through the use of drones made visible the otherwise 'unseen' processes of slow eviction and land use change occurring behind both conserved and deteriorating facades. In Jose Carlos Mariategui, the maps produced enable a detailed understanding of the difficult terrain that has to be negotiated by the inhabitants and the practices that make the slopes habitable. It also captured the shifting borders, the increased risk and the threatened ecological infrastructure, as well as the apprehension of the ravine as a system of interconnected settlements. Here the mapping is a means to open up dialogue between the inhabitants and the institutions which have various uncoordinated programs and projects in this area. It also seeks to enable the integrated planning of such areas and also support the preservation and management of the 'protected' ecological infrastructure for the functioning of the city.



Figure 5- Drone flying in Jose Carlos Mariategui

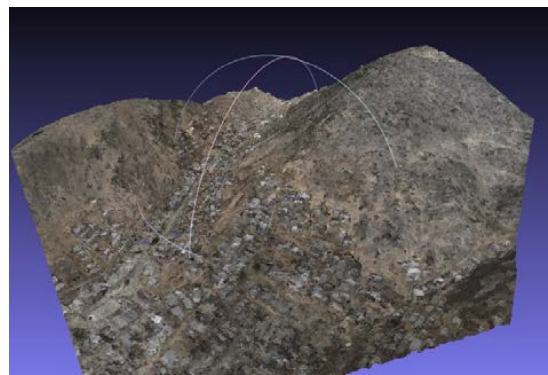


Figure 6- 3D output from drones

The outputs produced by the drones were then used as the basis for mapping from the ground. Moreover, the drone 3D outputs were sent remotely to London from Lima, in order to be 3D printed. These are converted into meshes and are then printed using a Makerbot 2 Replicator (Figure 7 and 8). The intention is to use the resulting models for community planning purposes in the two areas.

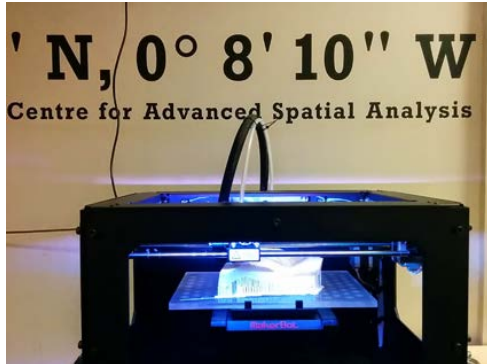


Figure 7- 3D printing of drone output



Figure 8- 3D model of Jose Carlos Mariategui

**Mapping from the ground** brought together men and women from each of the two settlements, to discuss and decide what to map, why and how. Mapping was discussed by the residents as a way to apprehend their territory and a means to document and denounce otherwise invisible processes of unwanted change in their neighbourhoods. It was also seen as a strategic activity to foster dialogue between stakeholders and to expand the room for manoeuvre to promote strategic interventions. Once decisions were made about what was important to map, pilot transects were traced to collect the required information. The fieldwork tested various crowd sourcing applications using mobile phones to collect data. Moreover, the pilot walks served to reflect upon and refine the mapping focus, the selected variables and methods for further mapping (Figure 9 and 10).



Figure 9- Preparing for a transect walk



Figure 10- Identification of illegal storage and evictions in Barrios Altos

In a subsequent phase, DPU and CASA are further developing a mapping methodology tailored to the unique contexts of Jose Carlos Mariátegui and Barrios Altos. This phase focuses on data collection as well as explore planning scenarios based on the printed 3D models captured by the drones. The process activated by 'Mapping beyond the Palimpsest' is also feeding into the other processes such as the practice module of the DPU MSc Environment and Sustainable Development, to deepen and consolidate this spatial mode of enquiry and action planning.

The main output, which is being developed, is a digital archive and mapping platform in José Carlos Mariátegui and Barrios Altos which would have information in various formats including text, photos, video and audio.

The latter will act as a repository for research on the urban Global South and an archive of knowledge accessible to those with interest in urbanisation, sustainability and justice. At the policy level, its intention is to encourage better planning practices, representing wider citizens' perspectives, ideals and aspirations. Moreover it seeks to stimulate public debate and awareness and be a showcase for cross faculty and cross global collaboration.

Two websites have been created to share the experience gained through the research 'Mapping Beyond the Palimpsest':

<https://www.bartlett.ucl.ac.uk/dpu/mapping-beyond-the-palimpsest>

<http://remaplima.blogspot.com/>

### **Acknowledgements**

The Bartlett Materialisation Grant for awarding £50 000 to the research.

Development Planning Unit-UCL

Centre for Advanced Spatial Analysis-UCL

Urban Lab- UCL

Drone Adventures

CENCA

CIDAP

Foro Ciudades Para La Vida

Over 30 community members from two case study sites - Barrios Altos and Jose Carlos Mariátegui - in Lima.

### **Biography**

Rita Lambert is a Teaching Fellow for the MSC in Environment and Sustainable Development and researcher at the DPU-UCL. She trained as an architect and has over 10 years of professional experience. She is currently undertaking a PhD Interrogating the role of mapping in planning in the urban global south.

Dr Adriana Allen is a professor of Development Planning and Urban Sustainability at the DPU-UCL, with over 25 years of international experience in academic research and applied work. Straddling the social and natural sciences and examining the interface between development planning, environmental change and urbanisation in the global south.

# Creating an Output Area Classification of Cultural and Ethnic Heritage to Assist the Planning of Ethnic Origin Foods in Supermarkets in England and Wales

Guy Lansley<sup>\*1</sup>, Yiran Wei<sup>†1</sup> and Tim Rains<sup>‡2</sup>

<sup>1</sup>Department of Geography, UCL

<sup>2</sup>J Sainsbury's plc

04 January, 2015

This paper presents a Cultural, Ethnic and Linguistic Output Area Classification for England and Wales built from clustering census variables which pertain to cultural identity. The study provides a quick insight into the broad patterns in ethnic segregation based on the residential geography recorded from the 2011 Census and is therefore a useful tool for supermarket planners seeking to identify areas where to target particular ethnic origin foods. To confirm this association, the classification has also been compared with the total sales of a selection of ethnic origin foods using supermarket customer loyalty data.

**KEYWORDS:** ethnicity, k-means clustering, census, food consumption

## 1. Introduction

Many minority ethnic and cultural groups in Britain have distinctive food consumption habits which emanate from their cultural origins (Uskul and Platt, 2014). With the ethnic minority population of the UK growing (Simpson, 2013), understanding a basic segmentation of ethnic compositions across England and Wales is useful to supermarket planners aiming to make their stores more appealing to local ethnic groups.

Ethnic groups can be considered as distinctive groups of individuals who share a common identity through kinship, religion, language, location, nationality and physical similarities from ancestry (Bulmer, 1996). However, each one of these domains can singularly be a defining characteristic of an individual's cultural identity. Therefore, the definition and classification of ethnicities have attracted on-going debate due to its multidimensional, subjective and complex nature (Mateos *et al*, 2009).

The 2011 Census for England and Wales identified that the population was becoming more ethnically diverse, largely due to immigration and higher fertility rates amongst most ethnic minority groups compared with the national average (Simpson, 2013). Typically many minority groups residentially cluster within urban areas due to a range of structural social and economic forces (Finney, 2013). While minority ethnic groups have been found to be dispersing more recently (Stillwell and Hussain, 2010), most metropolitan neighbourhoods still bare a more diverse ethnic composition than the rest of the country.

Using data at the output area level from the 2011 Census, this paper aims to identify the major spatial traits in ethnic identity across the residential geography of England and Wales by producing a Cultural, Ethnic and Linguistic Output Area Classification (CELOAC).

---

<sup>\*</sup> G.Lansley@ucl.ac.uk

<sup>†</sup> Yiran.Wei.13@ucl.ac.uk

<sup>‡</sup> Tim.Rains@sainsburys.co.uk

## 2. The 2011 Census for England and Wales

Output areas are the smallest geographical unit available from the 2011 Census data and have an average population of 309 (ONS, 2014). At this geography, over 400 variables relevant to the classification are available from seven Quick Statistics census tables. These cover key dimensions of cultural identity as labeled in Table 1.

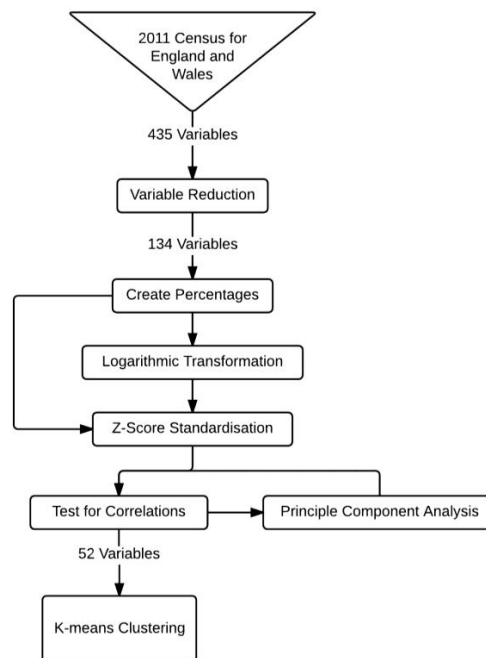
**Table 1** Census tables used in the classification

Census Table	Name
QS203EW	Country of birth (detailed)
QS204EW	Main language (detailed)
QS205EW	Proficiency in English
QS208EW	Religion
QS211EW	Ethnic group (detailed)
QS802EW	Age of arrival in the UK
QS803EW	Length of residence in the UK

Prior to the analysis an initial variable reduction was conducted to filter out variables which were not appropriate for a k-means classification. Variables with total populations below 10,000 were aggregated into broader groups based on their global regions of origin or removed altogether if they were considered too distinctive to merge.

## 3. Methods

The methodological approach for this study draws heavily on existing literature surrounding conventional geodemographic classifications (Harris et al, 2005), and most notably, the open source Output Area Classifications which are also produced exclusively from census data (Vicker and Rees, 2007). In a similar approach the data was standardised, tested for suitability and then clustered using the k-means algorithm, as displayed in figure 1.



**Figure 1** Flow diagram of methodological steps taken

The variables were standardised in order to reduce the effect of outliers across the dataset (Milligan, 1996). Much of the individual variables had positively skewed univariate distributions, largely due to low counts and a well-known tendency for cultural groups to cluster (Finney and Simpson, 2009). Therefore a natural log transformation for these cases was implemented, resulting in data set that is nearer to a normal distribution. In addition, a Z-score standardisation was applied so that each variable was presented on a common scale of standard deviations from mean.

Two steps were then undertaken to gauge the appropriateness of the remaining variables and to remove those which may unnecessarily skew the results. Firstly, a Pearson's correlation test was run between all variables to test for multicollinearity (Vicker and Rees, 2007). Of pairs of variables which correlated highly, either the smallest was removed or they were merged into 'other' groups if both variables were from the same census table and represented similar cultural groups, and did not correlate highly with a variable from a different table. Secondly, a Principal Component Analysis was undertaken to identify variables which may act erratically in the model (Rencher, 1998). Unstable cases were inspected and a handful were merged into broader variables or removed completely. Following these two steps many variables were aggregated and then retested.

In total 52 variables were selected for the classification (table 2). The variable with the smallest population out of the final selection, Russian language, represented over 67,000 persons.

**Table 2** 52 Variables used in the classification

Variable table	Variables used
Country of birth	China, Ghana, Hong Kong, Kenya, Middle East, Nigeria, Philippines, Romania, Somalia, USA, Other Central and Western Africa countries, Other EU accession countries, Other South and Eastern Africa countries, Other South-East Asia, Other Southern Asia
Main language	French, Russian, Turkish, African Language, East Asian Language, South Asian Language, West or Central Asian Language
Proficiency in English	Cannot speak English
Religion	Buddhist, Christian, Hindu, Jewish, Muslim, Sikh, No religion or not stated
Ethnic group	Afghan, African, Arab, Australian & New Zealander, Baltic States, Bangladeshi & British Bangladeshi, Black British, Caribbean, English/Welsh/Scottish/Northern Irish/British, Greek & Greek Cypriot, Indian & British Indian, Irish, Pakistani or British Pakistani, Polish, Sri Lankan, South East Asian, Other Eastern European, Other Western European
Age of arrival	0 to 4, 45 - 64
Length of residence	10 years or more, Less than 2 years

### 3.1 k-means clustering

The final 52 variables were clustered to create a composite classification using the k-means algorithm. K-means is an iterative allocation-reallocation method where the number of cluster groups (k) is predefined by the user (Harris et al, 2005). The approach creates distinctive cluster groups by attempting to minimise the sum of the distances from each case to their cluster centre based on the variable distributions. In its simplest form, the algorithm initially randomly seeds the cluster centres in a multidimensional space formed from the variables and every data case is allocated to its nearest centre. The cluster centres are then retested at the centroid of their current data allocation, and the process is repeated until the cluster centres cannot be moved as an optimum solution has been achieved (Harris et al, 2005).

With the intent of creating a classification with a relatively small number of clusters, a series of tests were run to help identify an appropriate number for the CELOAC. An 8 cluster solution was deemed to have the most appropriate average distance to the cluster centre and cluster size distribution.

#### 4. The Cultural, Ethnic and Linguistic Output Area Classification (CELOAC)

The CELOAC consists of 8 culturally distinctive groups. Two groups combined comprise of just over 70% of output areas in England and Wales, both contain higher proportions of the White British ethnic group than the remaining population, with rates of 88.5% and 96.2% respectively. As the focus of this research is on foreign origin ethnic groups the two white British clusters have been merged for the remainder of this paper (group G).

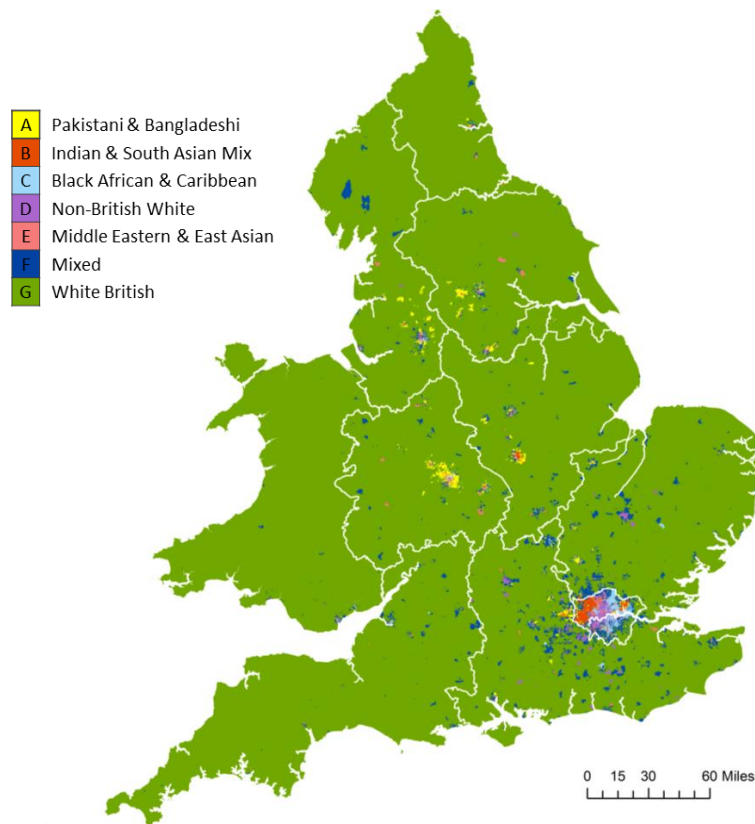
The remaining groups account for neighbourhoods where the proportion of ethnic minorities is above the national average, and these are largely in metropolitan regions. At the broadest level, the remaining groups identify four main characteristics, neighbourhoods with higher proportion of Asian ethnicities (groups A and B), neighbourhoods with higher proportions of black ethnicities (group C), and neighbourhoods with higher proportions of White, Middle Eastern and East Asian ethnicities (groups D and E). There is also a large group which represents a cosmopolitan mixture of ethnic groups (group F), the White British ethnic group is also well integrated here. The average proportion of key ethnic groups from the 2011 Census for each CELOAC group are presented in table 3.

**Table 3** The average percentage of ethnic groups by each CELOAC group

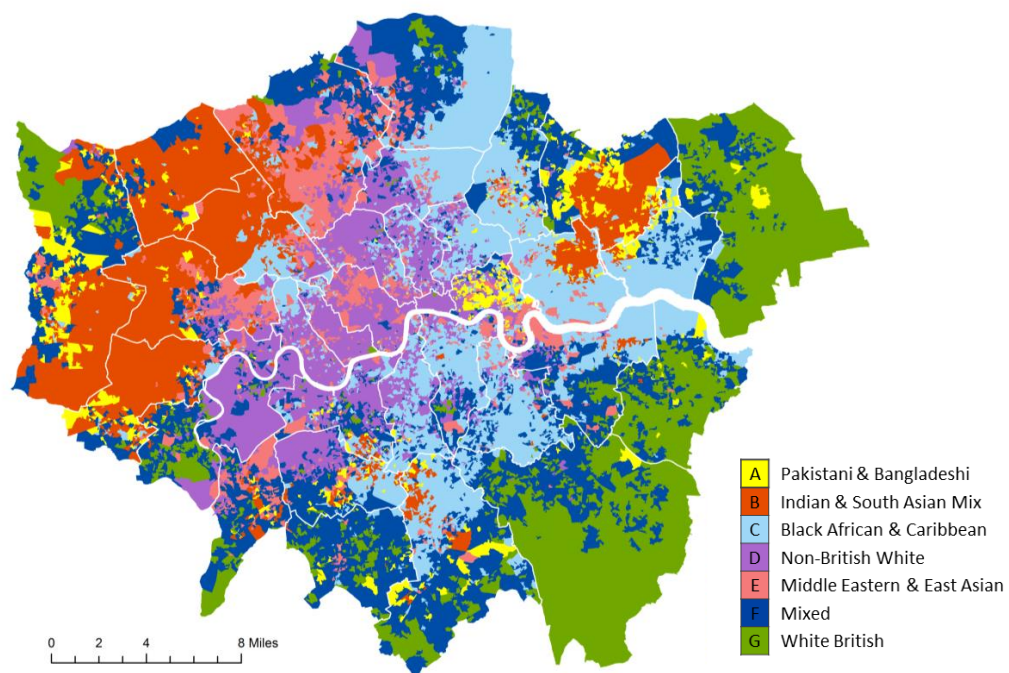
<b>Ethnic Group</b>	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>
White British	43.29	22.50	33.14	53.73	42.93	72.15	93.01
White Irish	1.21	1.70	1.80	2.82	1.71	1.47	0.64
Other White	4.69	8.62	12.40	20.63	14.24	7.96	2.06
Mixed & multiple	3.43	3.81	6.64	4.88	4.53	3.49	1.30
Indian	9.84	25.14	3.01	2.53	5.55	3.06	0.73
Pakistani	20.23	9.36	2.85	0.78	3.46	1.42	0.34
Bangladeshi	5.87	2.83	3.10	1.10	1.96	0.66	0.14
Chinese	0.66	0.90	1.35	2.09	5.89	1.30	0.30
Other Asian	3.11	10.24	3.92	2.88	5.66	2.49	0.43
Black	5.49	10.81	27.80	5.30	7.72	4.60	0.70
Arab	0.84	1.59	1.14	1.35	3.67	0.49	0.10
Other	1.25	2.38	2.69	1.85	2.60	0.77	0.17

The CELOAC groups have been labeled and mapped in figures 2 and 3. The labels only consider the key ethnicities which are overrepresented in each group and are only intended to aid interpretation for this study.





**Figure 2** A map of CELOAC in England and Wales



**Figure 3** A map of CELOAC in London



#### 4.1. Cultural heritage and ethnic origin food consumption

Ethnic and cultural identity can greatly influence consumption habits, and food especially (Kershen, 2002; Hamlett et al, 2008). The supermarket chain Sainsbury's provided the number of sales for six pre-selected grocery products by OA as recorded from their customer loyalty database, the data represented the total sales within a 52 week period commencing in May 2011. Each of the foods were chosen due to their distinctive cultural heritage with minority groups. As the total grocery expenditure per OA was not available, the data was standardised by the Census population. This data has been cross-tabulated by the CELOAC groups and the results are shown as location quotients whereby 100 represent an average penetration (table 4).

**Table 4** Sales of six ethnic origin foods by CELOAC group

Group	Black Eye Beans	Chickpeas	Chinese Leaf	Ghee	Halal	Ogorki
A	215.85	80.99	81.49	250.52	163.20	122.03
B	472.69	111.15	137.82	601.02	711.46	312.35
C	305.82	120.15	131.56	277.24	598.45	286.26
D	218.65	230.27	229.67	216.07	413.44	356.12
E	202.49	122.70	260.88	229.66	402.39	286.00
F	151.65	136.29	151.40	151.36	109.55	184.47
G:	50.40	87.58	78.99	44.77	19.14	49.37

The results identified substantial variations in the consumption of ethnic origin foods across the CELOAC groups. Generally, while each of the products may sell particularly well in one of the 'minority' clusters, they will often sell better in other minority clusters too, reflecting the cosmopolitan composition of their populations. The results are especially compelling considering migrant groups may be less likely to patronise Sainsbury's stores than the White British population.

#### 5. Conclusions

This study has presented an open-source output area classification of ethnic, cultural and linguistic characteristics for England and Wales and identified distinctive ethnic clusters. Whilst most rural and many suburban areas are homogeneously White British in composition. The inner cities of larger, more globally connected urban areas are composed of a more heterogeneous mix of cultural groups. As cultural groups cluster and segregate themselves from dissimilar communities, spatial mosaics of culturally distinctive neighbourhoods occur in major metropolitan areas, and London especially. This distinctive residential geography exerts an associated spatial variation in the consumption of ethnic origin foods, as identified by the supermarket data.

Whilst there are several disadvantages from devising a discrete categorisation of neighbourhoods, the approach presented provides an insightful snapshot into the contemporary cultural and ethnic geography of England and Wales.

#### 6. Biography

Guy Lansley is a Research Associate at the Consumer Data Research Centre, UCL, an ESRC Data Investment. His previous research at UCL has included exploring the temporal geo-demographics derived from social media data, and identifying socio-spatial patterns in car model ownership in conjunction with the Department for Transport. His current work entails exploring population data derived from large consumer datasets.

Yiran Wei is a recent alumnus of UCL. She studied the Geospatial Analysis MSc at the Department of Geography, UCL and specialised in geodemographics and ethnic clustering. She previously studied the

Environment and Development MA at King's College London and Environmental Science BSc at the University of Greenwich.

Tim Rains is a senior GIS analyst at J Sainsbury's plc. He also holds a GIS MSc degree from Birkbeck College and a Geography BA from the University of Plymouth.

## References

- Bulmer, M. (1996) 'The ethnic group question in the 1991 Census of population' in Coleman, D. and Salt J. (des.) *Ethnicity in the 1991 Census. vol.1 Demographic characteristics of the ethnic minority populations* London: HMSO
- Finney, N. (2013) "How ethnic mix changes and what this means for integration." In van Ham, M., Manley, D., Bailey, N., Simpson, L., MacLennan, D (Eds) *Understanding Dynamic Neighbourhoods*, New York: Springer
- Finney, N. & Simpson, L. (2009) *'Sleepwalking to segregation'? Challenging Myths about Race and Migration*, London: Policy Press
- Hamlett, J., Bailey A., Alexander, A. and Shaw G. (2008) Ethnicity and Consumption South Asian food shopping patterns in Britain, 1947 – 1975. *Journal of Consumer Culture* 8(1) 91-116.
- Harris R, Sleight P, Webber R. (2005) *Geodemographics: neighbourhood targeting and GIS*. Chichester, UK: John Wiley and Sons.
- Kershen, A. J. ed. (2002) *Food in the Migrant Experience*. Aldershot: Ashgate.
- Mateos P, Singleton A, Longley P (2009) Uncertainty in the analysis of ethnicity classifications: issues of extent and aggregation of ethnic groups. *Journal of Ethnic and Migration Studies* 35(9), 1437–1460
- Milligan, G. W. (1996), Clustering validation: Results and implications for applied analyses, in Arabie, P., Hubert, L. J. and De Soete, G. Eds., *Clustering and Classification*, Singapore, World Scientific.
- ONS (2012) 2011 Census, Population and Household Estimates for Small Areas in England and Wales, 23/11/ 2012. [Online] [http://www.ons.gov.uk/ons/dcp171778\\_288463.pdf](http://www.ons.gov.uk/ons/dcp171778_288463.pdf) (Accessed 19th November 2014)
- Rencher, A. C. (1998) *Multivariate statistical inference and applications*. New York: Wiley.
- Simpson, L. (2013) What makes ethnic group populations grow? Age structures and immigration. *Dynamics of Diversity: Evidence from the 2011 Census*. Centre on Dynamics of Ethnicity [Online] <http://www.ethnicity.ac.uk/medialibrary/briefings/dynamicsofdiversity/what-makes-ethnic-group-populations-grow-age-structures-and-immigration.pdf> (Accessed 11th November 2014)
- Stillwell, J and Hussain, S. (2010) Internal migration of ethnic groups in Britain, Chapter 5 in Stillwell, J., Finney, N. and Van Ham, M. (eds.) *Understanding Population Trends and Processes Volume 3: Ethnicity and Integration*, Springer, Dordrecht
- Uskul, A. K. and Platt, L (2014) A note on maintenance of ethnic origin diet and healthy eating in *Understanding Society - Institute for Social and Economic Research (ISER)*. Working Paper. [Online] <https://www.iser.essex.ac.uk/publications/working-papers/iser/2014-03> (Accessed 9th November 2014)

Vickers, D. and Rees, P. (2007). Creating the UK National Statistics 2001 Output Area Classification.  
Journal of the Royal Statistical Society: Series A (Statistics in Society) 170(2): 379-403

# Towards a Seamless World Names Database

Leak A<sup>\*1</sup>, Longley P<sup>†2</sup> and Adnan M<sup>‡2</sup>

<sup>1</sup>Department of Security and Crime Science, UCL

<sup>2</sup>Department of Geography, UCL

April 8, 2015

## Summary

This paper sets out to address limitations in a global database of personal names (worldnames.publicprofler.org: WND). A synthesis of 26 publicly available electoral register and telephone directory datasets. However, it has proven difficult to source further data that are representative of some other countries resident populations. Thus, this study seeks to evaluate the potential for proxy registers based on geotagged Twitter data and further, to devise a method for evaluating the datas' quality. The paper concludes with a discussion of the problems arising where there are no registers or directories against which population registers or directories might be compared.

**KEYWORDS:** GIS, Social Media Analysis, Surnames, Twitter.

## 1 Introduction

This paper sets out to address limitations in a global database of personal names (worldnames.publicprofler.org: WND) which is representative of approximately 2 billion of the Earth's population. A synthesis of electoral roll and telephone directory data from 26 countries, the WND provides a valuable resource in the analysis of populations (Longley et al., 2011; Mateos et al., 2007). However, it has proven difficult to source further data that are representative of many countries' resident populations, limiting the ability to perform truly global analyses. Thus, this paper seeks to assess the use of proxy population registers based on geotagged Twitter data. To this end, a framework for the creation of Twitter based population registers will be demonstrated and subsequently a method for their verification. The paper will conclude with a proposal as to how such registers may be verified in the absence of any reference data.

---

\*a.leak.11@ucl.ac.uk

†p.longley@ucl.ac.uk

‡m.adnan@ucl.ac.uk

## **2 Methods**

This section describes the methods employed in the construction of a Twitter based population registers comparable in structure to that of the UK Enhanced Electoral Roll (id, forename, surname and address). The method is applied to three countries, Poland, Spain and the United Kingdom such that three of the largest naming regions are represented. Each register is subsequently validated against a reference dataset drawn from the existing WND.

### **2.1 Data**

#### **2.1.1 Twitter**

The Twitter dataset comprised 1.4 billion geotagged tweets harvested using the Twitter sample stream API during the period December 2012 through January 2014. It is important to note that whilst the sample stream is commonly cited as being just 1% of all tweets, this figure is in reference to the total number of tweets which may be returned. As such, where only the geographically referenced content are specified (circa 1% of all tweets) the majority are returned (Morstatter et al., 2013).

### **2.2 Base population registers and geography**

The three baseline registers used were the 2011 UK Enhanced Electoral Roll (EER), the 2004 Spanish telephone directory and the 2005 Polish telephone directory. These data account for 53 million, 10.4 million and 8.2 million individuals for the UK, Spain and Poland accounting for 82%, 22% and 21% of the countries' populations respectively. In each case the data are referenced to GADM levels 0,1 and 2. Whilst the GADM boundaries lack the precision of many national administrative datasets, they have a consistent data structure.

#### **2.2.1 Estimation of residential location**

Prior to the extraction of users' personal names, the users believed to be resident within the study area were identified. For this exercise, the location information recorded in the users' tweets was used in preference to their declared locations. The assumption being made that the users are resident in the administrative areas in which they tweet most frequently. The process is as follows:

1. All unique users to tweet within the country boundary are identified.
2. All tweets, regardless of location, by the previously identified users are extracted from the primary dataset.
3. Extracted tweets are spatially joined to GADM administrative geography.

4. Users are ascribed a location where they have 5 or more tweets and greater than 50% of all their tweets within a single spatial unit.
5. Only those users who are assigned a location within the country are included in the final register.
6. Users who do not meet the inclusion criteria are omitted.

The accuracy of the residential location ascription algorithm, assessed in the UK at GADM level 2, was 84% based on a stratified sample of 1073 users.

### **2.2.2 Extraction of personal names**

Having identified those users believed to be resident within the target area, the next phase was the extraction of the users' probable given and family names. Rather than the distinct given and family name attributes found in conventional population registers, the full name is provided as a single string. This string was split using heuristics based on an enhanced version of a name extractor developed by Muhammad Adnan of UCL. The process is as follows:

1. All non-alpha characters are removed from the user's screen name.
2. Screen names are split into multiple tokens based on white spaces.
3. Tokens are compared against a list of titles and surname prefixes.
4. Surname prefixes are affixed to the family name segment.
5. Cleaned names are recorded against the user's id.

### **2.2.3 Population scaling**

Lastly, whilst not critical in terms of population composition, it is necessary to scale the Twitter population for inclusion in the WND. To achieve this, the correct distribution of the countries populations was ascertained based on the best available data and in turn, this data was used to scale up or down the Twitter population such that it accurately represented the resident population distribution.

## **2.3 Assessment of population registers**

Whilst the registers created are accurate in terms of their structure, little is known as to their representative capability. Subsequently, a testing framework is proposed which examines three key properties of the proxy registers; the most common names, population distribution and similarity in composition of surnames.

### 2.3.1 Location quotient

The comparison between observed and expected Twitter population size is performed using Equation 1, the Location Quotient.

$$LQ = \frac{p_i/p}{P_i/P} \quad (1)$$

- $LQ > 1$  indicates proportionally **more** Twitter users than expected.
- $LQ < 1$  indicates proportionally **fewer** Twitter users than expected.
- $p_i$  Local Twitter population count.
- $p$  Local Electoral population count.
- $P_i$  Total Twitter population count.
- $P$  Total Electoral population count.

### 2.3.2 Similarity of composition

The similarity in name composition is measured using Equation 2 the Morisita-Horn index of overlap (Horn, 1966). The index, developed in ecology, is recognised for ability to deal with populations of different sizes and diversities (Wolda, 1981). In this analysis, comparison is made between the resident and Twitter derived population for each administrative unit at each spatial resolution. Where an area had 100 or few individuals in either register, the area was declared null and omitted.

$$C_H = \frac{2 \sum_{i=1}^S x_i y_i}{\left( \frac{\sum_{i=1}^S x_i^2}{X^2} + \frac{\sum_{i=1}^S y_i^2}{Y^2} \right) XY} \quad (2)$$

- $C_H$  1 indicates equal composition of surnames.
- $C_H$  0 indicates no overlap in name composition.
- $S$  is the number of unique surnames shared between the two populations.
- $x_i$  and  $y_i$  are the number of individuals sharing a specific surname in region X and Y.
- $X$  and  $Y$  are the number of unique surnames in regions X and Y respectively.

### 3 Results

UK				Spain				Poland			
Surname	Res.	Tw.	Dif.	Surname	Res.	Tw.	Dif.	Surname	Res.	Tw.	Dif.
Smith	1	1	0	Garcia	1	1	0	Kowal	1	11	-10
Jones	2	2	0	Fernandez	2	6	-4	Nowak	2	2	0
Williams	3	3	0	Gonzalez	3	5	-2	Wojniak	3	9	-6
Brown	4	5	-1	Rodriguez	4	4	0	Kowalczyk	4	9	-5
Taylor	5	4	+1	Lopez	5	2	+3	Wozniak	5	15	-10
Davies	6	6	0	Martinez	6	7	-1	Mazur	6	12	-6
Wilson	7	7	0	Perez	7	8	-1	Kaczmarek	7	11	-4
Evans	8	8	0	Martin	8	10	-2	Krawczyk	8	4	+4
Thomas	9	9	0	Gomez	9	9	0	Zajac	9	-	-
Johnson	10	11	-1	Ruiz	10	11	-1	Krol	10	5	+5

Table 1: Comparison of surname ranks for the 10 most common family names between the resident and Twitter population registers for the UK, Spain and Poland.

UK		Morisita-Horn			Location Quotient			
GADM Level	Valid Locations	Min	Mean	Max	Min	1 <sup>st</sup> Qu.	2 <sup>nd</sup> Qu.	Max
0 n=1	416,819	-	0.98	-	-	-	-	-
1 n=4	413,461	0.91	0.97	0.99	0.94	-	-	1.08
2 n=192	346,900	0.10	0.67	0.97	0.45	0.77	1.08	8.77

Spain		Morisita-Horn			Location Quotient			
GADM Level	Valid Locations	Min	Mean	Max	Min	1 <sup>st</sup> Qu.	2 <sup>nd</sup> Qu.	Max
0 n=1	261,131	-	0.91	-	-	-	-	-
1 n=18	248,190	0.68	0.86	0.94	0.36	0.82	1.22	2.14
2 n=51	241,706	0.57	0.82	0.94	0.33	0.68	1.32	0.28

Poland		Morisita-Horn			Location Quotient			
GADM Level	Valid Locations	Min	Mean	Max	Min	1 <sup>st</sup> Qu.	2 <sup>nd</sup> Qu.	Max
0 n=1	13,178	-	0.37	-	-	-	-	-
1 n=16	12,721	0.00	0.04	0.16	0.38	0.49	1.38	2.75

Table 2: Results of the Location Quotient and Morisita-Horn Similarity Analysis for the UK, Spain and Poland.



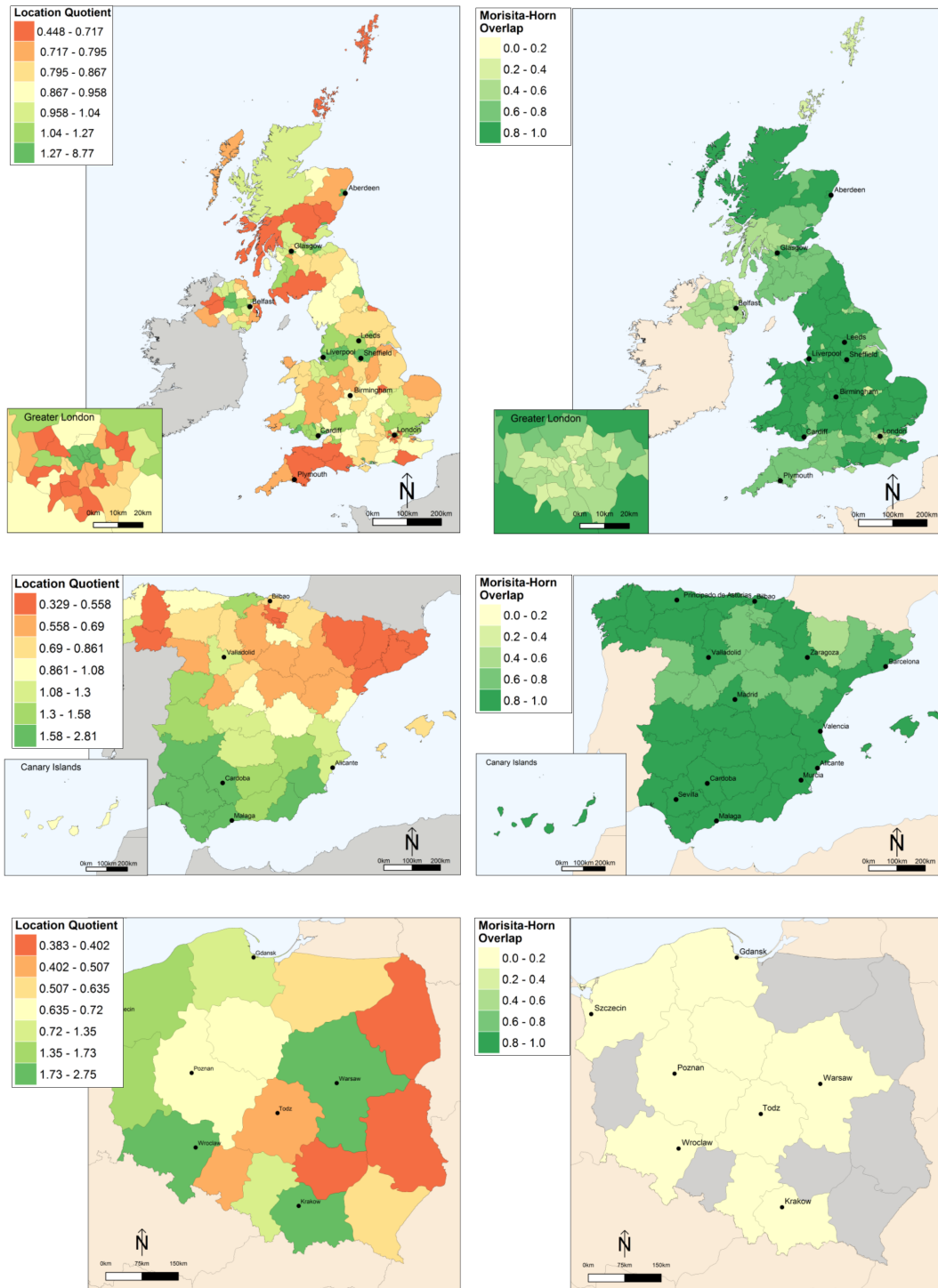


Figure 1: Location Quotient (left) and Morisita-Horn Similarity (right) maps for the UK, Spain and Poland at GADM levels 2, 2 and 1 respectively

## 4 Discussion

The results presented in this report provide a number of early indicators as to the potential utility of social media based population registers. The performance in both the UK and Spain was impressive at both national and regional scales. However, the results for Poland highlighted several potential limitations of the approach. Unlike its counterparts, Poland exhibited poor surname composition similarity at all spatial resolutions. Table 1 provides an early indicator as to the success of the register creation methodology. However, whilst strong rank similarity may be observed in the UK and Spain, this is not reflected for Poland. The disappointing performance in Poland is further observed in Table 2 which provides a breakdown of the Morisita-Horn and Location Quotient results. Of particular significance is the low national surname similarity observed in Poland. A further observation was the high Location Quotient value observed in the City of London, UK, which is likely a consequence of the highly transient population. The geographic distribution of Location Quotient and Morisita-Horn analysis are illustrated in Figure 1.

A key concern to arise from the preliminary analysis was as to how a country could have been deemed suitable for a Twitter based population register in the absence of baseline data. That is to say that whilst the UK and Spain may have been well represented by a Twitter based register, Poland most certainly was not; a fact that may not have been evident without some form of reference. Thus, there is a requirement for a series of diagnostic tools which may be applied to the Twitter derived registers in the absence of a reference dataset. To this end, Twitter population registers are to be created for all countries presently in the WND. In turn, a series of diagnostic measures will be performed on each of the proxy register to determine if, and to what extent, they are indicative of the registers representative capacity versus their WND counterparts. These measures will include the proportion of identified users to the true population, surname diversity and population structure as determined by the conformity to power laws. These diagnostic tools may also assist in the selection of an appropriate spatial resolution for each new country to be included in the WND.

## 5 Conclusions

Overall, the study has demonstrated how population registers may be created using geotagged Twitter data and how such registers may be validated. The preliminary analysis provided strong positive evidence to support the methodology though it was unfortunate how poorly Poland performed throughout the analysis. That being said, the results of the Polish analysis have allowed for validation of the register creation method in a Slavic country and helped determine how a country's suitability may be determined in the absence of baseline data. Finally, the work has taken the first steps in creating arguably the most complete record of human population ever created; a resource of significant value in academia.

## 6 Acknowledgements

This work was completed as part of the DSTL National PhD scheme (12/13NatPHD 61) and the EPSRC research Grant The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds (EP/J005266/1).

## 7 Biography

Alistair Leak is a Ph.D. student in the department of Security and Crime Science at University College London. His research interests include the analysis of names and the application of geographical information systems to ‘Big’ and open data.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include 14 books and more than 125 refereed journal articles and book chapters. He is a former co-editor of the journal *Environment and Planning B* and a member of four other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.

Muhammad Adnan is a Senior Research Associate at Consumer Data Research Centre, University College London. His research interests are in data mining, social media analysis, and visualisation of large spatio-temporal databases.

## References

- Horn, H. S. (1966). Measurement of “overlap” in comparative ecological studies. *American naturalist*, pages 419–424.
- Longley, P. A., Cheshire, J. A., and Mateos, P. (2011). Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum*, 42(4):506–516.
- Mateos, P., Webber, R., and Longley, P. (2007). The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names.
- Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from twitters streaming api with twitters firehose. *Proceedings of ICWSM*.
- Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, 50(3):296–302.

# Land-use Simulation at Large-scale using Big Data

Dan Li<sup>\*1</sup>

<sup>1</sup> College of Environmental Science and Tourism, Nanyang Normal University, China

January 4, 2015

## Summary

Land use/cover changes at larger scales have crucial impact on large-scale environmental problems. Cellular Automata (CA) has become a main tool to simulate and predict land use changes. But large-scale land-use change simulation with high-resolution data requires a large amount of data, complicated computing processes, and very long execution time. The data storage size can be hundreds of megabytes or even several gigabytes, we could consider these data as a kind of “Big Data”, therefore these big data also lead to the problems of computational capability. A computing model called GPU-CA model is proposed to use the graphics processing unit (GPU) high-performance technique to execute and accelerate such simulations. The comparison indicates that the GPU-CA model is faster than traditional CA by 30 times. Such improvement is crucial for land-use change simulations at large-scales using big data.

**KEYWORDS:** Land-use Simulation, GPU, Cellular Automata, Large-Scale, Land Use Change

## 1. Introduction

The basic research object of Geospatial System is an open, complex, giant system that composes of natural, social and other elements with multi-scale and other characteristics of complex systems. In past study, the theory of complex systems was introduced to model and interpret the geospatial system. Cellular Automata (CA) is one of the most powerful tools in complex system theory, which has been used to simulate the geographical phenomena, such as urban development (Batty and Xie, 1994), disease expansion, fire spread, and etc. CA has become a main tool to simulate and predict the urban expansion, especially in the field of land use/cover change studies. In the previous study, Xia Li (2009, 2010a) have successfully developed Geographical Simulation and Optimization Systems (GeoSOS) via combining the theories of land-use change simulation and spatial optimization and extending the theory framework and range of application in related research fields.

## 2. Solution for land-use simulation at large-scale with big data

Land-use changes at larger scales (e.g. provincial, national, or even at global scale) have crucial impact on large-scale environmental problems, such as global climate change, food safety, carbon recycling, and so on. When simulating and predicting large-scale land-use changes, selecting proper data and experimental techniques are very important to generate reasonable results.

Raster data are usually used in spatial simulation experiments, thus, selecting proper data precision is considerable. For low-resolution data, a single data pixel represents multiple land-use types, and the composition of these types is expressed as percentages. This may affect the precision of model analysis results. However, for high-resolution data, each pixel can completely represent the dominant

---

\* danl\_163@163.com

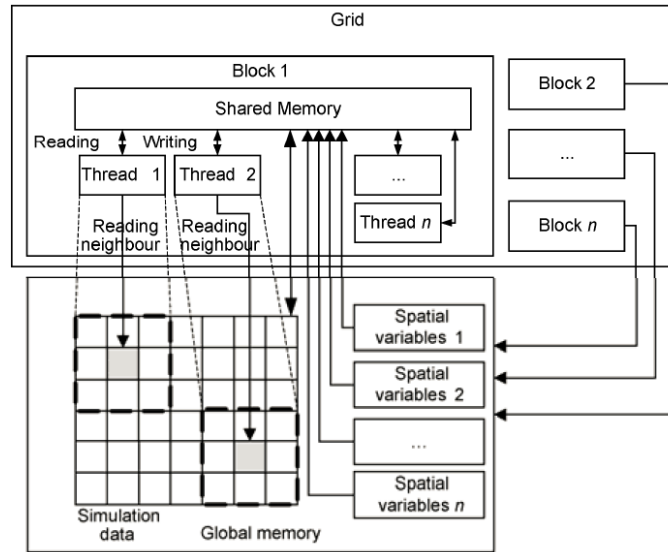
land-use type within the involved space. Moreover, low-resolution data lose local information and non-linear features of geographic patterns, thus using high-resolution data in the simulations can ensure that micro-scale information will not be omitted. CA use a bottom-up approach to explore the emerging behavior of complex systems, it is preferable to use fine-grain data in simulations. Thus, it is essential to apply high-resolution data to large-scale land-use change simulations. However, large-scale land-use change simulation with high-resolution data requires a large amount of data, complicated computing processes, and very long execution time. For example, the data storage size can be hundreds of megabytes or even several gigabytes. We could also consider these data as a kind of “Big Data”, which become more and more important and essential consideration in such large-scale simulations.

But Big Data also leads to the problem of computational efficiency. It is difficult for a PC to perform such simulation experiments because of its limited computing capability and CPU-based serial computing pattern. Parallel computing is an optimal method used to improve computational capability. There have been some studies on parallel computing-based land-use change simulation (add references) e.g., load balancing-based parallel CA simulation, grid computing-based CA simulation (Xia Li et al. 2010b), and others. However, the shortcomings of these methods lie in high computational cost, complex configuration, and lower performance relating to its computing acceleration. Thus, a new parallel computing pattern with low cost, simple configuration, and better acceleration performance is necessary, to provide better computational capability for large-scale land-use change simulations.

The graphics processing unit (GPU) is high-performance technique which can be used to CA simulation. GPUs make use of computer graphics card to execute general-purpose computations (Owens, 2008). It has characteristics of low cost, and a high degree of parallelization, programmability, and flexibility. The “CPU + GPU” computing pattern is a trend representing the future development of high-performance computing techniques. The TIANHE-1 supercomputer is also made by this hybrid computing architecture, which was the top 1 supercomputer in 2010 - 2011. NVIDIA proposed the CUDA (Compute Unified Device Architecture) computing platform in 2007, which provided a GPU-based general-purpose computing environment and software architecture that can be developed using C-like language. With this platform, general-purpose computing tasks can be performed with any CUDA-supported computer graphics cards, so the computational cost is relatively low. Presently, GPU based high-performance computing is extensively applied in fields such as physical simulation, image processing, three-dimensional terrain generation, signal processing, artificial intelligence, and others (Ferrando et al., 2011). There are greatly improves on the computational efficiency comparing the original CPU-based pattern. Geographic problems are generally complex, thus, applying the GPU computing technique to geographic simulations of large-scale and/or high-resolution land-use changes would be of great importance.

### 3. GPU-CA model

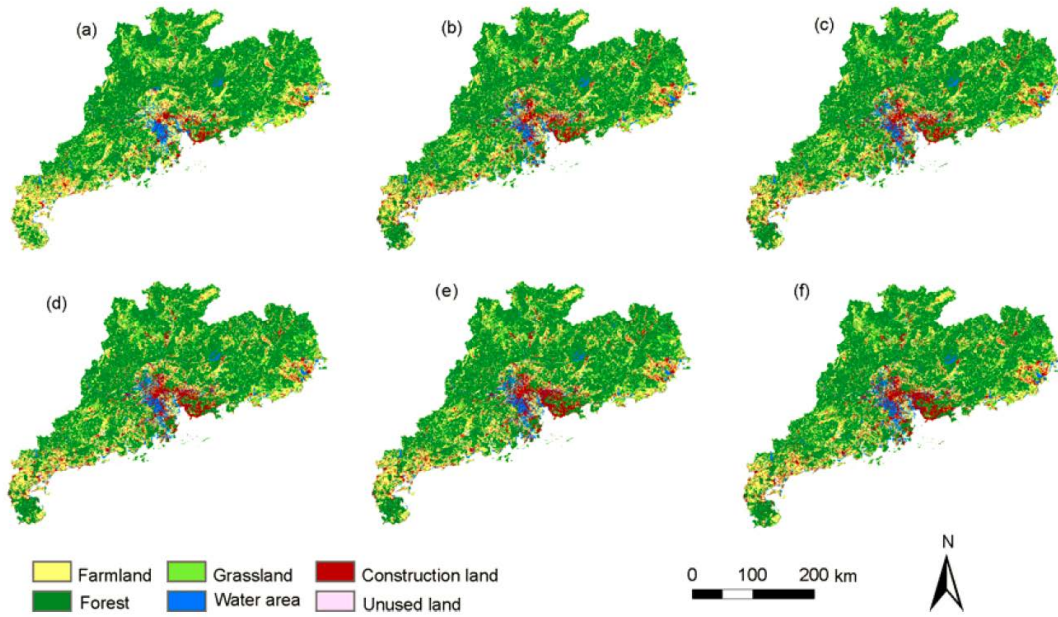
In this study, a computing model called GPU-CA model is proposed to map the CA computation procedure to the GPU programming and memory models. GPU general-purpose computing can be logically divided into three levels. The smallest computing unit is the Thread. Multiple threads compose a Block (one block has one shared memory), and one or more blocks compose a Grid. The GPU-CA computing model uses a data parallel computing pattern, which maps a thread to a cell in the cellular space, and this thread is responsible for the computation of the corresponding cell. The data with relating space variables are read by graphics cards from the computer mainframe (host) memory to global memory. These data are then copied to each shared memory of the blocks, where each thread reads the data according to the cell it maps. Since the CA model requires information about the neighborhood of each cell, each thread simultaneously reads data from the neighbors of the mapped cell. We adopted the 3×3 Moore neighborhoods. When the data reading is complete, a computing iteration is carried out using the multiple thread units, and resulting data are written back to global memory. The next iteration then begins, until the termination condition is reached. Figure 1 shows the GPU-CA computing model architecture.



**Figure 1** GPU-CA computing model architecture

#### 4. Experiments and Results

We used the GPU-CA model to simulate the urban expansion in a rapid urbanization area as the study area, which is Guangdong Province, China, with a land area of 179800 km<sup>2</sup>. The land use classification data of Guangdong Province in 2000, 2005, and 2006 were acquired, as well as related space variable data. The land-use classification data are raster data files in ArcGIS ASCII format, with 80 m spatial resolution and the raster size is 9792×7376. The single data document has a storage size of approximately 500 MB, so the total amount of land-use classification data and space variable data exceeded 3GB, which is excluded the large amount of intermediate result data. There are six land-use types, namely, farmland, forest, grassland, water area, construction land, and unused land. The space variables include distance to city center for each raster cell, and distances to the railway, highway, and roadway. Based on the logistic regression CA model, the land-use classification and space variable data in 2000 and 2005 were used to determine the conversion rules, which define the conversion of non-construction to construction land in Guangdong Province. Subsequently, the land use data of 2005 was used as initial data to simulate the land conversion process in the period 2005–2006. After the simulation result was produced, it was compared with the 2006 land-use classification data using two accuracy assessment approaches to validate the simulation result, which are the point-to-point and naked-eye approaches. The general CA provided by GeoSOS software (GeoSOS, <http://www.geosimulation.cn/>) was used to simulate the same land-use change process, for analyzing the degree of improvement in computational efficiency by GPU-CA. Finally, the construction land conversion process in Guangdong Province during 2010 and 2015 was predicted. Figure 2(a)–(c) describes land use in Guangdong Province in 2000, 2005, and 2006, respectively. Figure 4(d) describes simulation results for 2006, using land use transition patterns extracted from the data of 2000 and 2005. Figure 4(e) and (f) shows predicted results for 2005–2010 and 2005–2015, respectively.



**Figure 2** Diagram showing the simulated results of GPU-CA

To compare with the general CA using the CPU serial computing pattern, we adopted a powerful hardware platform configuration. This platform consists of 2× Intel Xeon E5620 2.5 GHz CPU, 24 GB mainframe memory, Windows 7 64-bit professional edition operating system, and Tesla 1060C graphics card manufactured by NVIDIA Inc., which has 4GB graphic memory and 1.30 GHz core frequency. The GPU computing environment is CUDA 3.2, and the development environment is Microsoft Visual C++ 2008.

By using the point-to-point comparison method which commonly used in image consistency evaluation, a confusion matrix is generated to test the simulation accuracy. The results indicate that the overall accuracy of GPU-CA simulation is 82.9%, indicating that the GPU-CA model is very effective and can be applied to large-scale land-use change simulations.

The concept of speedup is commonly used to evaluate the performance of parallel computation, which uses the ratio of parallel computation time to CPU serial computation time, to represent the acceleration performance of parallel computing. It shows that GPU-CA can remarkably accelerate the CA simulation, with the greatest speedup reaching 33.97. This is a much better improvement on the execution efficiency than the CPU computing pattern, indicating that GPU-CA is very efficient in computation. We also compare the acceleration performances of GPU-CA with the load balancing-based parallel CA. Parallel CA uses a 3×3 neighborhood to simulate land conversion in the Pearl River Delta region, with the computation time approximately 1200–1600s and a speedup around 2. The simulation time using GPU-CA is approximately 15s, and speedup around 18. Therefore, GPU-CA shows much improved computing performance over parallel CA. Moreover, parallel CA requires 8 PCs for parallel computing. GPU-CA only requires one computer with a graphics card supporting the CUDA environment, thus, the computational cost is much lower than parallel CA. This result fully demonstrates that GPU-CA has a low computational cost and high computing capacity.

## 5. Discuss and Conclusion

As described above, GPU-CA model is suitable, efficient and low-cost for large-scale land-use change simulations, which uses big data of geographical information and need powerful computation

capability. Therefore, these are still some issues should be discussed in future. First, CA obtains conversion rules by statistical algorithms, usually the sampling data for statistics are 10% or even less of the whole spatial data. This is proper for small or medium spatial scale studies. But in large-scale land-use simulations, spatial heterogeneity will lead to the sampling data maybe only reflect the global law of the entire study area, but lose the local laws in different spatial areas. So the spatial-division algorithms should be applied to divide the whole study area to different zones according to the different spatial and humanity factors. Second, in present work we read all the computing data to GPU at once, for the amount of these data is not beyond the size of the GPU memory. But in practice, this condition will not be guaranteed all the time, so data-partition method as a more parallel approach should be used in later simulations. That will divide data to regular blocks, and then read some or all blocks to GPU memories at once according to the GPU memory size. This is a flexible framework which can be self-adaptive for different hardware environment. Also parallel computing protocols, such as OpenMP, can be used to construct a computing cluster that composed of multiple GPU graphics cards. It can improve computing efficiency and expand the data processing scale. These are all the subjects of future research.

## 6. Biography

Dan Li, got a doctoral degree of geographical information science at Sun Yat-sen university in 2011. Mainly interested in urban development simulation, high-performance geographical computing, GIS software development.

## References

- Batty M and Xie Y (1994). From cells to cities. *Environment and Planning B-Plan Design*, 21(7), 531–548.
- Li X, Liu X P, He J Q, et al. (2009). A geographical simulation and optimization system based on coupling strategies (in Chinese). *Acta Geographical Science*, 64(8), 1009–1018.
- Li X, Li D, Liu X P, et al. (2010a). The implementation and application of geographical simulation and optimization systems (GeoSOS) (in Chinese). *Acta Sinientiarum Natralium University Sunyatseni*, 49(4), 1–5.
- Li X, Zhang X H, Yeh A G et al. (2010b). Parallel cellular automata for large-scale urban simulation using load-balancing techniques. *International Journal of Geographical Information Science*, 24(6), 803–820.
- Owens J D, Houston M, Luebke D, et al. GPU Computing. *Proceedings of IEEE*, 2008, 879–899.
- Ferrando N, Gosalvez M A, Cerda J (2011). Octree-based, GPU implementation of a continuous cellular automaton for the simulation of complex, evolving surfaces. *Computer Physical Communication*, 182(3), 628–640.



# UK Internal Migration by Ethnicity

Nik Lomax<sup>\*1</sup> and Philip Rees<sup>†1</sup>

<sup>1</sup>School of Geography, University of Leeds, UK

March 2015

## Summary

Migration is a key component of population change for local authorities in the United Kingdom (UK). This paper assesses internal migration during the first decade of the 2000s, disaggregated by ethnic group, drawing upon data reported in the 2001 and 2011 Censuses and a time series of migration data for years between these censuses estimated by Lomax (2013). The patterns, trends and changes for the decade are identified and mapped and presented alongside an interpretation and discussion. These internal migration estimates are one component of a wider project tasked with projecting ethnic group populations in the UK (entitled NewETHPOP) and a brief summary of this project and its proposed outcomes will be offered.

**KEYWORDS:** Ethnicity; Migration; United Kingdom; Population Projection

## 1. Introduction

This paper forms part of a series of work which discusses and analyses the inputs to a model for projecting ethnic group populations in the UK entitled NewETHPOP. Migration is a key component of population change for local authorities in the United Kingdom (UK), so understanding patterns and propensities by ethnic group is crucial for accurate construction and calibration of the model. Information about the original ETHPOP model can be found in Rees *et al.* (2011; 2012). The paper proceeds as follows: the next section outlines the required migration by ethnic group input data for NewETHPOP; it then provides an overview of some patterns and propensities for migration by ethnic group which serve to highlight the importance of gaining an accurate understanding of these patterns; finally some conclusions are offered.

## 2. Migration inputs for NewETHPOP

The desired input data specification for the NewETHPOP projection is in- and out-migration for each one of the 406 local authorities in the UK by single year of age (0 ... 101+), gender (Male, Female) and ethnic group (1 ... 12). Table 1 outlines the ethnic groups used for this project. These ages, gender and the ethnic groups will be consistent with the fertility and mortality components of the model.

Because these data are not consistently available, they will need to be estimated for most areas. To provide a consistent measure, we choose as our leading indicator the Gross Migration Rate (GMR) which allows us to draw the best available data from both national and small area tables. GMR is the sum over all ages of the age specific migration rate (ASMR) which will be estimated for both in- and out-migration for each local authority. GMR measures the expected number of moves that an individual will make in their lifetime, assuming that the individual survives to the oldest age group (Boyle *et al.* 1998). The desired schedule, adapted from Pandit (1997):

---

<sup>\*</sup> n.m.lomax@leeds.ac.uk

<sup>†</sup> p.h.rees@leeds.ac.uk

$$GMR_{ge} = \sum_{xge} M_{xge} I_{xge} \quad (1)$$

Where  $M_{xge}$  is the migration rate for age group  $x$ , gender  $g$  and ethnicity  $e$  and  $I_{xge}$  is the number of years in group  $xge$ .

**Table 1** The required ethnic groups for analysis

Code	Groups included
WBI	White: British, Irish, Gypsy, Irish Traveller
WHO	White: Other White
MIX	Mixed/Multiple Ethnic Groups
IND	Asian/Asian British: Indian
PAK	Asian/Asian British: Pakistani
BAN	Asian/Asian British: Bangladeshi
CHI	Asian/Asian British: Chinese
OAS	Asian/Asian British: Other Asian
BLA	Black/Black British: African
BLC	Black/Black British: Caribbean
OBL	Black/Black British: Other Black
OTH	Other Ethnic Group

Aside from the model input, for further analysis of the time series, the flow between origin and destination (OD) is required. For this, three datasets will be drawn upon: the OD tables reported in the 2001 Census, the 2011 Census and a time series of OD data estimated by Lomax *et al.* (2013) by age and sex. The addition of the ethnic dimension will allow for substantial analysis to be undertaken on the patterns and propensities of migration. The method used will be Iterative Proportional Fitting for those years where the Origin-Destination data are not available, but the total in- and out-migration by age, sex and ethnicity for each local authority are. See Lomax (2013) for a comprehensive discussion of the requirements for estimating missing data in a matrix using IPF. The following section outlines some initial results which take advantage of the recently released 2011 Census Special Migration Statistics for England and Wales, disaggregated by ethnic group.

### 3. Ethnic migration: patterns and propensities

Using the 2011 Census, we can derive a range of indicators for migration by ethnic group. This provides us with an understanding of the migration patterns and propensities for our 12 ethnic groups which will inform decisions used in the projections. The propensities and patterns for the groups are very different so this section demonstrates that if the migration component were treated the same for all ethnic groups we would lose a lot of detail (especially as migration is such an important component of change).

We present three summary measures in this section, which reveal that migration by ethnic group varies considerably. Turnover and Churn are measures utilised by Dennett and Stillwell (2008) and provide an overview of the level of stability or instability in an area. High turnover and Churn indicate higher instability. Turnover can be specified as:

$$TO_i^e = \left( \frac{D_i^e + O_i^e}{P_i^e} \right) 1,000 \quad (2)$$

Where  $TO_i^e$  is turnover for a given area by ethnic group,  $D_i^e$  is inflow for that group to area  $i$ ,  $O_i^e$  is outflow for that ethnic group in area  $i$  and  $P_i^e$  is the population for that area group in the area.

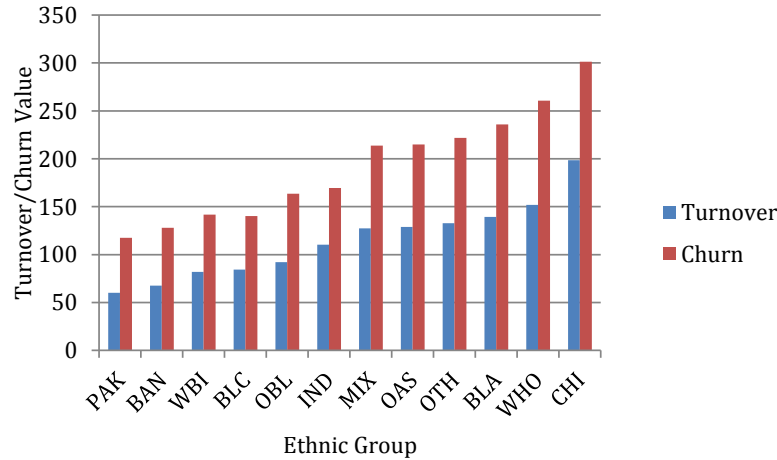
Churn ( $CH_i^e$ ) is similar, but with the addition of moves within area ( $W_i^e$ ):

$$CH_i^e = \left( \frac{D_i^e + O_i^e + W_i^e}{P_i^e} \right) 1,000 \quad (3)$$

Finally, Crude Migration Intensity ( $CMI$ ) provides a simple but effective measure of the proportion of the population at risk for each ethnic group ( $P^e$ ) who migrate ( $M^e$ ):

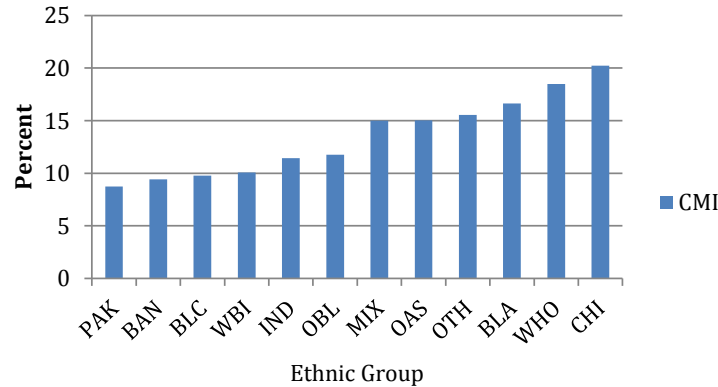
$$CMI = \left( \frac{M^e}{P^e} \right) 100 \quad (4)$$

All three measures are reported here as aggregate (whole country) measures but Turnover and Churn are available by individual local authority. Figures 1 and 2 demonstrate that for the year preceding the 2011 Census, there are clear differences by ethnic group in the rates of Churn, Turnover and CMI. Figure 1 demonstrates that both Churn and Turnover are highest for the Chinese ethnic group, while the lowest values can be seen for the Pakistani and Bangladeshi groups.



**Figure 1** Aggregate Turnover and Churn for each Ethnic Group

When CMI is considered in Figure 2, we can conclude that the Chinese are the most mobile (over 20 per cent of people in this group migrated in the year before the census), closely followed by the White Other group (at 18.5 per cent). The least mobile are the Pakistani, Bangladeshi, Black Caribbean and White British groups, where fewer than 10 per cent of the population migrated during the census period.

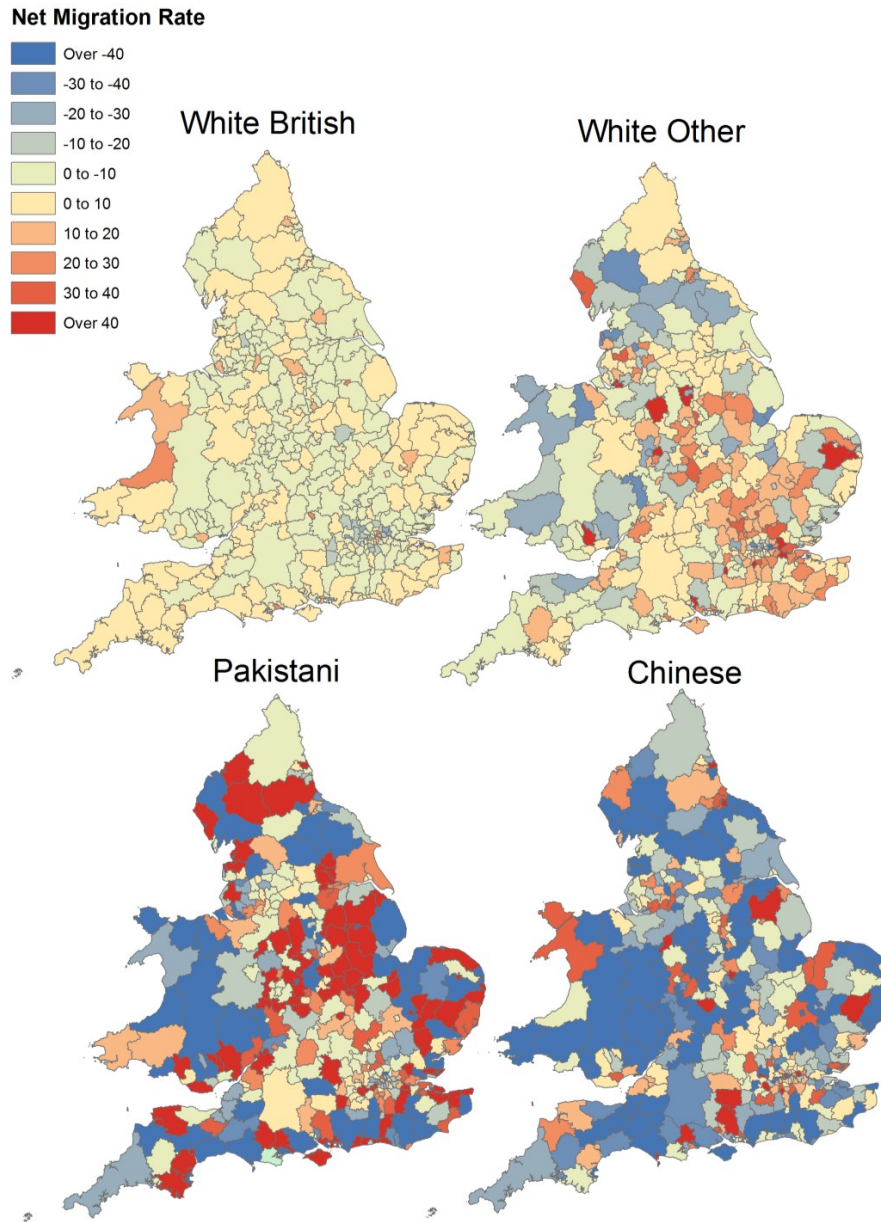


**Figure 2** Crude Migration Intensity for each ethnic group

These aggregate patterns mask substantial variation at LAD scale however, which is why it is so important to consider each ethnic group at an appropriate spatial disaggregation. Figure 3 presents the net migration rate (NMR) for four ethnic groups which are at different ends of the scale for CMI, Churn and Turnover as identified above.

$$NMR = \left( \frac{D_i^e - O_i^e}{P_i^e} \right) 1,000 \quad (5)$$

Where  $D_i^e$  is the total in-migration to area  $i$  for ethnic group  $e$ ,  $O_i^e$  is the total out-migration and  $P_i^e$  is the total population for that ethnic group in the area. It is expressed as a rate per 1,000 people.



**Figure 3** Net migration rates for four ethnic groups in 2011

Clear variances can be seen in Figure 3, both in terms of the values presented and the distribution of net gain and net loss across England and Wales. NMRs for the White British group are lower than the other groups, with evidence of population loss from London (where the majority of Boroughs have a rate of -20 to -30 per 1,000 people). The White Other group demonstrates similar net loss in London, but there are substantial gains to those areas surrounding the capital and to some local authorities on the south east coast. The Pakistani group shows some substantial net gain in some rural areas, especially Lincolnshire, North Derbyshire, East Yorkshire and the North of England. The Chinese group show net gains in a number of London Boroughs, and in the North East, with substantial net loss from a large number of local authorities.

There is apparently some difference in NMR for each group by size of local authority, so this final analysis assesses the relationship between Turnover, Churn and population density. Density is often used as a proxy for the level of urbanness or rurality for an area (e.g. Lomax *et al.*, 2014).

**Table 2** The correlation between population density, churn and turnover for each ethnic group

Ethnic Group	Turnover	Churn
WBI	.656**	.654**
WHO	.375**	.447**
MIX	.000	.078
IND	-.021	-.003
PAK	-.205**	-.177**
BAN	-.147**	-.137*
CHI	.321**	.369**
OAS	.070	.100
BLA	-.425**	-.385**
BLC	-.328**	-.279**
OBL	-.317**	-.289**
OTH	-.175**	-.100

\*\* . Correlation is significant at the 0.01 level (2-tailed).

\* . Correlation is significant at the 0.05 level (2-tailed).

n = 326

Table 2 demonstrates that there are differences when Turnover, Churn and population density are compared, although not all results are significant. The comparison takes in to account in-, out- and within-area migration so provides a measure of stability or instability for an area's ethnic population. All black groups show a significant negative correlation between population density, Turnover and Churn suggesting that the population is less stable in rural areas for these groups. This pattern is also true for the Pakistani and Bangladeshi groups. The White British, White Other and Chinese groups demonstrate a positive correlation, suggesting that there is a higher rate of churn and turnover in urban areas than in rural areas. These results may be influenced by the fact that some more rural areas have low numbers of people in certain ethnic groups, so these populations are by default fairly stable.

#### 4. Conclusions

This paper has outlined the migration component of a larger ethnic group population projection model entitled NewETHPOP. It has specified the required inputs to the model, made a case for understanding the patterns and propensities of migration disaggregated by ethnic group and presented some results from the recently released 2011 Census migration data by ethnic group. These results show that there is considerable variance in migration patterns and propensities by ethnic group, especially where the local authority scale is considered. There also appears to be some difference where local authorities are considered on an urban-rural continuum. Further data are being processed to produce a consistent time series estimate of migration by ethnic group which will be disaggregated by sex and by single year of age.

## 5. Acknowledgements

The work producing the illustrative ethnic projections was funded by the ESRC (Grant Ref ES/L013878/1, 2015-2016) as part of project *Evaluation, Revision and Extension of Ethnic Population Projections – NewETHPOP*.

## 6. Biography

Nik Lomax is a lecturer in Population and Migration. His research focuses on the demographic composition of local areas, which influences policy and resource allocation decisions. This work incorporates measurement and estimation of migration, births and deaths as well as assessment of how these patterns change over time.

Philip Rees is Emeritus Professor of Population Geography at the University of Leeds, with interests in ethnic population projections, health outcomes and ageing of the population.

## References

- Boyle, P., Halfacree, K. H., & Robinson, V. (1998). *Exploring contemporary migration*, Harlow : Longman
- Lomax, N (2013) Internal and cross-border migration in the United Kingdom: harmonising, estimating and analysing a decade of flow data. PhD thesis, University of Leeds. <http://etheses.whiterose.ac.uk/5839/>
- Lomax, N., Norman, P., Rees, P. and Stillwell, J. (2013) Subnational migration in the United Kingdom: producing a consistent time series using a combination of available data and estimates. *Journal of Population Research*, 30(3): 265-288. DOI 10.1007/s12546-013-9115-z
- Lomax, N., Stillwell, J., Norman, P. and Rees, P. (2014) Internal migration in the United Kingdom: analysis of an estimated inter-district time series, 2001-2011. *Applied Spatial Analysis and Policy*, 7(1): 25-45. DOI 10.1007/s12061-013-9098-3
- Pandit, K. (1997). Demographic Cycle Effects on Migration Timing and the Delayed Mobility Phenomenon. *Geographical Analysis*, 29(3), pp.187-199.
- Rees P., Wohland P., Norman P. and Boden P. (2011) A local analysis of ethnic group population trends and projections for the UK. *Journal of Population Research*, 28(2-3): 149-184. DOI: 10.1007/s12546-011-9047-4.
- Rees, P., Wohland, P. and Norman, P. (2012) The demographic drivers of future ethnic group populations for UK local areas 2001–2051. *Geographical Journal*, 179(1): 44-60. DOI: 10.1111/j.1475-4959.2012.00471.x

# Mapping Interactive Behaviour in Wildlife from GPS Tracking Data

Jed A. Long <sup>\*1</sup>

<sup>1</sup>Department of Geography & Sustainable Development, University of St Andrews

November 6, 2014

## Summary

Wildlife researchers now routinely collect detailed data on animal movement using GPS tracking. Methods for studying interactive (e.g., social) behaviour in tracked animals remain limited. I propose three new methods for *mapping* interactive behaviour from GPS tracking data, drawing on fundamental geographical concepts, most notably Hägerstrand's time geography. I demonstrate each method on simulated data and will use examples from my research on white-tailed deer tracked with GPS collars to further exemplify each method. My analysis suggests that how interaction is represented in a GIS leads to different interpretations of wildlife behaviour, but also unique opportunities for further spatial analysis. Open-source software (in R) is provided for other researchers wishing to implement the proposed methods.

**KEYWORDS:** dynamic interaction, movement, time geography, wildlife telemetry, spatial-temporal analysis

## 1 Introduction

The study of wildlife movement ecology has been enhanced by the development of sophisticated tracking devices, most notably those utilizing GPS (Cagnacci et al., 2010; Tomkiewicz et al., 2010). Modern wildlife tracking studies now frequently collect data on multiple individuals being tracked simultaneously, with increasing spatial and temporal resolutions. These advances are facilitating new research questions relating to joint movement behaviour – often termed dynamic interaction (Kernohan et al., 2001). While many methods exist for studying complex spatial-temporal patterns in individual-level movement, methods for studying joint movement behaviour in animal tracking data remain limited in both scope and sophistication (Long et al., 2014).

Studying interactive behaviour is important to many areas of wildlife ecology (e.g., the spread of disease). Typically methods for studying interactive behaviour from wildlife tracking data have simply focused on testing whether or not interaction exists (Doncaster, 1990; Kenward et al., 1993). Moving beyond tests for the presence of interaction, researchers strive to associate interactive behaviour with underlying geographic variables. Thus, new methods for mapping where wildlife interactions

---

<sup>\*</sup>jed.long@st-andrews.ac.uk

occur across the landscape are essential to discovering relationships between interactive behaviour and spatially-heterogeneous geographic variables (e.g., landcover).

## 2 Methods

Here I propose three methods for mapping interactive behaviour from wildlife tracking data that result in three different GIS representations: point, path, or polygon.

### 2.1 Contact Points

Consider two tracking datasets  $A$  and  $B$  each comprising of GPS fixes recorded at discrete times. Two fixes ( $a_i$  and  $b_j$ ) are considered simultaneous if they are recorded at times within a pre-defined critical temporal threshold ( $t_c$ ) of each other ( $|i - j| < t_c$ ). Two fixes ( $a_i$  and  $b_j$ ) are considered proximal if they are located within a pre-defined spatial distance ( $d_c$ ) of each other ( $\|a, b\| < d_c$ ). A contact is defined as occurring when two fixes are both temporally simultaneous and spatially proximal. To map the contact point, we first define the contact vector ( $C$ ) connecting the two contact fixes and identify the mid-point of this vector, which we define as the *contact point* ( $c_k$ ; Figure 1).

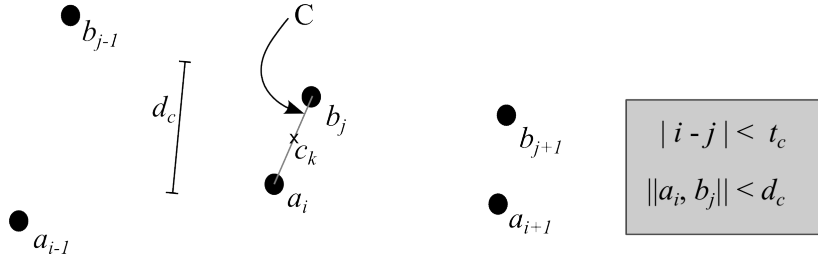


Figure 1: Contacts are defined when two fixes  $a_i$  and  $b_j$  are both temporally simultaneous ( $|i - j| < t_c$ ) and spatially proximal ( $\|a, b\| < d_c$ ). A contact point ( $c_k$ ) is then defined by the mid-point of a contact vector ( $C$ ) and contact points can be mapped across the study area.

### 2.2 Interaction Paths

I extend the contact point method from 2.1 to continuous time in order to map *interaction paths*. For any time point  $\tau$ ,  $a_\tau$  ( $b_\tau$ ) is an interpolated estimate of the location of animal  $A$  (resp.  $B$ ) along its movement path. From two location estimates, compute whether a contact point ( $c_\tau$ ) occurs at time  $\tau$  identically to the method from 2.1. If a contact point occurs at  $\tau$ , we add this point to the interpolation path, if it does not, we move to the next  $\tau$  (Figure 2). Consecutive periods of interaction behaviour are stored as separate lines within the interaction path, representing different periods of interactive behaviour.



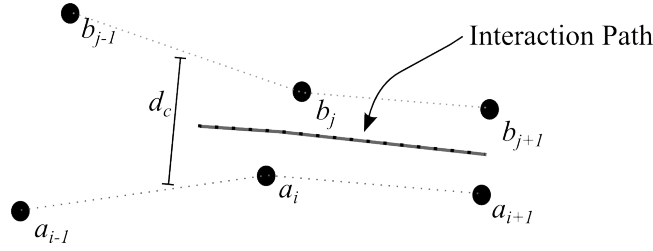


Figure 2: Interaction paths extend contact points to continuous time by interpolating animal locations along their trajectory. Interaction paths represent independent periods of interactive behaviour as separate paths (lines) on the map.

### 2.3 Social Interaction Spaces - Joint Potential Path Area (jPPA)

Drawing from existing movement theory from Hägerstrand's (1970) time geography, I use space-time prisms in order to delineate the social interaction space (Farber et al., 2013) of any two animals. *Social interaction spaces* are defined by the intersection of two (or more) individual space-time prisms (Figure 3), which delineate the movement opportunity space of an individual based on known movement locations (i.e., GPS tracking fixes) and an upper bound on mobility, termed  $v_{max}$ . Individual space-time prisms can be estimated from GPS tracking using the rigorous mathematical definitions from Miller (2005) and have been applied to wildlife previously in order to estimate home ranges (Long and Nelson, 2012). The social interaction space can be delineated simply by intersecting the space-time prisms from Long and Nelson (2012). Projecting the social interaction space onto the geographical plane results in a spatial measure of joint movement opportunity – the *joint potential path area* (jPPA; Figure 3b).

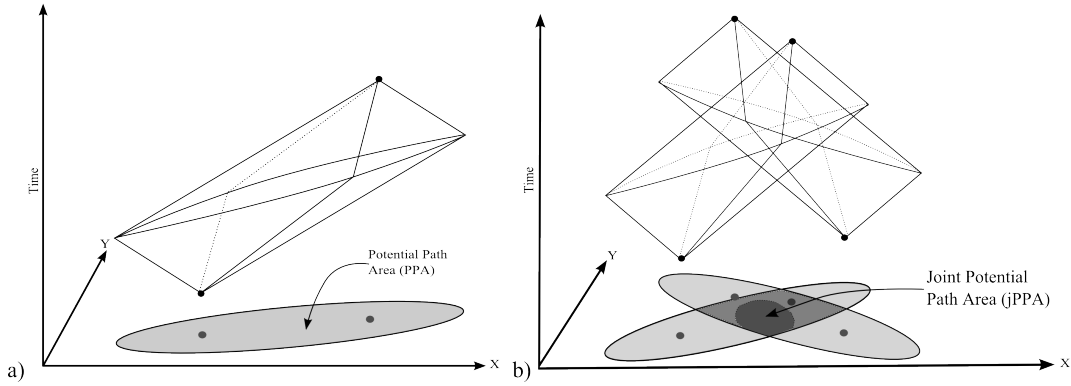


Figure 3: a) Space-time prism, between two known fixes, along with the potential path area (PPA) the projection of the space-time prism onto the geographical plane. b) Intersection of two space-time prisms representing the social interaction space. The projection of the social interaction space onto the geographical plane is a polygon termed the joint potential path area (jPPA).

### 3 Example: Simulated Data

To demonstrate each of the point, path, and polygon-based methods for mapping interactive behavior in wildlife tracking data I use simulated data consisting of two biased correlated random walks (Barton et al., 2009), where the bias in the second individual (Figure 4) is to the location of the first individual (following the procedure used in (Long et al., 2014)). The simulation approach was chosen as it allows for control of factors representing interaction strength and number of interaction episodes. In my presentation, I will also draw on examples from my research examining movement patterns in white-tailed deer tracked via GPS collars (Long et al., 2014).

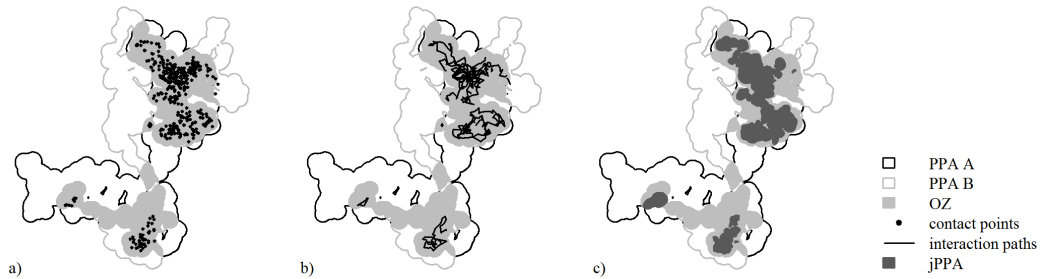


Figure 4: GIS mapping of contacts (using three different approaches) for a simulated dataset consisting of two biased correlated random walks. Mapped alongside the home ranges for each individual, and the home range overlap zone, are the a) contact points, b) interaction paths, c) joint potential path area (jPPA) polygons.

### 4 Discussion

The methods developed here focus explicitly on mapping where interactions occur within the landscape. In the past, methods have focused primarily on identifying only whether interactive behavior is present or absent (Doncaster, 1990; Kenward et al., 1993). Methods that are spatially explicit offer wildlife researchers the potential to explore spatial heterogeneity in interactive behavior across the landscape. Further spatial analysis of mapped contact points, interaction paths, or jPPA polygons will enable linkages between the patterns in interaction and wildlife behavior to be uncovered. For example, contact points can be analyzed as a spatial-temporal point pattern in order to examine spatial temporal clusters (i.e., hot spots) of interaction. Similarly, polygon shape metrics (e.g., number, area, shape complexity) applied to jPPA polygons can provide valuable insight into the behavioral processes generating observed interactions.

By mapping where interactions occur across the landscape researchers can begin to link interactive behavior to widely available datasets describing the landscape (e.g., from remotely sensed data). The spatial associations between mapped landscape variables (e.g., habitat types, topography) are likely to provide further insight into the environmental conditions associated with interactive behavior in wildlife. Further, point, path, and polygon-based measures of interaction can be compared directly to other discrete features existing on the landscape. For example, in many species it will be interesting to examine how interactive behavior is associated with linear features on the landscape (e.g., roads or cut-lines used for natural resource extraction activities, Latham et al., 2011).

The contact point method proposed can be easily extended. For example, attributes associated with the contact points (e.g., the  $\|a, b\|$  distance) could be appended to contacts facilitating more sophisticated spatial analysis of contact point maps (e.g., as a marked point pattern). The interaction path method utilizes a simplistic linear interpolation algorithm from which to estimate the location of the animal along the movement path. While linear interpolation has been utilized widely, both for reasons of ease of implementation and effectiveness, more sophisticated algorithms for interpolating paths (e.g., curvi-linear, Tremblay et al., 2006) could enhance the analysis, especially with certain species (e.g., marine mammals). Finally, the jPPA method is limited in that it maps areas where interaction potentially could have occurred. Incorporating probabilistic models (e.g., Buchin et al., 2012) in order to estimate contact probabilities will further enhance jPPA analysis.

## 5 Conclusion

The growth of wildlife tracking, using GPS, has expanded substantially in recent years and wildlife ecologists are now equipped with incredibly rich data from which to study animal movement patterns (Cagnacci et al., 2010). Objective and quantitative methods for extracting useful movement patterns are required to better understand movement processes and relate these patterns to underlying-contextual information (Purves et al., 2014). Here I provide three straightforward approaches (point, path, and polygon) for mapping interactive behavior as applied to the study of wildlife tracking data. The contact point method provides easy to interpret point-maps of where contacts occur. The interaction-path method identifies areas where sustained interaction periods occur. Finally, the jPPA polygons identify areas of potential interaction that are easily integrated into home range analysis and can be straightforwardly linked to other spatial variables. Each method can be easily computed in a GIS, and I have implemented each as part of the R package `wildlifeDI` (Long et al., 2014) in an effort to facilitate wildlife interaction mapping by other researchers.

## 6 Biography

Jed Long is a Lecturer in GeoInformatics in the Department of Geography & Sustainable Development at the University of St Andrews. His research interests span quantitative geographical analysis and spatial ecology. Much of his work focuses on studying spatial patterns in wildlife movement through the use of GPS tracking data.

## References

- Barton, K. a., Phillips, B. L., Morales, J. M., and Travis, J. M. J. (2009). The evolution of an intelligent dispersal strategy: biased, correlated random walks in patchy landscapes. *Oikos*, 118(2):309–319.
- Buchin, K., Sijben, S., Willems, E. P., and Arseneau, T. J. M. (2012). Detecting Movement Patterns using Brownian Bridges. In *ACM SIGSPATIAL*, pages 119–128, Redondo Beach, CA, USA. ACM Press.
- Cagnacci, F., Boitani, L., Powell, R. A., and Boyce, M. S. (2010). Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. *Philosophical Transactions of the Royal Society B*, 365:2157–2162.
- Doncaster, C. P. (1990). Non-parametric estimates of interaction from radio-tracking data. *Journal of Theoretical Biology*, 143:431–443.
- Farber, S., Neutens, T., Miller, H. J., and Li, X. (2013). The Social Interaction Potential of Metropolitan Regions: A Time-Geographic Measurement Approach Using Joint Accessibility. *Annals of the Association of American Geographers*, 103(3):483–504.
- Hägerstrand, T. (1970). What about people in regional science? *Papers in Regional Science*, 24(1):6–21.
- Kenward, R. E., Marcstrom, V., and Karlbom, M. (1993). Post-nestling behaviour in goshawks, *Accipiter gentilis*: II. Sex differences in sociality and nest-switching. *Animal Behaviour*, 46:371–378.
- Kernohan, B. J., Gitzen, R. A., and Millsaugh, J. J. (2001). Analysis of animal space use and movements. In Millsaugh, Joshua, J. and Marzluff, J. M., editors, *Radio Tracking and Animal Populations*, pages 125–166. Academic Press, New York.
- Latham, A., Latham, M., Boyce, M., and Boutin, S. (2011). Movement responses by wolves to industrial linear features and their effect on woodland caribou in northeastern Alberta. *Ecological Applications*, 21(8):2854–2865.
- Long, J. A. and Nelson, T. A. (2012). Time geography and wildlife home range delineation. *Journal of Wildlife Management*, 76(2):407–413.
- Long, J. A., Nelson, T. A., Webb, S. L., and Gee, K. L. (2014). A critical examination of indices of dynamic interaction for wildlife telemetry studies. *The Journal of Animal Ecology*, 83(5):1216–1233.
- Miller, H. J. (2005). A measurement theory for time geography. *Geographical Analysis*, 37(1):17–45.
- Purves, R. S., Laube, P., Buchin, M., and Speckmann, B. (2014). Moving beyond the point: An agenda for research in movement analysis with real data. *Computers, Environment and Urban Systems*, 47:1–4.

- Tomkiewicz, S. M., Fuller, M. R., Kie, J. G., and Bates, K. K. (2010). Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society B*, 365:2163–2176.
- Tremblay, Y., Shaffer, S., Fowler, S. L., Kuhn, C. E., McDonald, B. I., Weise, M. J., Bost, C.-A., Weimerskirch, H., Crocker, D. E., Goebel, M. E., and Costa, D. P. (2006). Interpolation of animal tracking data in a fluid environment. *Journal of Experimental Biology*, 209(1):128–140.

# Crowd sourced vs centralised data for transport planning: a case study of bicycle path data in the UK

Robin Lovelace<sup>\*1</sup>

<sup>1</sup>School of Geography, University of Leeds

November 13, 2014

## Summary

This paper seeks to test the often mooted hypothesis that distributed, user-contributed 'crowd sourced' GIS data will eventually supercede the traditional centralised geographic data model. The empirical basis used to explore this question is a couple of national-level datasets on a specific topic: bicycle paths in the UK. Open Street Map data represents the crowd-sourced model; Ordnance Survey's Urban Paths layer represents the centralised model. To assess the quality of each dataset, an array of tests was used, from narrow tests of accuracy against aerial photography, to more subjective tests of usability and practical utility. Overall it was found that the OSM data model won on the majority of criteria. However, it must be noted that this is a niche area. If the crowd-sourced data model is to triumph in more mainstream areas it needs to ensure much greater community 'buy-in', for example through compulsory engagement with Open Street Map for educational and citizenship purposes at school.

**KEYWORDS:** volunteered geographic information, Open Street Map, accuracy, bicycle paths.

## 1 Introduction

The relative merits of different geographic datasets has long been a source of academic interest. In the sixth century BC Herodotos, the 'first map maker', criticised the inaccuracies of military maps for (Roller 2010). M. F. Goodchild and Gopal (1989) brought these critiques into the 20<sup>th</sup> century, in an edited compilation of papers on accuracy in spatial data. This and other work has led to the emergence of 'testing spatial accuracy' as a sub-genre within the GIS discipline (e.g. Hunter and Goodchild 1995; Thapa and Bossler 1992; Arnold and Zandbergen 2011).

Throughout the vast majority of Geography's long history, documented geographic information has been created by only a tiny subset of the population. Millitary planners accross all cultures, Roman road builders and, more recently, cartographers armed with specialist and valuable equipment were among the 'geographic 1%' with the priveledge of filling in the blanks in humanity's perception

---

<sup>\*</sup>r.lovelace@leeds.ac.uk

of the world. Since Open Steet Map began freely accepting data submissions from amatures in 2002, we have been living in an era of Volunteered Geographic Information (VGI) (Goodchild 2007). Growth in these new data sources far now far outstrips the incremental and accretive growth in official map source originating from trusted organisations. In fact, the British Ordnance Survey and the US Geological Survey (two of the world’s most respective providers of geographic information), have already considered integrating the insights gathered from an army of ‘citizen sensors’ into their existing systems. The rapidly evolving landscape of VGI from Social Media (VGI-SM), which is becoming increasingly widespread with the penetration of GPS-equipped smart phones in emerging markets worldwide, only adds to the buzz surrounding crowd-sourced datasets.

Though the scope of these introductory comments are broad, the empirical aims and objectives of the study are narrow. The empirical aims of the project are as follows:

1. Describe the cycle path data in Open Street Map and test the hypothesis that it is a good source of data on cycle paths in Great Britain, suitable for academic research.
2. Compare the OSM cycle path data with a proprietary dataset.
3. Describe and explain the spatio-temporal distribution of additions to the the cycle path dataset and investigate how this corresponds with investment in cycle schemes overall.
4. Use the results to discuss the use of crowd-sourced data in mission-critical professional planning applications.

## 2 Data

### 2.1 Open Street Map data

The cycle paths were extracted from the up-to-date and compressed ‘british-isles’ .osm.pbf file (746 Mb), downloaded from OSM services provider Geofabrik.de. This was converted for processing into a raw .osm (14.7 Gb) text file, a variant of the xml filetype, using the command-line program **osmconvert**, accessed through the Linux command line.

To take only navigable paths, the first stage was to subset all *ways*. This led to a 4.9 Gb file containing 3.35 million features. Of course, only a small sample of this large dataset is related to cycling (Fig. x). A careful sampling strategy was therefore used to ensure all bicycle paths, according to OSM data, were captured and whilst minimising data that do not represent bicycle paths. As shown in Fig. x, the majority of the route network in OSM is *potentially* cyclable, with most residential, service and unclassified highways being suitable for bicycles. However, we are not interested in where one *can* cycle in this dataset: the focus is on where local authorities and other organisations have invested in sustainable transport by constructing bicycle paths. The ‘cycleway’ tag is the 11th most commonly used tag, comprising 1.6% of all highway tags in Great Britain (Figure 2).

Tagged cycleways comprise only a small proportion of recommended *cycle routes* in OSM, which are defined using a wide variety of both tags and relations. This means that extracting all cycle routes cannot be acheived with a simple one-line operation. The following command, using the

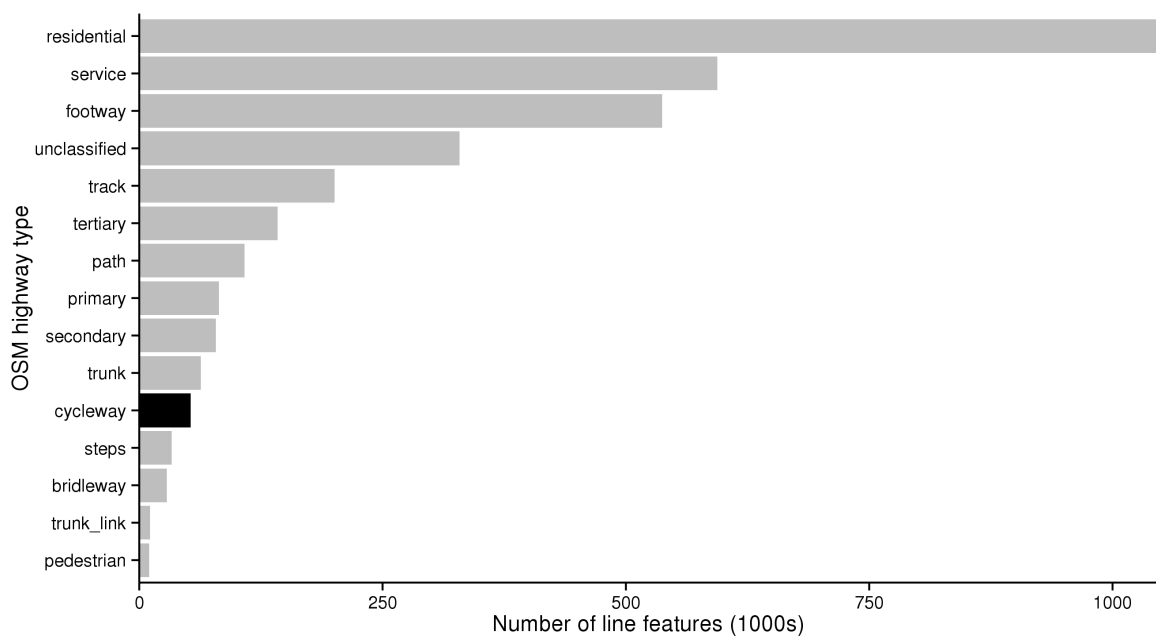


Figure 1: Number of line features by the high level ‘type’ tag in the UK. The only type guaranteed to be a cycle path is ‘cycleway’, although many other ‘types’ can also be tagged as bicycle paths in other ways.



command-line Java program `osmfilter`, for example, will extract only a sample of the total length of cycle paths stored in Open Street Map:

```
osmfilter data.osm --keep="highway=cycleway" >bcways.osm
```

This is because the ‘cycleway’ attribute is only one of the possible *tags* used to represent cycle paths. There are many additional tags from a variety of variables that can but used to identify cycle paths, each with subtly different (and sometimes overlapping) meanings (Table 1).

Table 1: Overview of the most commonly used cycling-related variables and tags in Open Street Map. Note that percentages in bold refer to the proportion of the *network* (tag = \* is a wildcard) while the denominator for the specific tags is the count of non-empty rows *for that variable*. Quoted text is taken from the OSM wiki.

Variable	Tag	Number	Percentage	Meaning
<b>highway=</b>	*	3351300	100.0	
	cycleway	52672	1.6	Dedicated bicycle paths
<b>bicycle=</b>	*	128493	3.8	How permitted are bicycles on the path
	yes	79586	61.9	Bicycle can ride here, not necessarily bike path
	no	24162	18.8	Bicycles are not permitted
	designated	16036	12.5	“Where a way has been specially designated”
	dismount	4015	3.1	You must push your bike by law
<b>cycleway=</b>	*	34085	1.0	“an inherent part of the road”
	no	16472	48.3	No bicycle path
	track	6926	20.3	Cycle path separated from the road
	lane	8831	25.9	On road/shared use bicycle path
	shared	1560	4.6	“Cyclists share space with other traffic here”
	opposite_lane	806	2.4	A ‘contraflow’ cycle lane
	share_busway	502	1.5	Bicycles share space with buses
	opposite	414	1.2	Bicycles can travel either way
<b>lcn=</b>	*	15320	0.5	Local cycle routes
	yes	13193	86.1	Completed
	proposed	1536	10.0	Proposed/under construction
<b>ncn_ref=</b>	*	12684	0.4	Reference of National Cycling Network paths
	1	1179	9.3	NCN route 1
	5	1003	7.9	NCN route5
<b>ncn=</b>	*	3313	0.1	Is part of the National Cycling Network
	yes	2368	71.5	Completed
	proposed	888	26.8	Proposed/under construction
<b>towpath=</b>	*	2795	0.1	The route is along a canal towpath

It is important to note that although the tags presented in figure x are *related* to cycling, they certainly do not all imply the presence of a bicycle path.

Based on a careful reading of the ‘official’ tag description from the OSM wiki page, and visual inspection of the paths overlaying aerial photography, a list of these ‘cycle infrastructure’ tags was compiled. These were:

```
"highway" = 'cycleway' OR "bicycle" = 'designated' OR
"cycleway" = 'track' OR "cycleway" = 'lane' OR
"cycleway" = 'shared' OR "cycleway" = 'opposite_lane' OR
"cycleway" = 'opposite_track' OR "cycleway" = 'segregated' OR
"cycleway" = 'shared_lane' OR "cycleway" = 'yes' OR
"cycleway:left" = 'lane' OR "cycleway:left" = 'track' OR
"cycleway:right" = 'lane' OR "cycleway:right" = 'track' OR
"cycleway:oneside" = 'lane' OR "cycleway:otherside" = 'lane' OR
"path.bicycle" = 'designated'
```

## 2.2 Ordnance Survey data

The Ordnance Survey (OS) data was obtained from their central office following completion of a non-disclosure agreement. Following substantial software challenges, requiring expensive proprietary software, the entire OS dataset was loaded for the UK and could be analysed.

The full attributes of the OS data will be described in the final paper, as will the methods, and results. Preliminary analysis suggests that where OS and OSM datasets do overlap (in a minority of the paths’ lengths for both datasets), the correspondence between them is good, with the OS dataset having higher spatial resolution (Figure 2).



Figure 2: Example of partial correspondence between OSM and OS cycle path data. Note this is not at all representative of the UK.

### 3 Conclusion

This is the first study to our knowledge which decisively finds a free and crowd-sourced dataset to be more appropriate to a critical application than a professionally produced proprietary dataset from a national cartographic organisation. In this case, Open Street Map was found to provide more coverage, attributes, continuity and recency than the commercial dataset provided by the UK's national geospatial data provider Ordnance Survey.

Many future research directions are opened-up by this study, including more systematic and 'real time' tests of different geographical datasets to provide 'health checks' of suitability for different applications; other niche areas where crowd sourced geographical data may be more appropriate than centralised sources; and the potential for national mapping bodies such as Ordnance Survey to incorporate the richness of crowd-sourced offerings into their official products. All of these themes intersect with wider questions about the 'democratisation' of academia (Berry and Moss, 2006) and open up exiting possibilities for combining educational benefits with improved public data administration.

### 4 References

- Arnold, Lisa L., and Paul A. Zandbergen. 2011. "Positional Accuracy of the Wide Area Augmentation System in Consumer-Grade GPS Units." *Computers & Geosciences* 37 (7) (July): 883–892. doi:10.1016/j.cageo.2010.12.011. <http://www.sciencedirect.com/science/article/pii/S0098300411001063><http://www.sciencedirect.com/science/article/pii/S0098300411001063/pdf?md5=a6e3fa02e5aa8a8208c04a9533764729&pid=1-s2.0-S0098300411001063-main.pdf>.
- Berry, D. M., & Moss, G. (2006). Free and open-source software: Opening and democratising e-government's black box. *Information Polity*, 11(1), 21–34.
- Goodchild, Michael F. 2007. "Citizens as Sensors: the World of Volunteered Geography." *GeoJournal* 69 (4) (November): 211–221. doi:10.1007/s10708-007-9111-y. <http://link.springer.com/10.1007/s10708-007-9111-y>.
- Goodchild, Michael F., and Sucharita Gopal. 1989. *The Accuracy Of Spatial Databases*. CRC Press. <http://books.google.co.uk/books?id=HL206J-XtLAC>.
- Hunter, Gary J, and Michael F Goodchild. 1995. "Dealing with Error in a Spatial Database: A Simple Case Study." *Photogrammetric Engineering and Remote Sensing* 61 (5): 529–537.
- Roller, D. 2010. *Eratosthenes' "Geography"*. Princeton University Press. [http://books.google.co.uk/books?id=8peKyWK/\\_SWsC](http://books.google.co.uk/books?id=8peKyWK/_SWsC).
- Thapa, Khagendra, and John Bossler. 1992. "Accuracy of Spatial Data Used in Geographic Information Systems." *Photogrammetric Engineering and Remote Sensing* 58 (6): 835–841.

# Strategies in the Use of Referring Expressions to Describe Things Urban

William Mackaness<sup>1</sup>, Phil Bartie<sup>2</sup> and Philipp Petrenz<sup>1</sup>

<sup>1</sup>The University of Edinburgh, School of Geosciences

<sup>2</sup>University of Stirling, Biological and Environmental Sciences

## Summary

In the context of wayfinding technologies, there is increasing interest in dialogue based systems that use description of landmarks as a way of guiding people through cities. In the absence of maps and photographs, the challenge for automated systems is the production of descriptions of things in the field of view that are unambiguous and easily interpreted. We are therefore interested in the mechanisms used by humans to create and interpret descriptions of things in the urban vista. Here we report on a web based experiment in which we explored the veracity of human generated referring expressions in order to better understand the most successful strategies for directing people's gaze.

**KEYWORDS:** psycholinguistics, referring expressions, surface realisation, wayfinding, urban

## 1. Psycholinguistics and Referring Expressions

Landmarks are one aspect of the environment frequently referenced, as they assist in forming mental representations of space (Hirtle and Heidorn 1993, Tversky 1993), and in way-finding tasks (Werner *et al.* 1997, Lovelace *et al.* 1999, Caduff and Timpf 2008, Winter *et al.* 2008, Duckham *et al.* 2010). Landmarks are defined as identifiable features in an environment, whose saliency may be calculated by comparing scores for particular attributes (e.g. their size) and identifying those which deviate from the mean (Raubal and Winter 2002, Elias 2003a, Elias and Brenner 2004). These are the buildings unlikely to be confused with others, either which appear very different to their surroundings (e.g. churches) or are well known major international brands (e.g. Starbucks). The focus of this paper is not on modelling landmarks. Instead its focus is on i) determining what governs choice of the characteristics that allow landmark identification in 'vista space' (Montello 1993), and 2) how those characteristics are formed into a string of words that constitute a referring expression – a process called 'surface realization' (Jurafsky and Martin 2008).

### 1.1. Referring Expressions and common ground

Referring expressions (RE) can be 1) optimally informative (e.g. 'the small apple' – Figure 1a), 2) under-informative ('the apple') or 3) over/hyper- informative ('the small green apple'). There are various reasons why subjects, when asked to write referring expressions, might create these different forms. They may fail to undertake a full visual scan of the image and fail to see a need to further differentiate. But probably more important than this, the creation of a referring expression depends on a shared conceptualisation between the subject describing the object, and the viewer who interprets and is thus able to locate the object in the scene; this is referred to as 'common ground' (Horton and Keysar 1996). This is very pertinent in the context of the urban (Figure 1b) where reality can be conceptualised (and so described) at very different levels of granularity. Where there is thought to be little common ground or uncertainty between the subject and viewer, we might expect the referring expression to be hyper informative, and therefore contain information that may essentially be redundant.



Figure 1: Simple and complex worlds: a) two apples and b) an urban vista

### 1.2. Influence of Vista, Relatum and Distractors on Referring Expression Generation

The context of this research is the growing interest in dialogue-only based interaction in which the user is both hands free and eyes free to explore the environment as they move through it (Bartie and Mackaness 2006; Mackaness et al. 2014). In a context of only having spoken text as the description, the question arises: ‘how best do we describe urban objects that are in the field of view?’. Too brief and we risk uncertainty; too verbose and the cognitive effort is unnecessarily high. To that end we draw inspiration from Grice who wrote ‘make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange’ (Grice 1975) and from which we can apply the maxims of *quantity*, *quality*, *relation* and *manner* as useful frameworks by which to assess ‘the perfect description’.

The task of referring to something is easy if its uniqueness is readily apparent; ‘distractor’ is the term given to objects of the same class or form, and their presence requires disambiguation (eg the need to distinguish between the two apples in Figure 1a). We see in Figure 1b) there is a large number of distractors. What constitutes a distractor in this urban vista is complex; buildings are of similar shape, colour and size (eg the large beige buildings), or they could be deemed as distractors because they perform a similar function (eg churches). We might therefore envisage that longer RE descriptions are required to disambiguate.

Figure 1b also makes clear that some objects are far more salient than others – unambiguous in their form (eg ‘the castle’). It is common to use such features as an anchor by which we might describe the (less salient) object, and thus differentiate it from its distractors (eg ‘The church immediately right of the castle’). Here the castle acts as a *relatum*. We note that the relatum is acting to reduce the area of search in the scene (the solution space) – provided of course, that the subject and viewer have the same conceptual understanding of castle (common ground). We also note that the larger the vista, the greater the likelihood of distractors. From perusal of the literature in psycholinguistics (the processes by which we understand utterances) (Aitchison 2011) and work in qualitative spatial reasoning (Freska 1991) we can identify the various ways in which referring expressions can be combined both to reduce the solution space, and differentiate between distractors (Table 1). One can readily imagine a large permutation in these choices in order to direct someone’s gaze.

**Table 1** Ways of reducing the solution space or referring to objects in an urban scene

Form of reference	Ambition	Example
Absolute distance	Reducing search	In the <i>far distance</i> ...
Egocentric description (self to object)	Reducing search	To your <i>left</i> you will see..
Allocentric description (object to object)	Reducing search	The library is <i>to the right of</i> the museum
Colour, colour tone, texture, size	Description	The <i>large</i> house with the <i>dark blue</i> door...
Architectural age	Description	..the <i>Victorian</i> looking house
Cardinality	Description	The <i>south</i> facing windows...
Landmark brand	Description	..between <i>MacDonalds</i> and <i>Subway</i>
Landmark type	Description	The <i>library</i> ..., the <i>church</i> ...
Via relatum	Description	Two doors down from the <i>Fire station</i> ....
Relatum type <network, region>	Description	On the right side of the <i>river</i> ..., beyond the <i>park</i> ...
Composition by form	Description	..comprising <i>steps</i> leading to a <i>set of columns</i>
Composition by shape	Description	..large <i>block</i> with a <i>pointy</i> top
(superlative) adjectives, proper nouns	Description	...the <i>taller</i> of the two <i>grand</i> towers on the Houses of Parliament
Topological, container descriptions	Description	..immediately <i>next to</i> the pub, which is <i>in</i> the park

## 2. The Experiment

A web based internet based experiment was set up to identify the most common strategies and types of natural language phrases used to describe the location of an urban feature. We wished to understand 1) how subjects dealt with distractors, 2) reduced the solution space, 3) whether there were any patterns in the ordering or choice of variables, 4) what are the implications in the design of databases necessary to support to automatic generation of RE.

Subjects were asked to write descriptions for each of five images chosen randomly from a choice of 32 images, which varied widely in vista. A press of a button temporarily placed an outline over the building in question. Subjects were asked to provide a text based description sufficient for another person to be able to identify the highlighted feature. In a second phase of the experiment, subjects were presented with a second set of random images together with textual descriptions (provided via previous subjects). Based on the description, subjects were asked to identify the feature (by clicking on the object in the image). Subjects could record if they were uncertain or the description was ambiguous. The images with descriptors were presented to at least three viewers. This provided a means of assessing the veracity of a description. The experiment was first promoted via Crowdfunder but with mixed results. Greater success came from promotion at conference, via Facebook and to subjects involved in a previous experiment (Mackaness et al. 2014). Subjects were incentivised by an opportunity to win Amazon vouchers. The number of participants grew quickly and within a month close to 200 participants provided a total of 800 annotations distributed over the 32 images.

## 3. Results

Figure 2 is an example of an image, with a list beneath of a subset of the descriptions that successfully led viewers to identify the target (shown by the green dots).



Large, modern glass fronted building, butted up against traditional Victorian terrace, slightly set back from road, and with facing bowed frontage.
The target is the Festival Theatre on North Bridge. It is a large glass fronted modern building slightly set-back from the road.
Just look at the first building from the left, the one with really big and nice glass walls.
A large modern building with a totally glass front.
A rather square-shaped, glass walled building, which shows no resemblance to its environment due to its complete different, modern style.
Festival Theatre. The glass-fronted building with obvious posters advertising shows.
Festival Theatre - large glass fronted building with theatre posters in the windows. To the left of Rymans, as you're looking at it.

**Figure 2:** Image with a subset of successful descriptors illustrating the breadth of techniques used.

We observe that most of the descriptors are hyper informative (despite the absence of distractors). For this reason, and contrary to expectations, across the images we could not discern a relationship between vista and length of descriptor. In the case of Figure 2, we surmise that subjects might feel this object is not prototypically ‘theatre’ and this has led to hyper informative descriptions. The issue of granularity is very apparent - subjects utilise detail where it is discernable. Where the object reveals less detail (eg Figure 3), the subject is forced to use alternate strategies.



**Figure 3:** Cropped image of an urban vista showing viewers who selected the object (green) or distractors (red dots)

In Figure 3 the target is a grand building with columns and two distractors. Among the very few successful descriptors was one that anticipated the confusion and likely fixation upon the prominent



central building. Their solution was to direct the gaze away from the distractor (‘Not the first building with the stone pillars but the one behind it’), the other was to use the distractor as a relatum (‘the second building with the columns, the farthest away one’).



**Figure 4:** A monument that is ‘cathedral like’.

Figure 4 is interesting because it has a large solution space and a number of distractors. Of the successful descriptors, 60% used superlatives (dominating, taller, enormous, emblematic). Optimally ‘monument’ proved sufficient. 20% used their knowledge of the city to name the monument. 60% used the trees as a container relatum (eg ‘the spire among the trees’). The term ‘spire’ was used by 50% of the subjects – but ironically this *created* distractors since there are other spires in the image, (whereas it is the only *monument*). This required additional disambiguation and a lengthier descriptor.

### 3.1 Failed Descriptors

A review was also made of descriptions which were deficient. Some were ascribed to poor English, were puerile or directed the viewer to the wrong target. Some used terms that lay outside common ground (eg ‘mansard roof with dormer’, ‘triangular pediment’, ‘neo-classical’). Some used names requiring local knowledge (eg ‘Story telling Centre’), and others linked the description to events or people (eg ‘where J K Rowling writes’). Most frequent were instances where the subject failed to discriminate between the target and other discriminators. In some cases the relatum had distractors which meant the viewer searched in the wrong part of the image (eg ‘second block to the right of the church’, where the image contained two churches).

## 4. Conclusion

This research contributes to research in generating referring expressions (GRE). The web based experiments did not reveal a gender bias, nor a preferred scan direction (vertical or horizontal), nor a relationship between vista (complexity) and length of description. It did reveal how a mix of strategies were used to reduce the solution space, and direct the user’s gaze.

Some might argue that the experiment is of poor design given its multi variate complexity. We would counter that the simple scene analysis typical of psycholinguistic experiments (eg Baltaretu et al 2013; Hanna et al 2003) has little relevance to the generation of RE in urban vistas. We argue that the experiment shows that a critical step prior to surface realisation is scene analysis in which both relatum and distractors are identified. Applying Grice’s norm depends on 1) a deep understanding of the context (objects in the vista), and 2) an understanding of the level of detail discernible to the naked eye. Identifying appropriate relatum requires assessment of its saliency (since relatum may have their own distractors!) Determining whether something is a distractor is fraught with difficulty, since the granularity of the feature will govern the viewer’s perception of whether it is a distractor or not. We would argue that this work is of use as a corpus to the psycholinguistic community.

The implications for a database that supports automatic generation of referring expressions is intriguing. It would appear that multi scale (or multi resolution) representations (Burghardt et al. 2014) are required of each potential candidate object in order to cope with both ‘close up’ RE and ‘vista based’ RE, as well as RE in between!

### Acknowledgements

We are very grateful for EU funding FP7/2007-13 via the SpaceBook project (No 270019). Our thanks too to the participants in the web based experiments.

### Biography

William Mackaness is a senior lecturer at The University of Edinburgh in the School of GeoSciences. Phil Bartie is a lecturer in the Biological and Environmental Sciences at University of Stirling.

### References

- Abella A, Kender J.R. (1999) From images to sentences via spatial relations. Proceedings of the Integration of Speech and Image Understanding.
- Aitchison, J (2011) *The Articulate Mammal: An Introduction to Psycholinguistics*. Routledge, London.
- Bartie, P. and Mackaness, W.A. (2006) Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS* 10: 63-86
- Baltaretu, A.A. Krahmer E.J. Maes, A. (2013) Factors influencing the choice of relatum in referring expressions generation: animacy vs. position. *Proceedings of the CogSci workshop on the production of referring expressions PRE-CogSci 2013*.
- Burghardt, D. Duchêne, C. and Mackaness, W.A. 2014 Abstracting Geographic Information in a Data Rich World Methodologies and Applications of Map Generalisation. Springer.
- Caduff D, Timpf S (2008) On the assessment of landmark salience for human navigation. *Cognitive Processing* 9: 249-267
- Duckham M, Winter S, Robinson M (2010) Including landmarks in routing instructions. *Journal of Location-Based Services* 4 28-52
- Elias B, Brenner C (2004) Automatic generation and application of landmarks in navigation data sets. IN Fisher, PF (Ed.) *Developments in Spatial Data Handling*. Springer, Berlin
- Elsner, M., Rohde, H. Clarke, A.D.F. (2014) Information Structure Prediction for Visual-world Referring Expressions. *EACL2014* April 26-30<sup>th</sup> Gothenburg, Sweden. <http://aclweb.org/anthology/E/E14/E14-1055.pdf>
- Freska, C. (1991) Qualitative Spatial Reasoning in *Cognitive and Linguistic Aspects of Geographic Space*, D.M. Mark & A.U. Frank (eds.), 361-372.
- Grice, P (1975). "Logic and conversation". In Cole, P.; Morgan, J. *Syntax and semantics*. 3: Speech acts. New York: Academic Press. pp. 41–58.
- Hanna, J.E., Tanenhaus, M.K. and Trueswell J.C. 2003 The effects of common ground and perspective on domains of referential interpretation *Journal of Memory and Language* 49: 43-61
- Hirtle SC, Heidorn PB (1993) The structure of cognitive maps: Representations and processes. *Behavior and Environment: Psychological and Geographical Approaches*: 170-192
- Horton, W.S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59, 91-117.
- Jackendoff, R. (1992) *Languages of the Mind*, Cambridge, MA: MIT Press.
- Jurafsky, D. and Martin J.H. (2008) *Speech and Language processing*, Prentice Hall.

- Lovelace K.L., Hegarty M., Montello D.R. (1999) Elements of good route directions in familiar and unfamiliar environments. IN Freksa, C, Mark, D (Eds.) *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*. Springer Berlin / Heidelberg.
- Mackaness, W.A. Bartie, P. Dalmás, T., Janarthanam, S., Lemon, O., Liu X., Webber B., (2014) Talk the Walk and Walk the talk: Design, Implementation and Evaluation of a Spoken Dialogue System for Route Following and City Learning, *Annual Conference of the Association of American Geographers*, Tampa Florida, 7-13 April.
- Mackaness, W.A. Bartie, P. and Sanchez-Rodilla Espeso, C. (2014) Understanding Information Requirements in 'Text only' Pedestrian Wayfinding Systems, *GIScience 2014*, Vienna.
- Montello D (1993) Scale and multiple psychologies of space. *Spatial Information Theory A Theoretical Basis for GIS*: 312-321
- Moratz, R. and Tenbrink, T. (2006) Spatial Reference in Linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations *Spatial Cognition and Computation* 6(1), 63-107.
- Raubal M, and Winter S (2002) Enriching wayfinding instructions with local landmarks IN Egenhofer, MJ, Mark, DM (Eds.) *Second International Conference GIScience*. Springer, Boulder, USA
- Richter, K-F and Winter S. (2014) Landmarks: GIScience for Intelligent Services
- Tversky B (1993) Cognitive maps, cognitive collages, and spatial mental models. IN Frank, AU, Campari, I (Eds.) *Spatial Information Theory: A Theoretical Basis for GIS*. Italy, Springer-Verlag
- Werner S, Krieg-Brückner B, Mallot HA, Schweizer K, Freksa C (1997) Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. IN Jarke, M (Ed.) *Informatik '97 GI Jahrestagung*. Berlin, Heidelberg, New York. Springer
- Winter S, Tomko M, Elias B, Sester M (2008) Landmark hierarchies in context. *Environment and Planning B: Planning and Design* 35: 381 – 398

# Using Mobile Phone Traces to Understand Activity and Mobility in Dakar, Senegal

Ed Manley, Adam Dennett and Michael Batty

Centre for Advanced Spatial Analysis (CASA), University College London  
Gower Street, London, United Kingdom

## Summary

With the emergence of mobile phone trace datasets, new opportunities have arisen for improving the understanding large-scale mobility behaviours. The potential impact of these insight derived from these data is no more significant than in the developing country context, where existing data collection infrastructure is limited or non-existent. In this research, mobile phone data for Dakar, Senegal is used to better understand urban activity and mobility dynamics. To achieve this, a clustering method is introduced that extracts the spatial distribution, and the temporal characteristics, of the activities of individual mobile phone users. With this classification of individual locations of activity, citywide trends in activity and mobility over time are derived. The paper concludes in discussing the potential and limitations of this approach, and the outlook for associated analyses that employ mobile phone trace data.

**Keywords** Big Data; Mobile Phone Data; Mobility; Activity; Developing Countries.

## 1 Introduction

With mobile phone use nearly ubiquitous in both parts of the world, interest is gathering around the potential for using derivative trace data to better understand human behaviour. The potential impact of these datasets is no greater than in the developing world, where datasets considered standard in many countries are unavailable due to a lack of resources. One significant potential avenue of research relates to establishing patterns of activity and mobility in developing countries.

The increasing use of mobile phone traces for the exploration of activity and mobility patterns has led to the development of a range of novel methodologies. In one study located in Tallinn, Estonia, strong similarities were observed in the temporal dynamics in the activities of different individuals (Ahas et al., 2010). Other applications have involved the automatic classification of land use from mobile phone activity patterns (Toole et al. 2012) and identification of trip purpose (Phithakkitnukoon et al, 2010). Network analyses applied to communication interactions derived from mobile phone in identifying the location of ethnic boundaries in Ivory Coast (Amini et al, 2014).

In this paper, urban activity and mobility dynamics are derived through exploration of individual behaviours, capture within mobile phone traces recorded in Dakar, Senegal. The paper introduces a method for the identification of locations of individual activity over different times of day. On an aggregate basis, the identification of activity clusters enables an analysis population density, possible land use, and mobility within between different areas of the city. The paper first outlines the data used in this study, and the method used for the identification of activity locations, and then moves on to exploring citywide trends in activity and mobility.

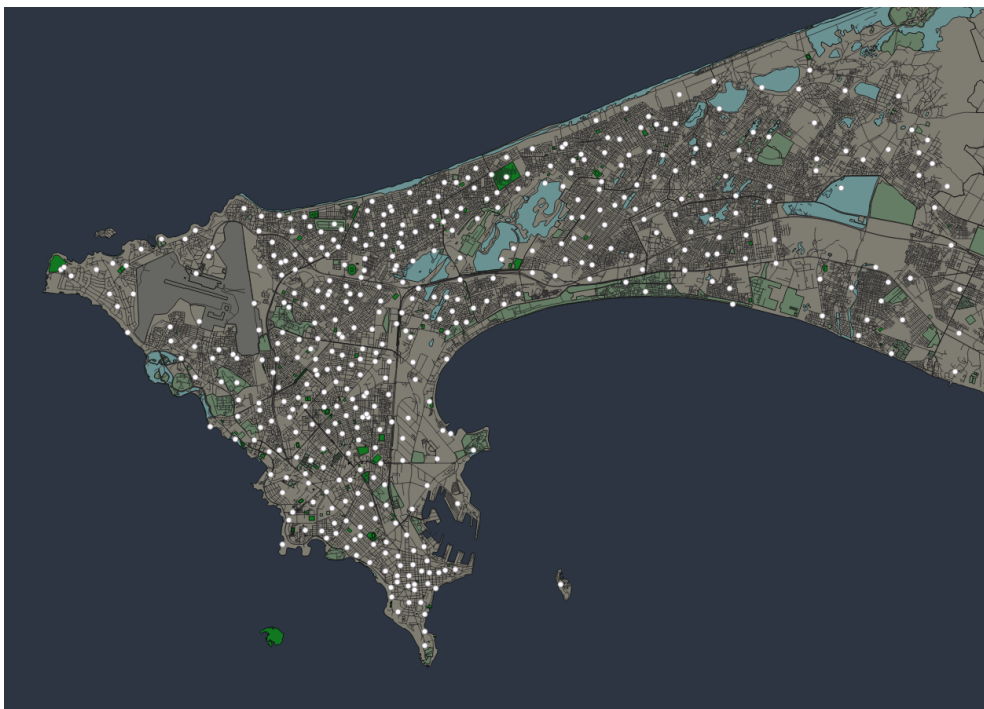
## 2 Study Area and Data Description

The study is located in Dakar, the capital city of Senegal, West Africa. The majority of the city is located on a peninsula on the Pacific Ocean coastline, and contains a population of 1.08 million

inhabitants. Like many cities in developing countries, data collection facilities are not as established nor comprehensive as elsewhere. Although the last census was carried out in 2013, only limited data has since been released.

The study utilises mobile phone usage data provided as part of the 2014 Orange Data for Development challenge. The scheme made multiple mobile phone datasets available for research purposes, recorded for selected Orange mobile phone users in Senegal during the course of 2013.

The dataset used for this study describes the mobile phone usage patterns of individual users during three-week periods over the course of the year. Within this dataset, whenever individuals used their mobile phone, the cell tower through which the call or text message was sent was recorded. The cell tower used for transmission is usually that nearest to the individual, unless they are crossing between zones covered by two towers. Within the area of Dakar, indicated in Figure 1, there are 435 cell towers, with an average minimum distance to the nearest next cell tower 404 metres.



**Figure 1:** The Dakar study area, showing the location of the 435 cell towers.

### 3 Identifying User Activity Locations

An analysis of individual mobile phone usage behaviour was undertaken in order to identify areas of *regular activity*. The location – a cell tower in this case – at which an individual is observed on a regular basis can be assumed to hold a strong significance to that individual. The timeframe within which these activities take place may be indicative of the type of activity being undertaken by the individual.

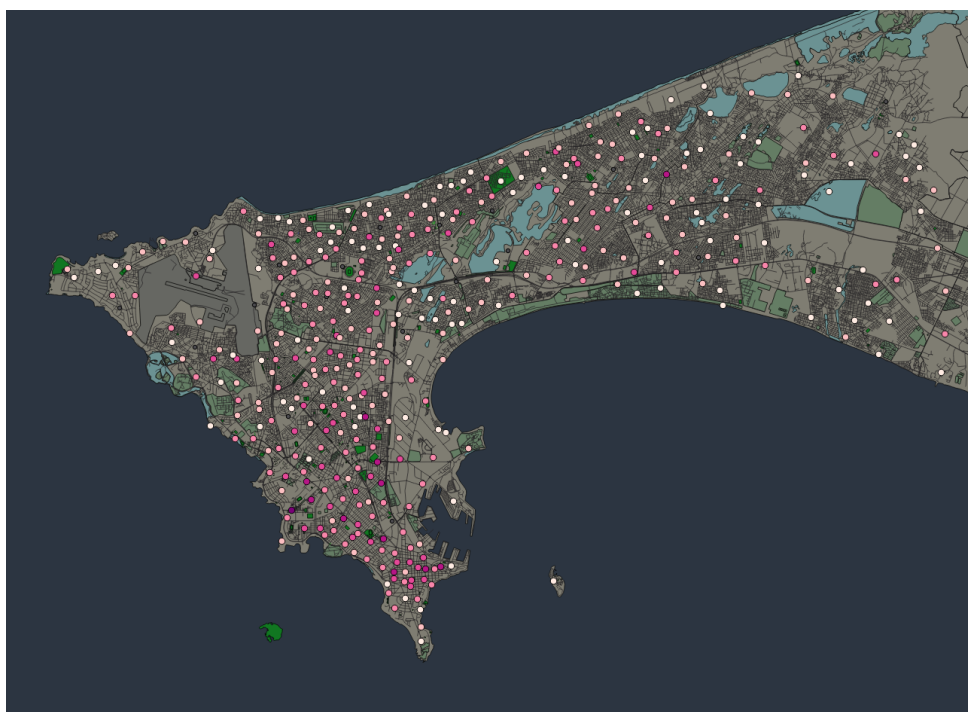
The definition of regularity in this case is defined as the demonstrable presence of an individual at a location, within a given timeframe, on at least 30% of the days they are observed. The 30% threshold ensures the location is visited relatively often and across multiple days, and accounts for the possibility that individuals will not use their device at that location each time they visit.

Locations of activity are identified through clustering the time points at which the individual appears at each unique location. For this purpose, DBSCAN is used, ensuring flexibility in cluster size. The DBSCAN algorithm is specified to only cluster points that fall within 60 minutes of an existing cluster, with all other instances classed as outliers. The result is a set of locations for each individual at which they are observed at around the same time on a regular basis. Each location cluster is defined with a minimum, maximum and mean time.

#### 4 Spatial and Spatiotemporal Variation in Activity

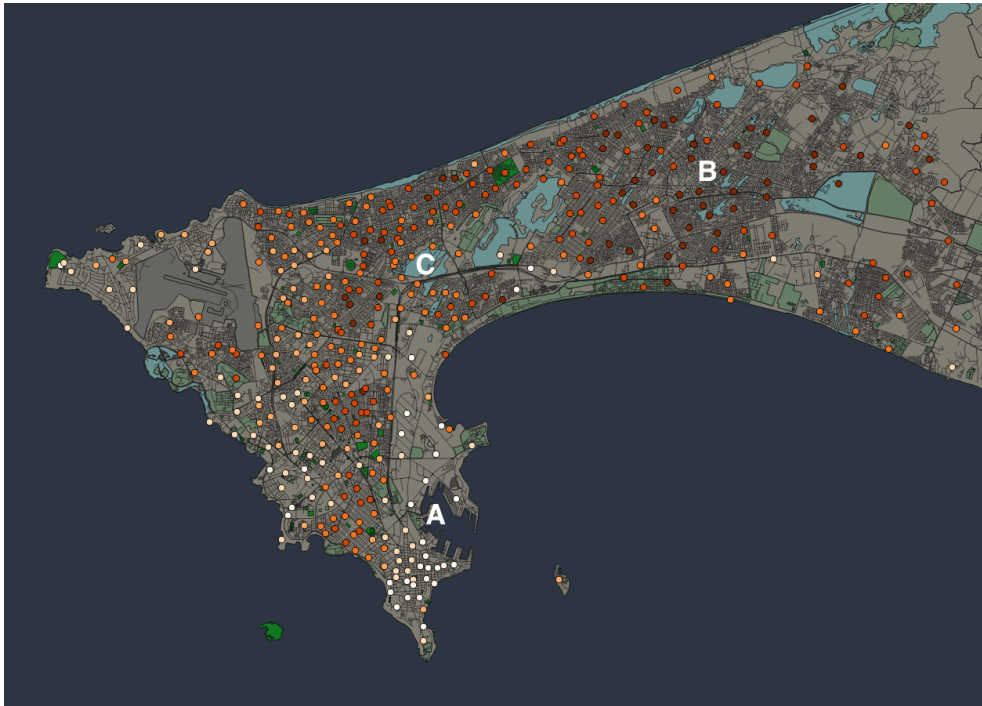
Through identification of individual activity locations over time, one can aggregate to draw out population-level indicators of spatial and spatiotemporal variation in activity across Dakar. This nature of these trends may provide some indication of the types of activity being undertaken at each location.

The first, simplest approach is to extract the sum number of people shown to be visiting each location. This demonstrates the volume of individuals dwelling in each area across the city. This distribution is shown in Figure 2, and indicates a higher proportion of this behaviour around the southern and western areas of the city.



**Figure 2:** Number of individuals visiting each location on regular basis.

Moving onto temporal variation, a useful initial indicator is to extract the mean time at which individuals, observed regularly, are active at each location. Across all cell towers, the mean regular activity time is 16:13, indicating an evening bias towards phone activity. Spatial variation in mean regular activity time, as shown in Figure 3, should be judged relative to the overall mean.

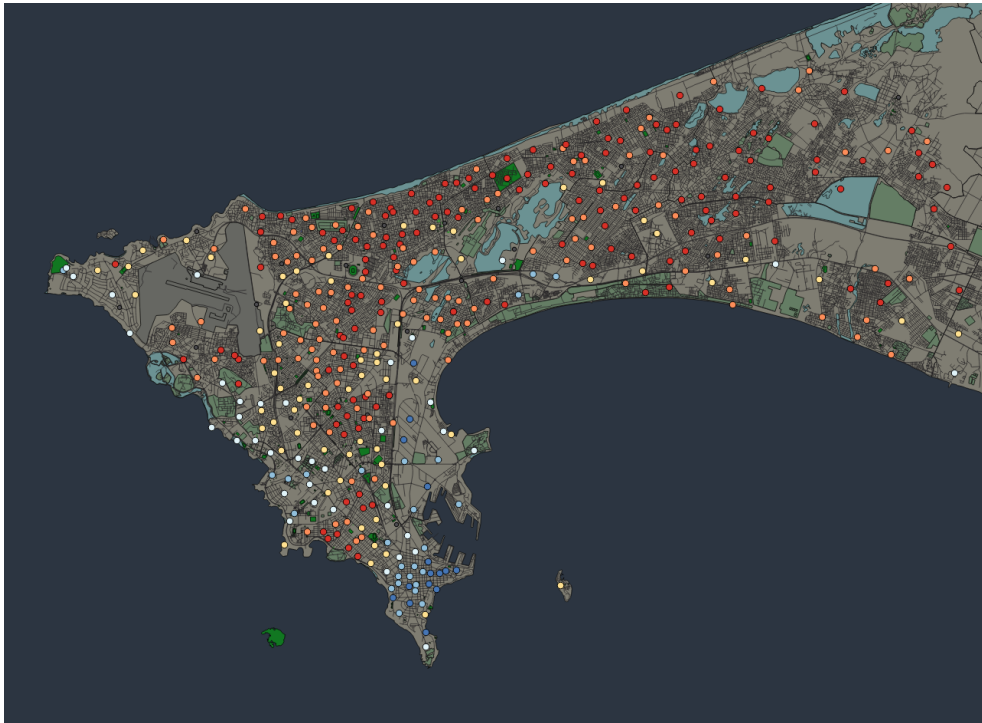


**Figure 3:** Mean regular activity time for each cell tower in Dakar.

As indicated in Figure 3, regions of earlier and later regular activity are present. Most notably in south-western parts of the city (in the area marked A), the mean time dips to around 2pm, the lowest being cell tower ID 224, with a mean time of 13:32. Conversely, in the north-west mean regular activity times are later (area marked B), with far western areas reaching 17:30, the latest mean is found at cell tower 384, where the mean time is 17:45. Likewise, the area marked C also reflects a later mean activity time than the rest of the city.

Average time, however, may mask the presence of conflicting early and late peaks in activity at each location. An alternative viewpoint is to explore the proportion of regular users present at a location within a specific time period, relative to all other times of day. Two time periods are examined – daytime, defined as 10am to 4pm, and evening, defined here as 6pm to 12am – and are shown in Figures 4 and 5.

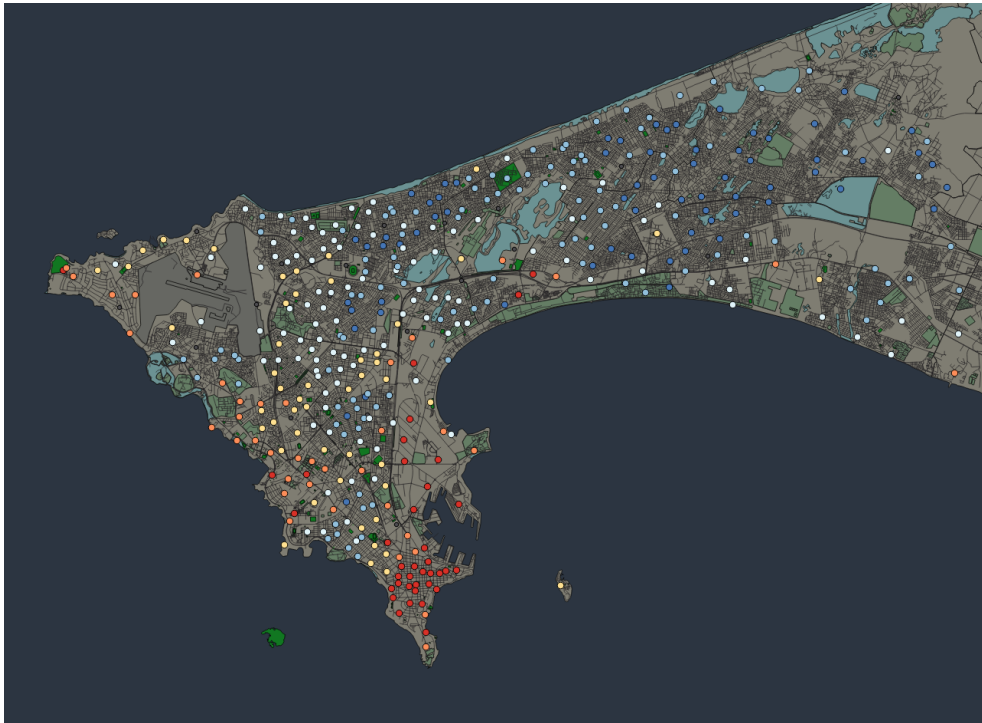




**Figure 4:** Spatial variation in regular activity during daytime hours (10am to 4pm).

The maps provide a little more insight into the spatiotemporal variation in activity across Dakar. It is clear that the southern tip of Dakar is predominantly active during the daytime, where many of the cell towers see 60-70% of their users during the daytime. During the evening, activity in the southern areas drops significantly, dispersing to the northern and western areas. In these areas, 50-60% of all regular users are observed during the evening hours.





**Figure 5:** Spatial variation in regular activity during evening hours (6pm to 12am).

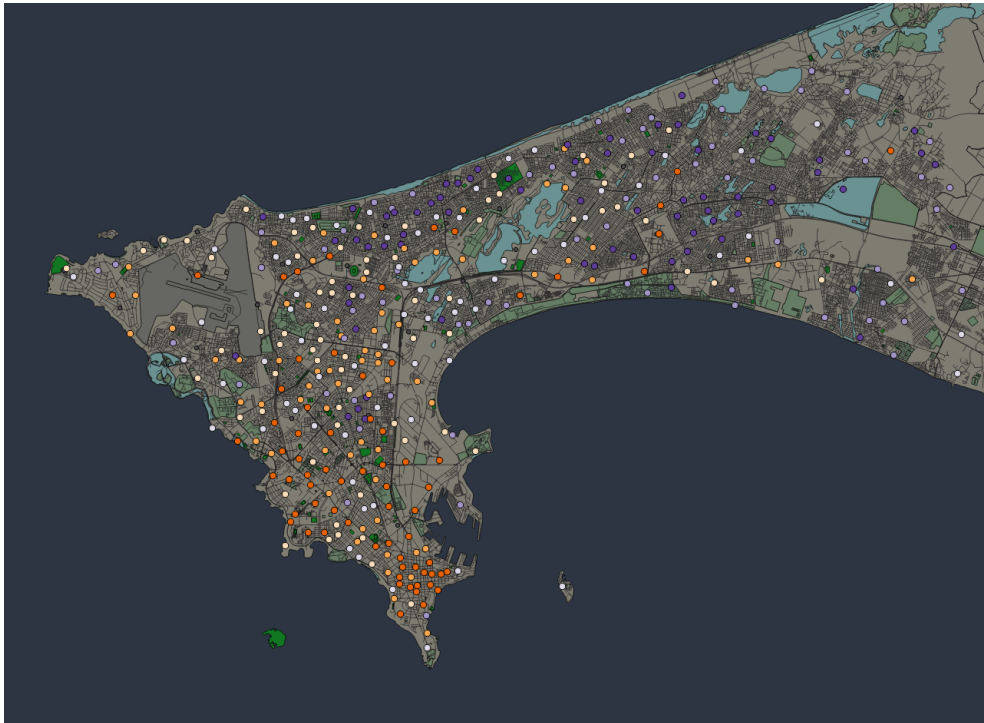
## 5 Mobility Indicators

In addition to identifying location of activity, it is possible to explore the degree of connectivity between cell towers, as demonstrated by the movement of users. This is particularly interesting with respect to identifying movements between ‘home’ regions and other areas of the city.

In exploring this process, a definition of ‘home’ is required. Once more the locations of regular activity are used. In this case, however, the time periods are adjusted to incorporate only individuals observed regularly at locations between 9pm and 5am. These rules identify regular night time locations for 708463 individual users. From this point we assume these are the home locations of these users.

With the definition of night time location assigned to each user, the locations to which these users travel away from this area is identified. To counter the potential that users may be tracked travelling to these locations, only those cell towers at which individuals are shown to dwell (as identified through clustering) are included here. These dwell locations are not required to fit within a certain time period, nor do they need be observed on multiple occasions (as specified in identifying home locations). However, only cell towers outside of a 500m radius of the home location are included, in order to ignore very local movements.

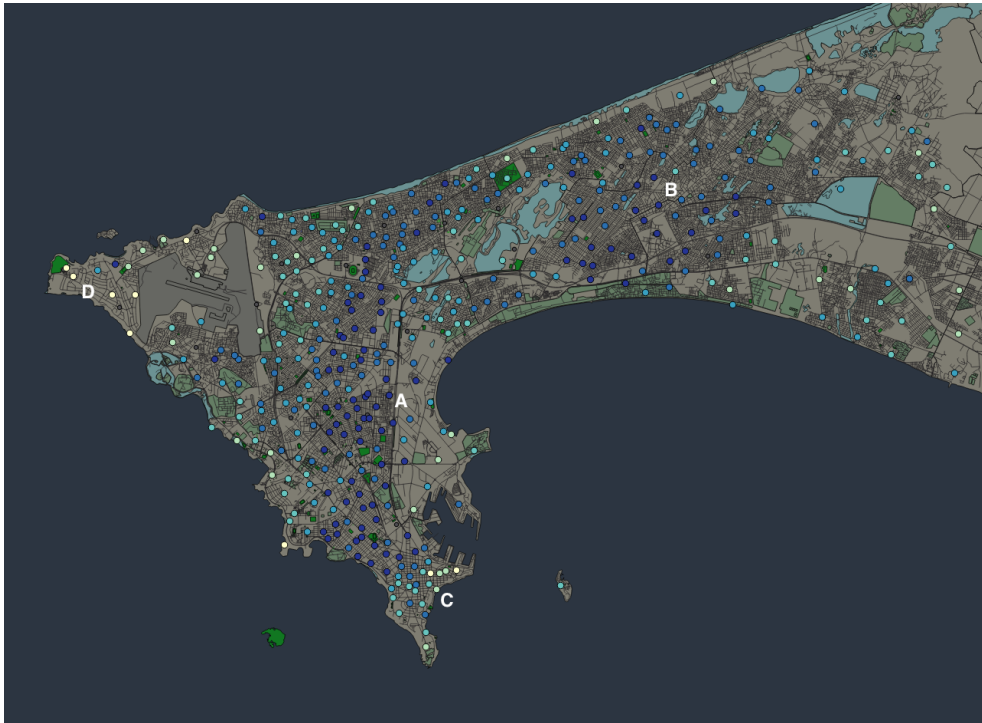
The result is an indication of the mobility of all individuals living at each location, and the locations to which they travel. Through the construction of these networks of mobility for each location, initial insight can be gained by extracting the most popular destinations for all travellers. Using these data, Figure 6 shows the balance between the number of individuals visiting a location and number indicated to be living in that area.



**Figure 6:** Differences in individual living at each location against the number regularly visiting each location

As can be seen from these results, many of the southern areas of the city are popular destination locations relative to the number of people living in these areas. This trend appears to align with the earlier identified patterns, whereby these same areas were more often visited during daytime than the evening. The highest imbalance between residents to visitors is found at cell tower 77 on the western coast. Maps of Dakar suggest that this tower is positioned on the University of Dakar campus, and so these results would align with expectations.

Another alternative measure than can be drawn from exploring mobility is the mean total mobility of individuals originated at each location. These measures provide an indication of the degree of mobility of individuals living at each location. The results for each location are shown in Figure 7.



**Figure 7:** Spatial variation in average total mobility of inhabitants at each cell tower.

The results indicate considerable variation in average total mobility across the city. In areas such as A and B, as indicated in Figure 7, total mobility is relatively low on average, at between 5km and 8km during the time period. This increases to upwards of 20km on average in the regions marked C and D. This latter trend is interesting as these areas are equally those of high attraction to visitors. It may be concluded that the individuals living in these attractive areas are more affluent, and there have more opportunity to travel than others in the outer suburbs.

## 6 Discussion and Conclusions

This paper has demonstrated how mobile phone usage data can be used to better understand activity and mobility patterns within a developing world context. Through the identification of individual locations of regularity – including those locations that may be assumed to be home locations – it has been possible to generate indications of activity for individuals residing at each cell tower. While these initial explorations currently lack validation, there are some clear, promising trends that should be further investigated.

Despite the promise of these initial findings, some notes of caution should be highlighted. The identification of user home locations is clearly problematic, as the repeated visit of a location at night does not necessarily indicate that that location is that person's home. Furthermore, within this dataset, location is only captured when the phone is used. This biases the location data for an individual towards places where they stop for a reason amount of time. It furthermore means it is not possible to fully capture the complete set of trips undertaken by an individual. As such, the macroscopic picture provided at this stage is likely the most appropriate use for the data. Finally, there are likely to be inherent demographic and usage biases within the dataset too. The requirement that an individual has a phone and uses it with some regularity requires that that person must have some free income. This likely leads to a more affluent study group than may be present within the wider population.

These limitations aside, the initial results described here indicate the potential for further exploration with these datasets. With little quantitative understanding of activity and mobility within developing world cities, mobile phone data may reflect a useful indicator for planners looking to better understand place and transportation needs.

## References

Ahas, Rein, et al. "Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data." *Transportation Research Part C: Emerging Technologies* 18.1 (2010): 45-54.

Amini, Alexander, et al. "The impact of social segregation on human mobility in developing and industrialized regions." *EPJ Data Science* 3.1 (2014): 6.

Phithakkitnukoon, Santi, et al. "Activity-aware map: Identifying human daily activity pattern using mobile phone data." *Human Behavior Understanding*. Springer Berlin Heidelberg, 2010. 14-25.

Toole, Jameson L., et al. "Inferring land use from mobile phone activity." *Proceedings of the ACM SIGKDD international workshop on urban computing*. ACM, 2012.

# A Spatiotemporal Population Subgroup Model of Radiation Exposure

Martin B<sup>\*1</sup>, Martin D<sup>†1</sup> and Cockings S<sup>‡1</sup>

<sup>1</sup>Geography and Environment, University of Southampton, SO17 1BJ.

March 13th, 2015

## Summary

Understanding the whereabouts of vulnerable population subgroups during emergencies can improve the targeting and implementation of countermeasures, including evacuation and sheltering. This paper uses spatiotemporal population density modelling and atmospheric dispersal modelling to estimate the radiation exposure of a specific population at different times of day, during the start of a hypothetical radiation accident scenario in Exeter, UK. The model outputs are analysed by GIS to discern spatiotemporal trends in population exposure, and to identify the times of day when population subgroups may be most at risk.

**KEYWORDS:** Spatiotemporal population modelling, Public health, Radiation protection, Risk.

## 1. Introduction

The UK began the world's first civil nuclear energy production programme in 1956 and has a successful legacy of power generation. Nuclear energy currently contributes toward 18.5% of the UK electricity portfolio and a new phase of reactors is anticipated, following publication of the Nuclear Industrial Strategy (Bolton, 2013, HM Government, 2013).

Emergency preparedness is an important feature of nuclear installation (NI) management. All UK NIs are required to have off-site emergency planning to comply with Radiation (Emergency Preparedness and Public Information) Regulations (REPPPIR). REPPPIR includes the testing of hypothetical scenarios to inform understanding of potential outcomes and to improve decision-making. Nuclear and radiation emergencies are low-likelihood but extremely high impact events which have long-term public health implications. Demographic studies of historical accidents, including Fukushima Daiichi (2011), Chernobyl (1986), Three Mile Island (1979) and Idaho National Engineering Laboratory SL-1 (1961) have been used to advise preparedness. Fortunately, the UK has not experienced an accident of equivalent scale to these accidents. This paper tests a hypothetical scenario which includes accurate spatial and temporal population profiles, to understand how the timing of the start of an accident may cause differential exposure to vulnerable population subgroups.

---

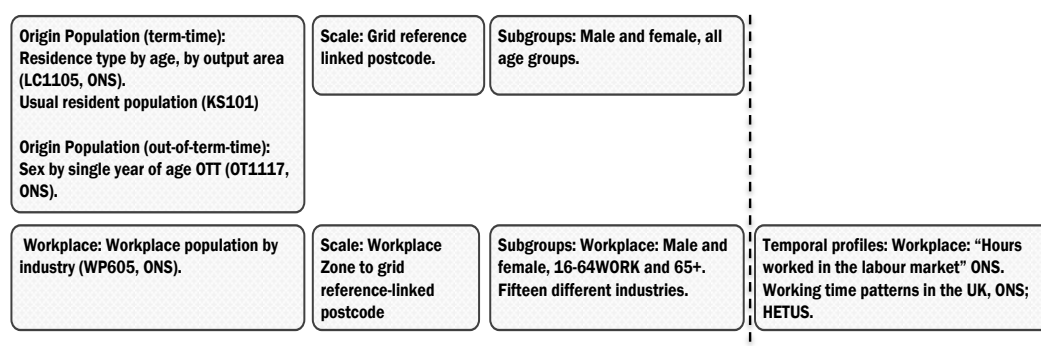
\* Becky.Martin@soton.ac.uk

† D.J.Martin@soton.ac.uk

‡ S.Cockings@soton.ac.uk

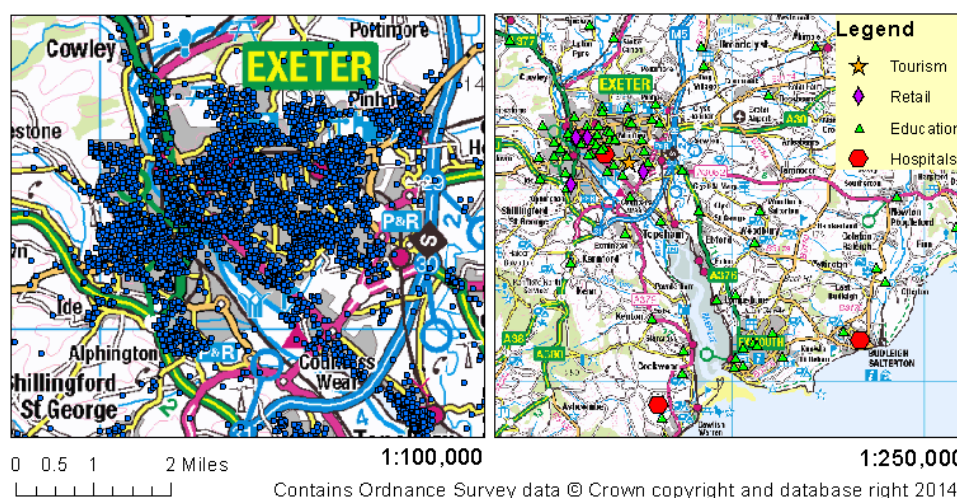
## 2. Methodology and Data

Gridded spatiotemporal population density modelling was combined with atmospheric dispersal modelling by GIS analysis. SurfaceBuilder was implemented to model spatiotemporal population density (Smith, 2013). Using this model, an adaptive kernel density algorithm was applied to redistribute population subgroups from individual postcode-based origin centroids, to destination centroids, and onto a transport network. The proportion and distance of the redistribution was determined by centroid density, catchment size, and time; and was dasymmetrically constrained to prevent inappropriate relocation. It is important to include different population subgroups, due to age and gender differences in daytime spatiotemporal activity patterns, which can result in differential exposure. There are also some physiological differences between body mass, respiration and susceptibility to the effects of radiation exposure, across age and gender subgroups (Shore, 2014, Simon and Linet, 2014). Spatiotemporal distribution profiles were constructed for six new age groups and two new genders with 2011 data. An example of population data sources, scales, subgroups and temporal profiles within this case study is shown by **Figure 1**.



**Figure 1:** A residential and workplace population example of data sources, scales, subgroups and temporal profiles.

However, the model also includes 2011 education, healthcare, retail, tourism, and leisure data to provide a comprehensive insight into the spatiotemporal whereabouts of different population subgroups during day-time. **Figure 2** shows the distribution of some of these activities, compared to the residential population distribution.



**Figure 2:** The original spatial distribution of residential population data (left) and activities (right).



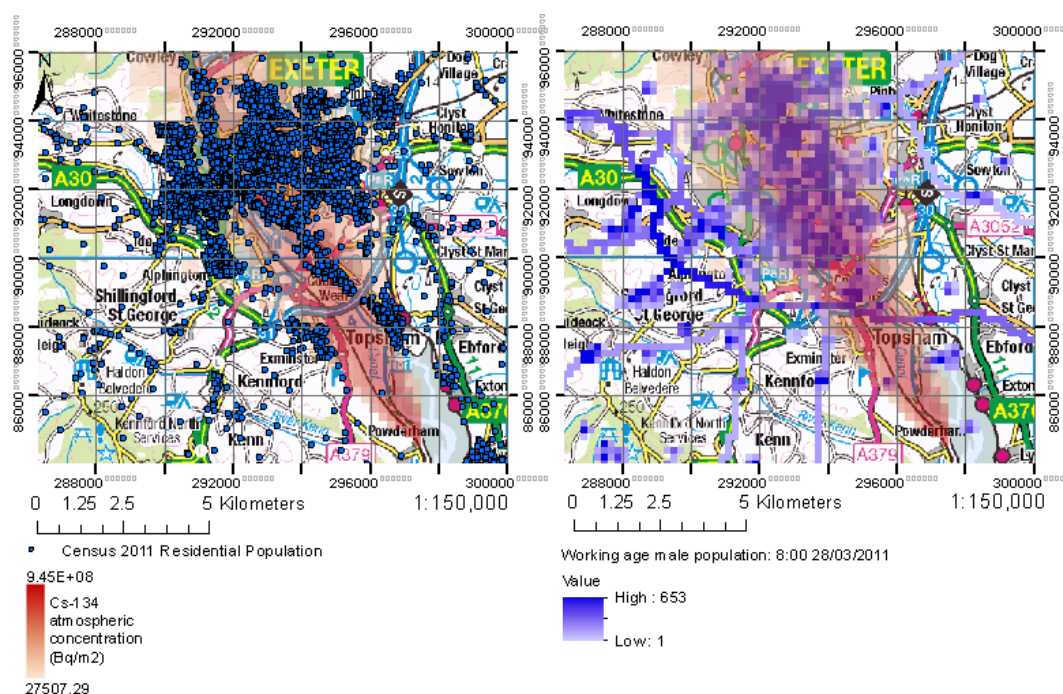
The Numeric Atmospheric Modelling Environment (NAME) is a Lagrangian model which uses Monte-Carlo random walk techniques to represent and predict turbulent atmospheric transport, and the deposition of airborne substances in the atmosphere within a stochastic framework. NAME was applied to model the dispersal of 0.47PBq of Cs-134 a source term of  $\approx 1\%$  of Chernobyl for this isotope. Archive MESUM5 meteorological data used to generate dispersal for an hours' time slice from 08:00 to 09:00 on Monday 28<sup>th</sup> March 2011. This slice is being used to represent the start of an incident, to investigate exposure differentials for different populations. Regional weather was dry with hazy sunshine and a peak temperature of 19°C, providing good conditions for dry deposition, which can be a source of external exposure in urban environments. However washout of atmospheric particles and gases may be a more significant exposure mechanism (IAEA, 1994). Dry deposition is also affected by deposition surface, but this is beyond the scope of this paper.

Atmospheric plume dispersal model and spatiotemporal population model data layers were combined using GIS to assess exposure likelihood, by concentration (Bq/m<sup>2</sup>) for each grid cell of residential population density at 08:00 and 20:00.

### 3. Results

A study area of 15km<sup>2</sup> centered upon the City of Exeter (X: 286000, Y: 079500) was selected. Exeter is a location without a nuclear installation (NI), and is therefore a suitable analogue site. The city includes national and international rail, road and air transport infrastructure, and has a residential population of approximately 117,770 individuals (ONS, 2014).

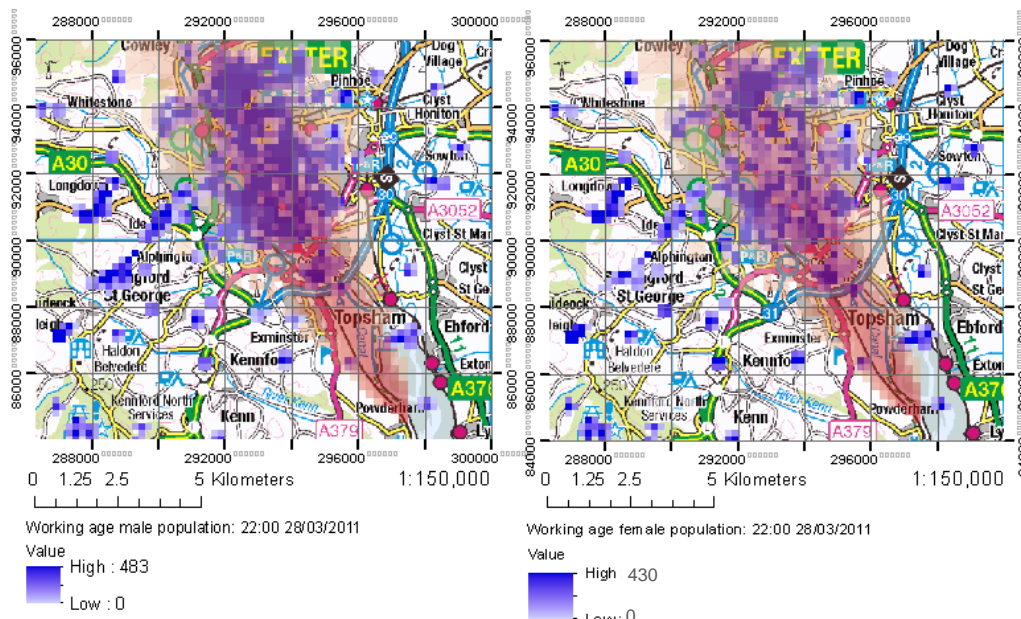
**Figure 3** shows some modelling results from the study area. Both images include NAME output for the dispersal of an atmospheric plume of Cs-134 from 08:00 to 09:00 on 28<sup>th</sup> March 2011. This has been combined with two different outputs of population distribution model, for assessment of radiation exposure.



**Figure 3:** GIS analysis of NAME Cs-134 plume dispersal has been combined with a) Census total residential population centroid distribution; and b) SurfaceBuilder spatiotemporal population density surface output for a working-age male population subgroup, at 08:00 on 28<sup>th</sup> March 2011.

**Figure 4** shows the difference between male and female spatiotemporal distribution for the working

age population subgroup, at 22.00 on 28<sup>th</sup> March 2011. It is evident that fewer females are present within the city centre at the time. Combining this information with the plume model output, significantly more males of working age are likely to be exposed than females of working age within this scenario. This may be due to more females working in occupations that do not require evening-shift working patterns in this region.



**Figure 4:** Comparison of male and female working age population distributions with equivalent NAME model output, at 22.00 on 28<sup>th</sup> March 2011.

Comparing the temporal profiles of the male and female working-age population subgroups confirms that more males than females should be anticipated within the city at 22.00 on a working weekday, and that therefore males are potentially more vulnerable to the effects of radiation exposure at the start of an incident, within this specific application.

#### 4. Discussion and Conclusions

The inclusion of spatiotemporal population density modelling offers improvements upon the traditional choropleth map, by revealing spatial population subgroup change through time. Whilst the hypothetical scenario of differential male and female exposure to radiation is interesting, the key purpose of this study is to demonstrate that spatiotemporal radiation plume dispersal modelling and population density modelling can be combined to offer new insights into the likelihood of subgroup exposure to radiation and its cumulative effects; providing substantial improvement to existing comparative study methodologies across different times, spaces, ages and genders, for any location where appropriate data is available.

Whilst this study provides a methodology for assessment of exposure at the start of a radiation emergency, there is still a need for a model which estimates the deterministic and stochastic health effects of radiation exposure to different populations in space and time.



## 5. Acknowledgements

The authors gratefully acknowledge the advice of Matthew Hort and Laura Burgin at the Met Office, and Stephanie Heywood and Tom Charnock at Public Health England.

Data: OS 1:150,000 Scale Raster [WMS map service], Coverage: Exeter, Ordnance Survey/EDINA supplied service © Crown Copyright 2014. NAME modelling output and MESUM5 archive weather datasets © Copyright/database rights Met Office and Public Health England 2003-2014, Census Output Area Boundaries and Workplace Zones © Crown copyright 2011.

## 6. Biography

Becky Martin is a PhD researcher in Geography at the University of Southampton. Her research interests include spatiotemporal demography, public health and radiation protection.

David Martin is a Professor of Geography at the University of Southampton. His research interests are focused on social science applications of GIS.

Dr Samantha Cockings is an Associate Professor of Geography at the University of Southampton. She worked on AZTool development and provided methods for new Census 2011 Workplace Zones.

## References

- BOLTON, P. 2013. Nuclear Energy Statistics. House of Commons Library: UK Government.
- HM GOVERNMENT 2013. The UK's Nuclear Future. Her Majesty's Government: Department for Business, Innovation & Skills and Department of Energy & Climate Change.
- IAEA 1994. Modelling the deposition of airborne radionuclides into the urban environment. *In*: VAMP URBAN WORKING GROUP (ed.). Austria: IAEA.
- ONS. 2014. *Neighbourhood Statistics* [Online]. Available: <http://www.neighbourhood.statistics.gov.uk/> [Accessed 10th October 2014].
- SHORE, R. E. 2014. Radiation Impacts on Human Health: Certain, Fuzzy, and Unknown. *Health physics*, 106, 196-205.
- SIMON, S. L. & LINET, M. S. 2014. Radiation-Exposed Populations: Who, Why, and How to Study. *Health physics*, 106, 182-195.
- SMITH, A. D. 2013. 24/7 population modelling to assess exposure to natural hazards. *Colloquium on Spatial Analysis*. University of Copenhagen, Denmark

# Understanding the urban experience of people with visual impairments

Panagiotis Mavros<sup>1</sup>, Katerina Skroumpelou<sup>1</sup>, Andrew Hudson Smith<sup>1</sup>

<sup>1</sup> Centre for Advanced Spatial Analysis, The Bartlett, University College London

<sup>2</sup> School of Electrical and Computer Engineering, National Technical University of Athens

November 7, 2014

## Summary

One of the major issues visually impaired people face in everyday is the difficulty to navigate the city independently, which has implications for their wellbeing and health. As part of a research collaboration with the Guide Dogs for the Blind Association and Future Cities Catapult, we have employed mobile Electroencephalography (mEEG) to study the urban experience of visually impaired people. Our pilot study demonstrates the potential of such methods to provide insights both through the analysis of data, but also by using the visualisation of emotional experience of the city as a tool for empathy.

**KEYWORDS:** mobile EEG; Emotiv; urban mobility; pedestrians; visual impairment.

## 1. Introduction

According to recent RNIB figures<sup>1</sup>, in 2011 there were 1,865,900 people living with sight loss in the UK -- of which 223,500 live with severe sight-loss (blindness) -- a number which is expected to rise to a total of 2.4 million in the next twenty years. Visual impairments can have significant effects on physical, mental and emotional wellbeing of individuals and it is essential to maintain an independent and active life. In this context, as part of the Cities Unlocked project, our project with Guide Dogs for the Blind Association (henceforth Guide Dogs) and Future Cities Catapult applies novel research methods to investigate urban mobility and the experience of the city of blind and partially sighted individuals. The aim of this report is to present some key aspects of the ongoing study and highlight early results.

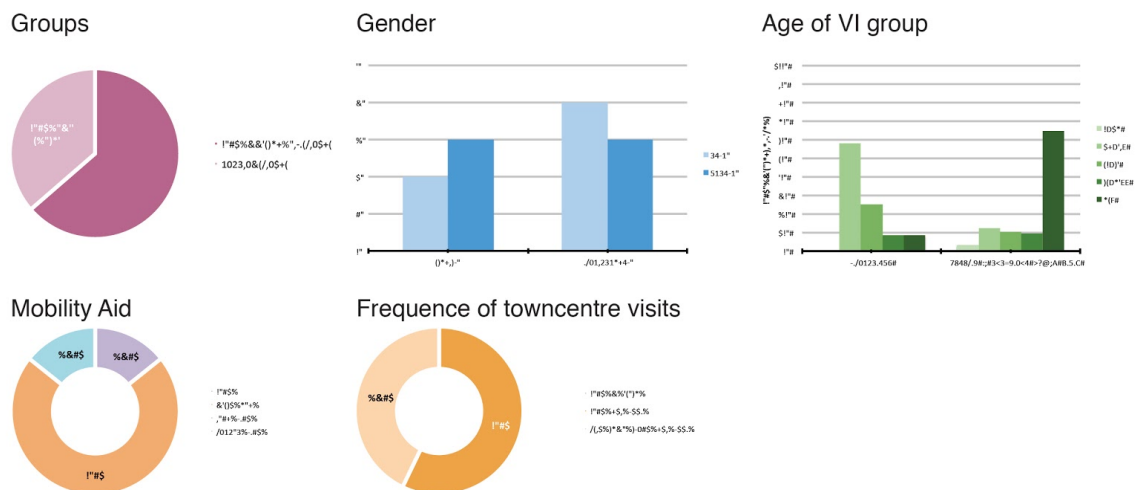
Walking and navigating in outdoor environments, including town or city centres, is a major part of an active lifestyle, but involves significant physical, mental and emotional challenges for visually impaired people: negotiating traffic, obstacles and street crossings, locating points of interest and mitigating difficulties. Secondary accessibility barriers also include the lack of appropriate signage or even

---

<sup>1</sup> RNIB Sight Loss Data Tool, version 2. Available at:  
[www.rnib.org.uk/knowledge-and-research-hub-key-information-and-statistics/sight-loss-data-tool/Sight\\_Loss\\_Data\\_Tool\\_Version\\_2.0.xls](http://www.rnib.org.uk/knowledge-and-research-hub-key-information-and-statistics/sight-loss-data-tool/Sight_Loss_Data_Tool_Version_2.0.xls)

architectural features such as spatial or urban legibility. On a first level, this study aims to tie the cognitive, emotional and functional (usability) dimensions of urban mobility to contribute insights for the development of new designs and services. On a second level, using data and visualisations to portray the similarities and differences in the experiences of sighted and visually impaired people, aims promote understanding, empathy and perspective taking among designers, stakeholders and citizens in general.

Exploring these questions, this study applies a battery of novel sensing technologies, such as mobile Electroencephalography (EEG) using the Emotiv EEG (Mavros et al., 2012; Aspinall et al., 2013), measurement of skin conductance as an indicator of psychophysiological arousal, as well as activity and location tracking using smartphone devices. These sources of data were combined with qualitative methods, such as established questionnaires tapping on emotion (Matthews et al., 1990), wellbeing (Tennant et al., 2007) and spatial cognition (Hegarty et al., 2002; Pazzaglia and Debeni, 2001) as well as semi-structured interviews.

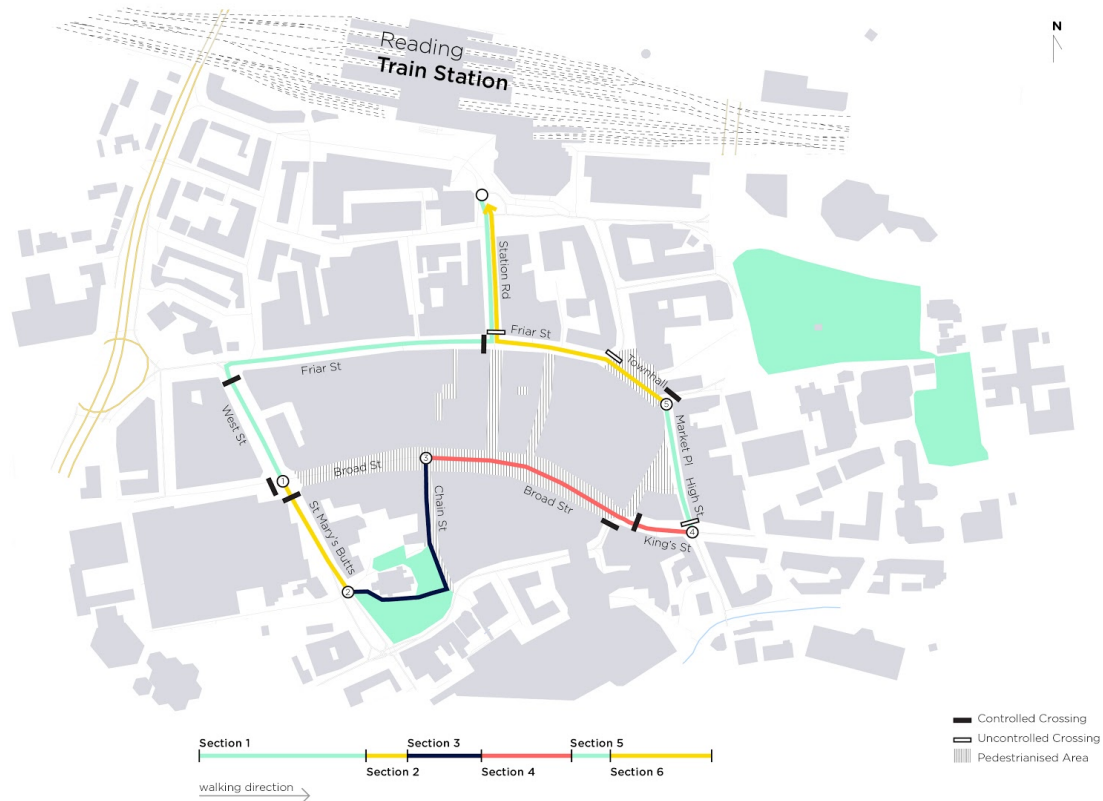


**Figure 1: Basic elements of the study**

## 2. The experiment

To understand how a visit to a familiar town centre environment (Bentzen et al., 2004) might affect people, twelve (12) sighted, partially sighted and blind volunteers (figure 1) were asked to complete various everyday mobility tasks, using their preferred mobility aid (long cane, symbol cane or guide dog) and then tell us in detail about their experience. Participants completed a circular, 1.8 km long ‘experimental route’ in central Reading. The route was designed to lead them through a variety of typical and everyday urban environments, such as streets with and without shops, a small park, narrow

and wide pedestrian areas, like Broad Street, and use various controlled with various accessible pedestrian signals, as well as uncontrolled crossings (figure 2).

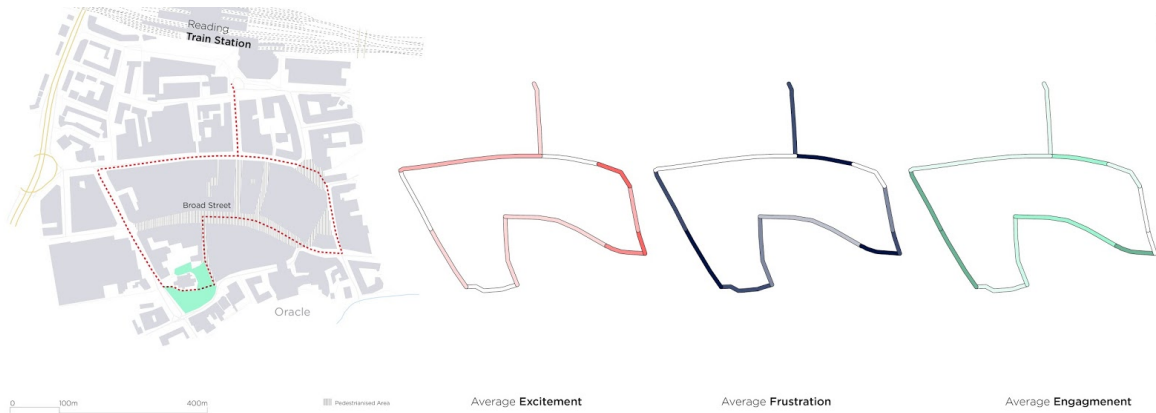


**Figure 2:** Map of the 1.8km experiment route. Participants were asked to walk a cyclical route starting and ending at Reading train station, through a series of instruction points (numbered). Black bars indicate the direction of controlled junctions, while empty bars indicate uncontrolled junctions that were traversed by participants.

### 3. Localising the urban affect

Initial analyses of participants' electrophysiological data (EEG) based on Emotiv's emotion detection algorithms, demonstrate the potential of these methods. Figure 3 shows maps of emotional states of visually impaired participants, aggregated per street, and allows us to explore the affective component of environmental conditions (e.g. motorised and pedestrian traffic) or the impact of open pedestrian areas. Further, figure 4 is a map visualising only the peaks of 'excitement', an EEG derived measure of arousal, and illustrates that crossing junctions as well as moments of verbal interaction with participants, stand out from the walking experience and can be easily distinguished through the data.

Further analyses will focus on the experience of streets segments versus junctions, and also compare the routes of the visually impaired and control group.



**Figure 3:** Emotional states of the visually impaired group of participants, aggregated by route segment.



**Figure 4:** Aggregated peaks of 'excitement' of the visually impaired group of participants, occur at junctions and street crossings as well as points of social interaction (verbal communication, obstacles).

Analysis of the self-reported data, on the other hand, highlight other aspects of the mobility. Both groups seem to 'agree' that walking through a green area, reduces perceived stress, inline with theories about the restorative qualities of green and natural environments. However, in contrast to the control group, when walking through a large pedestrianised area, such as Broad street, visually impaired participants reported higher perceived stress. This could be explained by the increased numbers of pedestrians and lower number of tactile paving (drop kerbs, tactile paving) that can act as cognitive landmarks.

Further, observational and interview data confirm the stress-inducing nature of everyday incidents, such as negotiating unpredicted obstacles on pavements, including overhanging branches, cyclists or parked cars blocking crossings. Other factors, include inconsistencies in urban infrastructure, e.g. junctions that include both controlled and uncontrolled sections, or unexpected deviations from familiar walking routes, e.g. due to roadworks or bus stop changes. These factors need to be addressed by relevant stakeholders as they are known to contribute to feelings of stress, anxiety and spatial confusion, that often lead visually impaired people to limit their levels of independent activities (Kitchin et al., 1998).

#### **4. Discussion**

To conclude, our initial analyses of the data suggest that novel methods such as mobile EEG and emotion analysis can capture the multiple facets of exploring the city, highlighting significant points during our participants walks, such as street crossings and verbal interactions. Data from our longitudinal study demonstrate the potential of quantified-self applications, like activity tracking to capture the extends and diversity of mobility patterns. For our presentation at GISRUK we will include a fine grained and statistical analyses of the various collected data will seek to reveal spatial and temporal patterns, issues, similarities and differences in the experiences of multimodal urban mobility.

#### **5. Acknowledgements:**

First of all, we would like to thank all the participants who have volunteered their time and energy to take part in the study. Further, we like to thank the Guide Dogs staff Karen Potter, Caoilfhionn Lee, Chris Yates, Susie Luff, Andy Gatenby for this project would not have been possible without the energy, efforts and knowledge. We would also like to thank Claire Mookerjee from the Future Cities Catapult, Jenny Cook from Guide Dogs for initiating and supporting this collaboration, as well all the other Futures Cities Catapult and Guide Dogs staff that assisted in various aspects of the study.

#### **6. Biography**

Katerina Skroumpelou is a PhD student at the School of Electrical and Computer Engineering at the National Technical University of Athens (NTUA). She is an Architectural Engineer of NTUA and holds an MRes on Advanced Spatial Analysis and Visualisation by the Centre for Advanced Spatial Analysis of UCL.

Panagiotis Mavros, is a PhD Candidate and his research is focused on the use of mobile EEG in the study of spatial cognition and behaviour. He was trained as an Architect Engineer at NTUA, and holds an MSc by Research in Digital Media and Culture by the University of Edinburgh.

Dr Andrew Hudson-Smith is Director of the Centre for Advanced Spatial Analysis (CASA) at The Bartlett, University College London. Andy is a Reader in Digital Urban Systems and Editor-in-Chief of Future Internet Journal, he is also an elected Fellow of the Royal Society of Arts, a member of the Greater London Authority Smart London Board and Course Founder of the MRes in Advanced Spatial Analysis and Visualisation and MSc in Smart Cities at University College London.

## 7. References:

Bentzen, B. L., Barlow, J. M., & Bond, T. (2004). Challenges of unfamiliar signalized intersections for pedestrians who are blind: Research on safety. *Transportation Research Record: Journal of the Transportation Research Board*, 1878(1), 51-57.

Hegarty, M., Richardson, A., & Montello, D. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence*, 30, 425–447. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0160289602001162>

Kitchin, R., & Blades, M. (2002). *The cognition of geographic space* (Vol. 4). IB Tauris.

Kitchin, R. M., Jacobson, R. D., Gollledge, R. G., & Blades, M. (1998). Belfast without sight: exploring geographies of blindness. *Irish Geography*, 31(1), 34-46.

Marques-Brooksopp, L. (2012). The broad reach of the wellbeing debate: Emotional wellbeing and vision loss. *British Journal of Visual Impairment*, 30(1), 50–55. doi:10.1177/0264619611428244

Matthews, G., Jones, D. M., & Chamberlain, A. G. (1990). Refining the measurement of mood: The UWIST Mood Adjective Checklist. *British Journal of Psychology*, 81(1), 17–42.

Pazzaglia, F., & Beni, R. De. (2001). Strategies of processing spatial information in survey and landmark- centred individuals, (September 2013), 37–41.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., ... Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes*, 5, 63. doi:10.1186/1477-7525-5-63





# Traffic Prediction and Analysis using a Big Data and Visualisation Approach

Declan McHugh\*<sup>1</sup>

<sup>1</sup>Department of Computer Science, Institute of Technology Blanchardstown

March 10, 2015

## Summary

This abstract illustrates an approach of using big data, visualisation and data mining techniques used to predict and analyse traffic. The objective is to understand Traffic patterns in Dublin City. The prediction model was used as an estimator to identify unusual traffic patterns. The generic model was designed using data mining techniques, multivariate regression algorithms, ARIMA and visually correlated with real-time traffic tweets. Using the prediction model and tweet event detection. The result is a high-performance web application containing over 500,000,000,000 traffic observations that produce analytical dashboard providing traffic prediction and analysis.

**KEYWORDS:** Big Data, Data Mining, Visualisation, Traffic Analysis, Twitter Analysis.

---

\*declan.mchugh@gmail.com

## 1 Introduction

The aim of this paper is to analyse traffic patterns for an urban city, Dublin and to provide a visual dashboard for analyzing traffic patterns. The data sets used vary from remotely sensed data and social media information from open data sources.

One of the challenges of this work is the Big Data (Four V's), Sheth (2014). Using data mining techniques for resolving issues around data quality and performing complex aggregation tasks in the form of Map Reduce enabled high-performance computation executions.

When visualizing the correlation between traffic-related tweets and adverse Traffic conditions, it is necessary obtain a prediction model. The prediction model provides an expected travel time. The actual travel time compared to the predicted travel time as a key performance indicator (KPI).

Variables from weather data, moving average and spatially related data, multivariate regression and statistical models is implemented. For each monitored traffic segment along with datasets to perform statistical and visual analysis such as the impact of weather conditions and spatial patterns.

The models were used to perform analysis on the volatility, the effect weather and prediction on travel times. Using visualization, the interpretation of the analysis is made simple using an analytic dashboard. The features of the dashboard allow the user compare inbound and outbound travel time, weather analysis, volatility analysis, twitter analysis for any hour of any day.

Using a classification model on Twitter data, traffic-related tweets were plotted on Google-Maps to identify Traffic events. The events can be visually correlated against the spurious Traffic events from the traffic prediction models. Tweets classified as traffic-related provided insights into the predictions the varied away from actual travel times.

The following sections provide more detail on the data sources, visualizations and conclusions.

## 2 Data Sources

Open data sources DubLinked (2014), Wunderground (2014) and Twitter (2014) were used in this work for the visualizations. The following section is a summary of the data used in the data collection and creation of the visualizations.

### 2.1 Traffic Data Sets

The prediction model was generated from data accumulated from open source portal known as DubLinked.

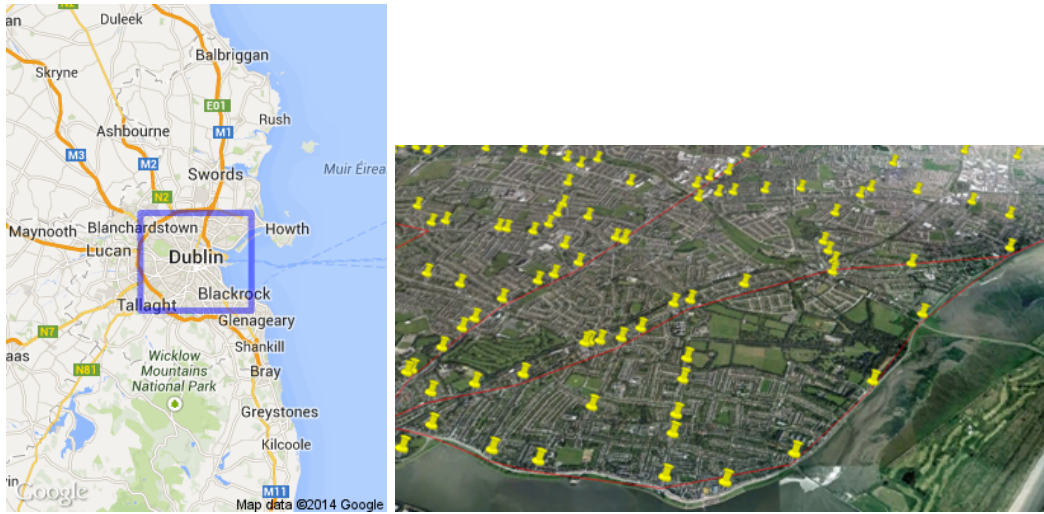


Figure 1: Dublin Traffic Data

### 2.2 Weather Data Sets

The Wunderground API is used to access three of the available historical weather data points 2.

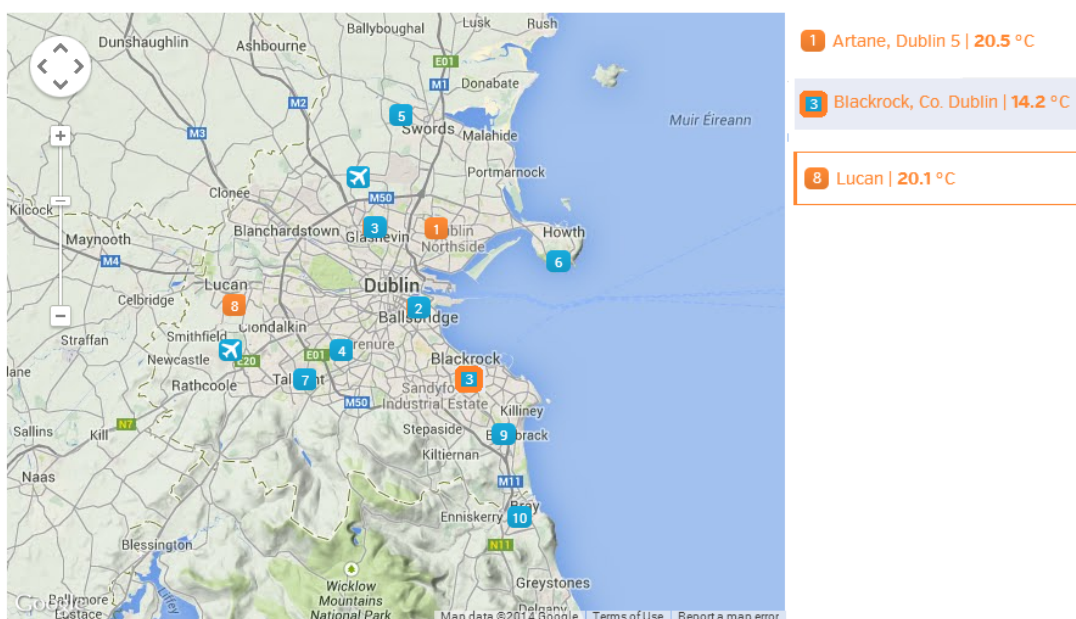


Figure 2: Open Data Weather Stations Dublin

## 2.3 Twitter Data Sets

Twitter provides an API for searching historical and real-time data. The historical data is used to as training for the classification model. Using traffic-related tweets from Twitter account, **#aaroadwatch** a training set is formed. The non-traffic tweets from the real-time data that contain geospatial data were used to make it possible to plot a tweet onto the map.

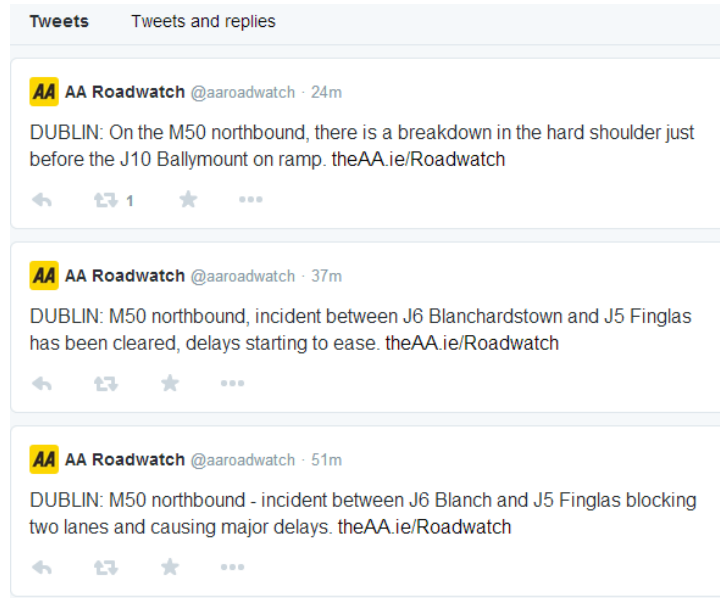


Figure 3: AA Roadwatch Tweets

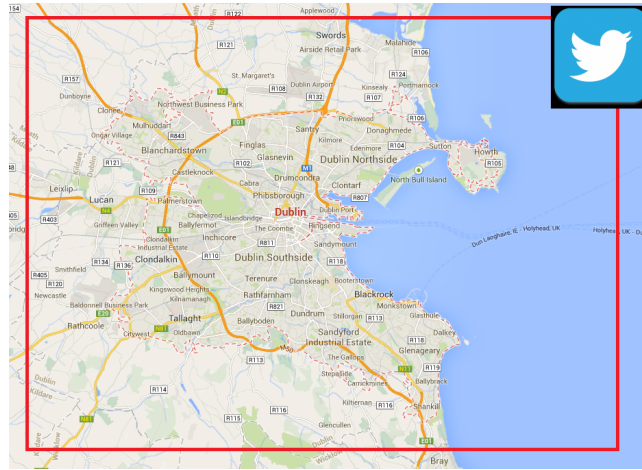


Figure 4: Monitoring of Twitter Stream

### 3 Visualizations

The analytics dashboard enables the users to identify the patterns of traffic visually as mentioned in the introduction 1. The visualisations are generated using Google Maps, JQuery along with a Python backend.

#### 3.1 Volatility

Volatility is a way of identifying an inconsistency in the travel time. With this users can identify areas that are prone to delays. Standard deviation can be considered a way of measuring the volatility according to a paper from Tulloch (2012). A range of colors provides the result of the standard deviation from low red equal to 0 and high purple equal 200+ see Figure 5

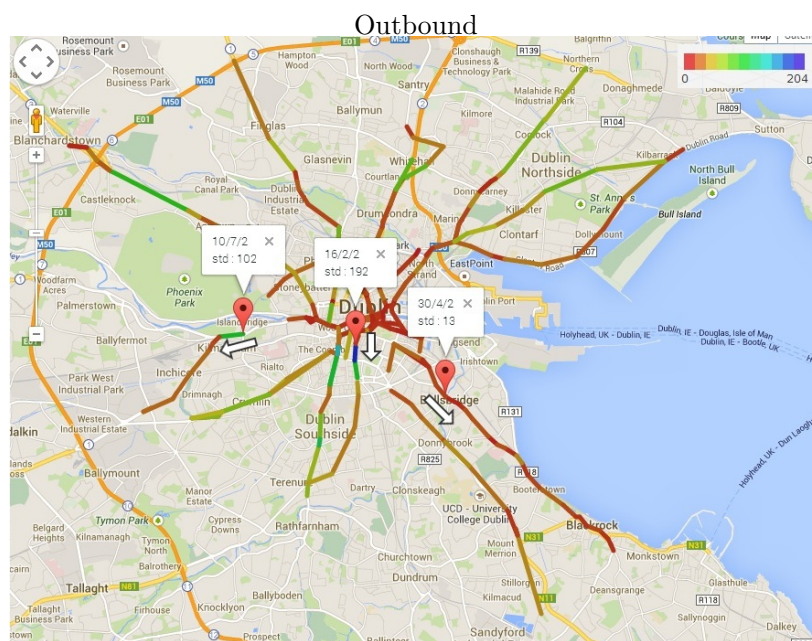
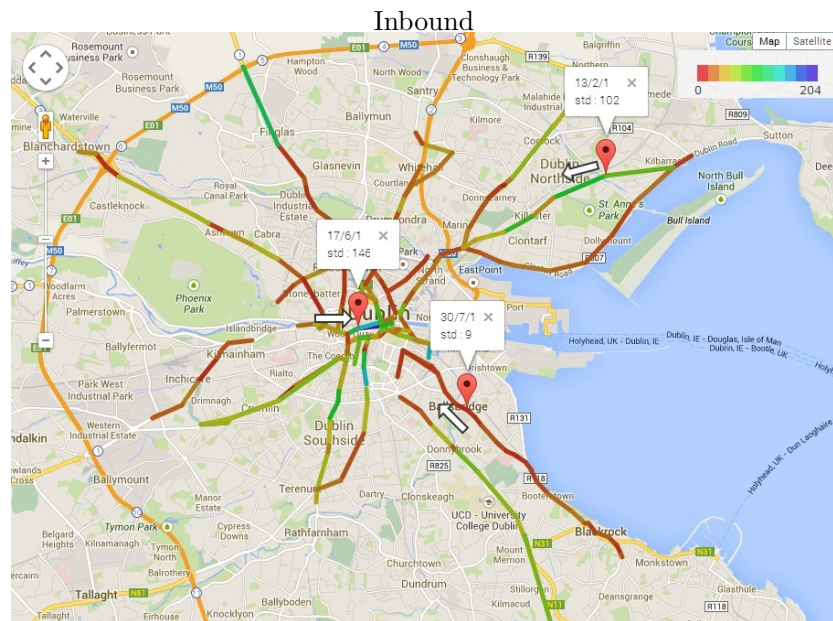


Figure 5: Inbound and Outbound Traffic Observations



### 3.2 Weather

The three weather stations selected for correlation analyses was performed (see section 2). The visualisation demonstrates there is a spatial relationship between the observing weather station. The triangles in Figure 6 represent weather stations. The size circles represent correlation values from -1 to 1, for example, zero correlation is not visible on the map. The circle uses transparency as the negative correlation indicator.

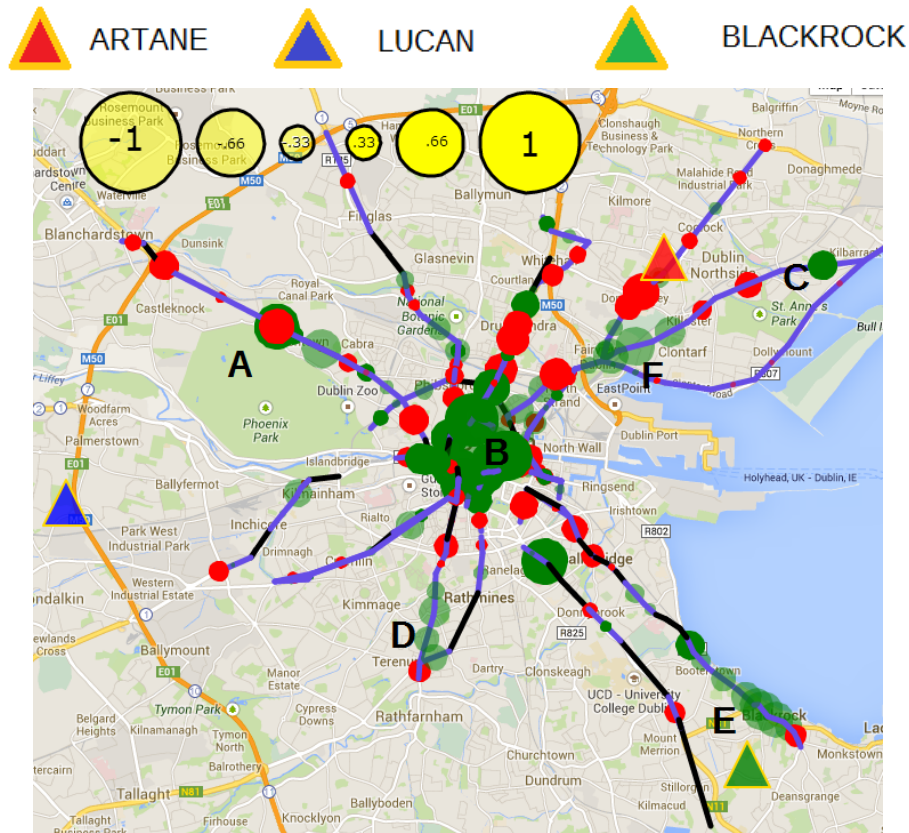


Figure 6: Correlation of Temperature Map Direction Inbound Peak Times

### 3.3 Prediction Model

During the exploration process, the highest correlated weather stations and spatial neighbour were used in the generic estimation data model. A generic data model provides the ability to reuse the process of building the data sets that would best fit the majority of the 512 observed locations while still keeping features that improve accuracy. As a result, the prediction algorithms behaved differently depending on the influence of features.



The best-performing algorithms for the least volatile road segments mentioned ?? are linear regression. Some road segments had little or no volatility. Other linear regressions, performed well that had volatility used a normalisation of feature to improve accuracy.

Road segments with highly volatility with features of insignificant correlation resulted in a non-linear Support Vector Machine with Fourier transform with the highest accuracy.

Bayesian Ridge linear regression algorithm performed very well for the prediction. It demonstrates that when noise accounts for the more linear the data becomes.

In figures, ?? and ?? shows that the area of Finglas and Glasnevin is the least affected by weather and is highly volatile. Where the city centre and Clontarf are volatile and highly affected by weather conditions becomes a linear problem, see figure 7.

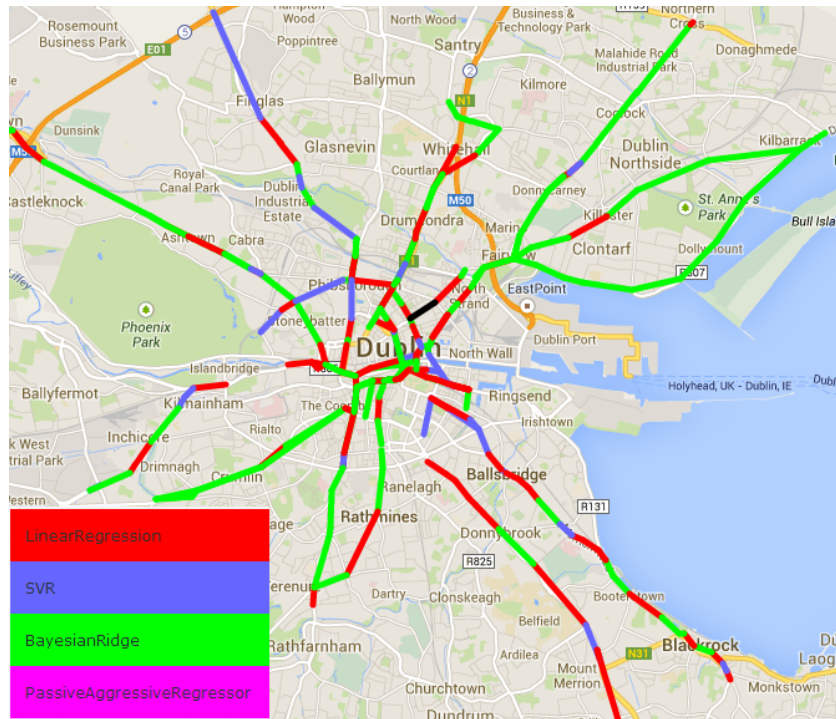


Figure 7: Off-peak and Inbound Algorithm Map

### 3.4 Twitter

The objective of twitter section is to analyse the approach of using the traffic domain tweets to extract tweets from real-time data that is related to the traffic domain. The tweets from the traffic domain do not contain geospatial information compared to the real-time tweet. Passive Aggressive Classifier is an example of one of the algorithms used to classify the real-time traffic tweets. Using

Tf-Idf scoring and Vectorizer algorithm. tweets are tokenised to fit the predictive model for the classifier. A sample of results fit on Table 1.

Table 1: Classified Tweets

Result	Text
True Positive	No better way to start your day with a car crash, and then forgetting about the banana in my pocket going through security...
False Positive	We're gonna crash vine if we keep doing this
False Positive	Lyndsay Lohan looks like a car crash.... She is wrote off #ChattyMan
True Positive	I bloody hate waiting #delays <a href="http://t.co/Yh55PrfQK3">http://t.co/Yh55PrfQK3</a>
True Positive	There's after been a crash outside my estate, 3 fire trucks and 3 ambulances

## 4 Results

The classification approach worked as a proof of concept. The real-time traffic tweet could be used to provide further analysis on traffic delays. In Figure 8 the dashboard demonstrates traffic related tweets as blue markers overlayed above the road segments and its estimated result. The red lines indicate delays, the green indicate better than expected while the grey is as expected. Each tweet marker is click-able to provide more informative details on the traffic conditions. Using the buttons on the left of the dashboard will display the different elements of the visualisations show in this abstract.

Using NoSQL to overcome the challenges of the four V's the approach stored volumes of data that on a single machine RDMS system would have been problematic 2.

Table 2: Data Volume

<i>Data Source</i>	<i>Items</i>	<i>No. of Documents</i>
Traffic Observations	501,402,840	8,356,714
Real-time Tweets	3,048,310	116
User Tweets	5,267	5,267
Weather Records	229,311	2,103

Issues such as in figure 8 the dashboard contains a some false positives, example "*Lyndey Lohan looks like a car crash.. she is wrote off #ChattyMan*" while true positive "*We hope our new display doesn't cause too many delays in Donnybrook .... <http://t.co/sDKrey1pJf>*" can be resolved in future work.

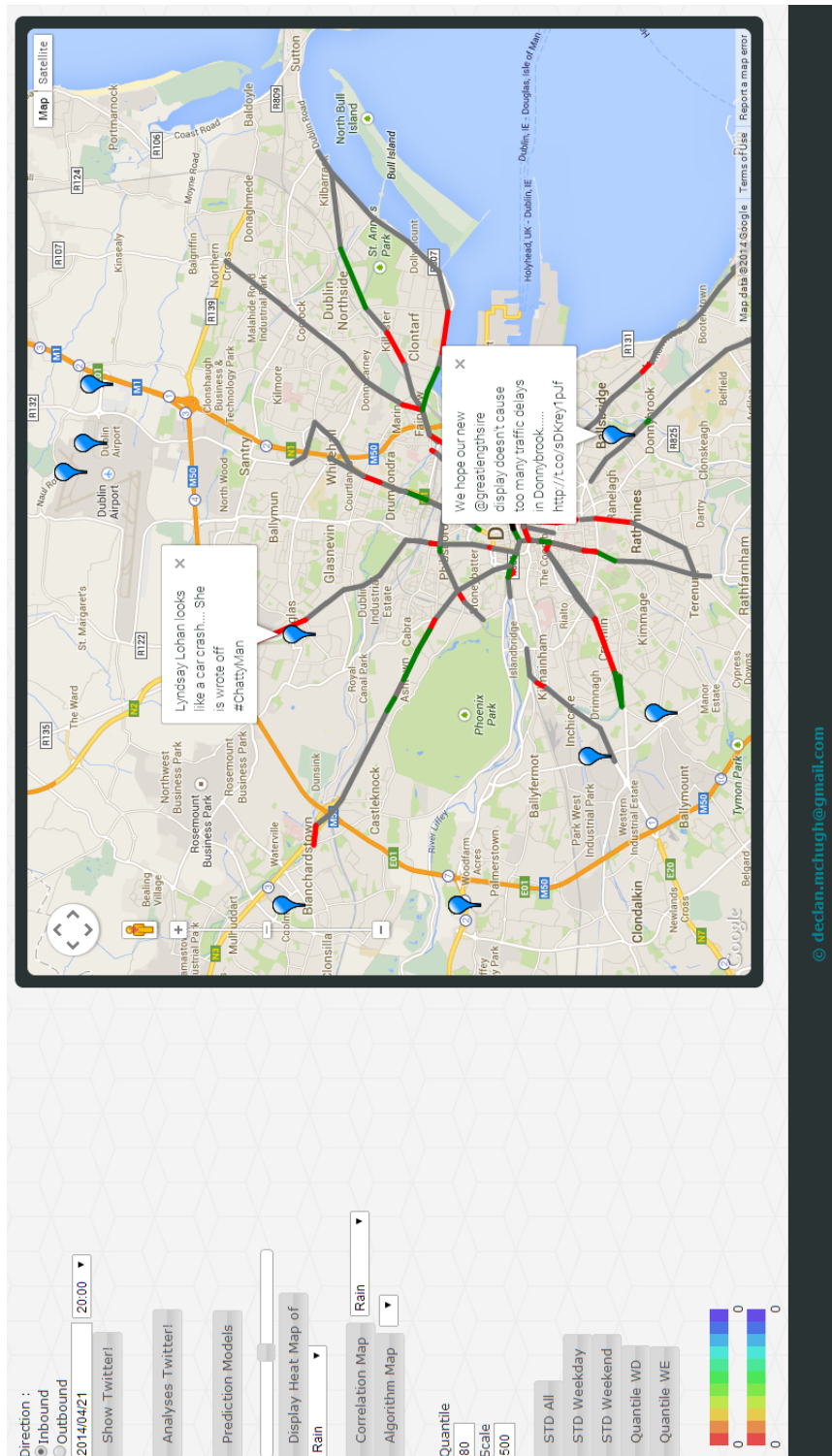


Figure 8: Dashboard Analysis for 21/04/2014 8pm to 9pm

## References

DubLinked (2012-2014). Trips data.

Sheth, A. (2014). Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies.

Tulloch, D. J. (2012). A garch analysis of the determinants of increased volatility of returns in the european energy utilities sector since liberalisation. *IEEE*.

Twitter (2014). Twitter.

Wunderground (2012 - 2014). Wunderground.

## 5 Acknowledgements

None of this work could have been achieved without data from Twitter (2014), DubLinked (2014) and Wunderground (2014)

## 6 Biography

Declan McHugh is a 12-year veteran of writing software. He has a broad exposure to developing in many technologies and is a keen advocate in all things analytical. Declan has recently obtained a 1st class honours Masters the Analytics is actively working projects with Big Data Analytics and Visualisation.

# Beyond Visualisation in 3D GIS

James Milner<sup>\*1</sup>, Kelvin Wong<sup>†1</sup> and Claire Ellul<sup>‡2</sup>

Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK, WC1E 6BT

November 02, 2014

## Summary

Although 3D visualisation is becoming more common in GIS, as of yet, there has been relatively little in the way of 3D editing and analysis functionality especially in the web. This research describes a first attempt at addressing this deficit, documenting a 3D Web GIS with the ability to select, edit, 3D buffer, measure and retrieve attributes. A small user evaluation was undertaken to assess aspects such as usability, consistency and responsiveness. The system developed was implemented using Three.js as a frontend 3D framework and PostGIS as a backend database. The GIS was successful in its execution but detected some issues in requirement of addressing in order to progress. It concludes with recommendations to improve performance and go further with 3D editing.

**KEYWORDS:** 3D, Web, GIS, WebGL, Editing

## 1. Introduction

The progression of 3D geographic information systems (GISs) has perhaps never been more necessary. The ever increasing demand for 3D GISs and 3D geographical data has been acknowledged (Gröger and Plümer, 2012) and recent research has evidenced real-world applications for 3D GIS such as visualisation of dynamic urban landscapes (Yin & Shiode, 2014), and mineral resource exploration (Wang et al., 2014). The primary aim of this research was to develop, implement and document fundamental functionality such as visualising, selecting, editing, measuring, 3D buffering and attribute revival in a 3D web GIS.

## 2. Background

2D GISs struggle to represent the three-dimensional world, unable to express multiple height values at one 2D point (Abdul-Rahman and Pilouk, 2008). 3D GISs are needed that are able to visualise, capture, structure, manipulate and analyse 3D data (Abdul-Rahman and Pilouk, 2008). Many 3D systems currently lag behind 2D oriented systems (Ellul and Haklay, 2009). This is even more prominent in browser based 3D web GIS (Ming, 2008), with most modern 3D web GISs (e.g. Esri CityEngine Web Viewer) focusing predominantly on visualisation and/or navigation, and not editing and analysis. Recent technologies such as HTML5 and WebGL, have made the possibility of 3D GIS natively in the browser a real scenario, however they present new problems such as handling various coordinate systems in these environments.

## 3. Data

The research uses 3D data for Newcastle (Table 1 and Figure 1) which was provided in Esri Shapefile format by Ordnance Survey. All layers were 3D ( $x, y, z$ ), with the data being originally derived by the

---

<sup>\*</sup> james.milner.13@ucl.ac.uk

<sup>†</sup> kelvin.wong.11@ucl.ac.uk

<sup>‡</sup> c.ellul@ucl.ac.uk

OS from aerial imagery using photogrammetric techniques. The ‘Walls’, ‘Roofs’ and ‘Bridges’ layers are at LoD2 with the ‘Iconic\_Buildings’ layer at LoD2 with some LoD3 elements.

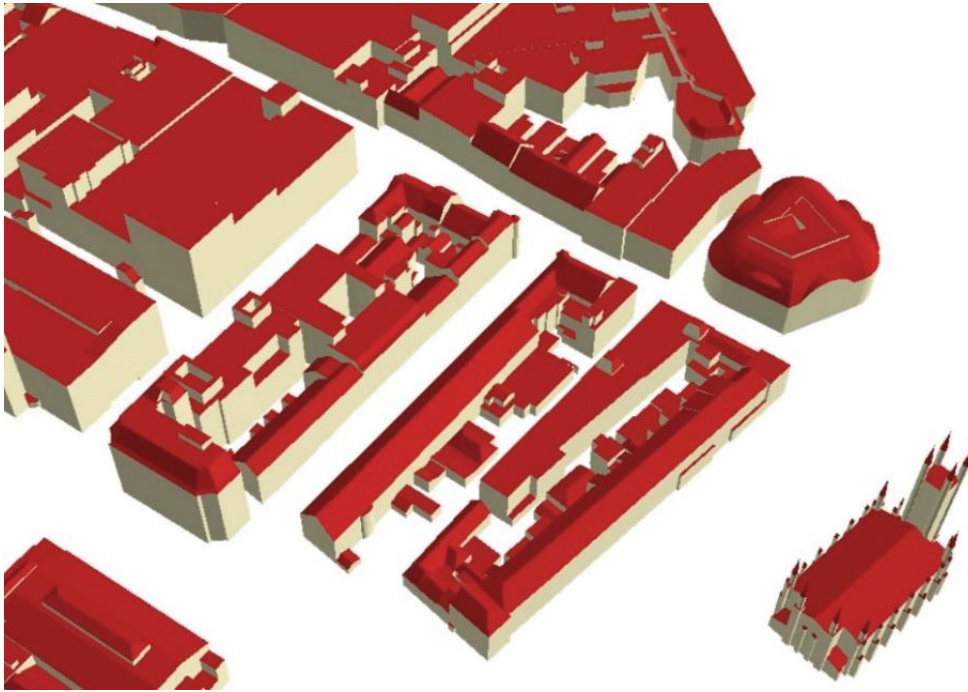
**Table 1** Feature Details for the Newcastle Dataset

Feature Name	Feature Geometry Type	Number of Features	Shapefile Size (kilobytes)
HeightedITNRoadNode	Point	1470	64
HeightedITNRoadLink	Line	2086	762
Vegetation	MultiPatch	139	10,476
HeightedOSMM_TopoLine	Line	8998	2,008
HeightedOSMM_TopoArea	Polygon	21469	22,506
Walls	MultiPatch	300	2,830
Roofs	MultiPatch	300	1,968
SimpleBuildingHeights	Polygon	5080	2,547
Bridges	MultiPatch	3	132
Iconic_Buildings	MultiPatch	4	1,784



**Figure 1** The Newcastle dataset layers displayed in ESRI's ArcScene 10.2.1





**Figure 2** Close-up of LoD2 buildings, consisting of wall and roof layers

Data was translated using SafeSoftware's 'FME Data Inspector' to a PostgreSQL 9.2.0 database using the PostGIS extension. Shapefile types were converted to the equivalent PostGIS Well-Known Text (WKT) primitives (i.e. Points into WKT Point Zs, Multipatch Shapefiles into WKT PolyhedralSurfaces and TIN Zs, etc). Each table (data layer) was updated with an ID column using a Big Serial primary key allowing for auto-incrementation. This allows the 3D GIS web system to uniquely identify geometries.

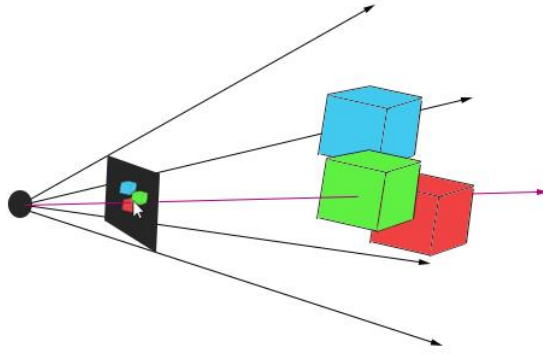
#### 4. Methodology

The systems layout was modelled on popular GISs such as ArcScene and QGIS to retain familiarity and client side technology such as HTML5, CSS, JavaScript and WebGL (through Three.js). The system is underpinned by a PostGIS database, and geometries (WKT) are brought into the scene from the database using PHP and Structured Query Language (SQL).

The WKT strings are converted into Three.js geometries in the scene. Some Three.js primitive geometries were used; for points `THREE.SphereGeometry` were used, for lines the `THREE.Line` primitive. For WKT TINZs and PolyhedralSurfaces custom Three.js geometries were created. The project makes use of Seidel's (1991) Polygon Near Linear (PNL) 2D triangulator implemented in JavaScript by Ahting (2014) for triangulation of arbitrary geometries. To overcome its 2D nature a micrometric epsilon value was used to distinguish between 3D points sharing  $X, Y$  values ( $0.0005\text{ m}$ ).

Single selection was implemented into the system through use of raycasting; firing a ray into the scene to determine if objects exist under a mouse click (Figure 3).





**Figure 3** Raycasting (purple arrow) into the scene from the camera in Three.js

Multi-selection used a marquee selection methodology. The marquee is variable sized HTML div element controlled by the mouse. Geometries in the field of view frustum are iterated through, with vertices unprojected to 2D coordinate screen space and checked to see if they are in the defined marquee. Attribute querying was implemented by taking selected object IDs (stored in an array) and passing these to the backend PostGIS database to pull out attributes and push them into a formatted table.

Point distance measurement was implemented using a simple formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (1)$$

Where  $(x_1, y_1, z_1)$  is the first supplied coordinate, and  $(x_2, y_2, z_2)$  is the second. All geometries are constructed from triangles in WebGL, and as such surface area measurement can be determined using Heron's Formula:

$$s = \frac{a + b + c}{2}$$

$$A = \sqrt{s(s-a)(s-b)(s-c)} \quad (2)$$

Where  $a, b, c$  are the lengths of the triangles sides,  $s$  is its semi-perimeter and  $A$  is its area.

Multiple 3D buffering options were implemented through use of three of Three.js's primitives Box, Sphere and Cylinder. Transparency and dimensions are optional parameters.

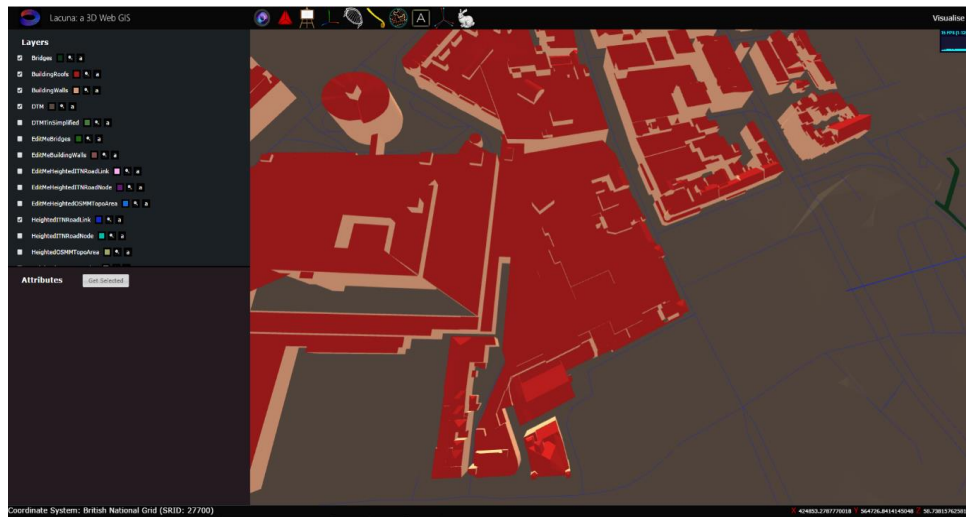
Object level editing such as deletion, copying, translation, scaling and rotation were implemented by use of custom functions and built-in Three.js methods. Geometries could not be transformed in place with built-in Three.js transformation functions and needed to be brought to the system origin  $(x, y, z = 0, 0, 0)$  by subtracting the objects center point. Edited geometries are then transformed and returned to their previous position and synced to the database using PHP/SQL.

Vertex level editing was possible through wireframing geometries, and a 'closest vertex to click'

function. This allowed for editing and updating of specific vertices. An UPDATE command is sent using PHP/SQL to update the associated geometries in the database (as WKT)

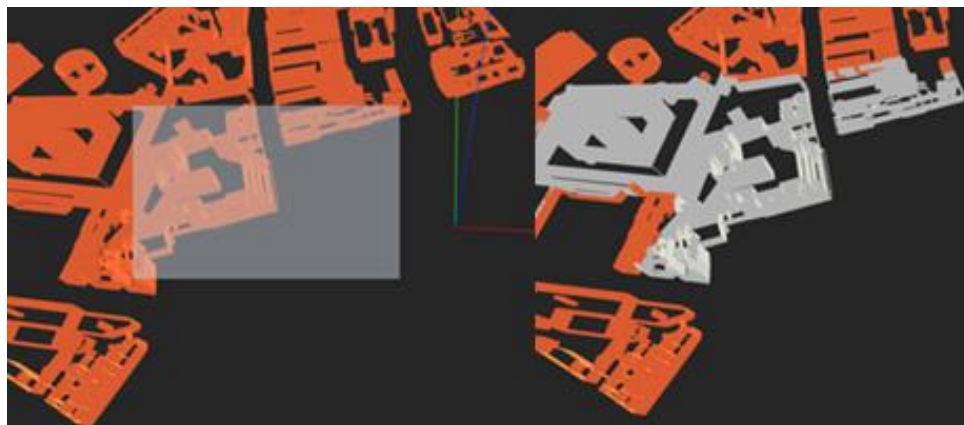
A short user evaluation was also conducted to find the strengths and weaknesses of the developed system. It consisted of a 'Think-out-loud protocol' assessment, three qualitative questions and ten quantitative, with ten participants.

## 5. Results



**Figure 4** The system displaying various geometries

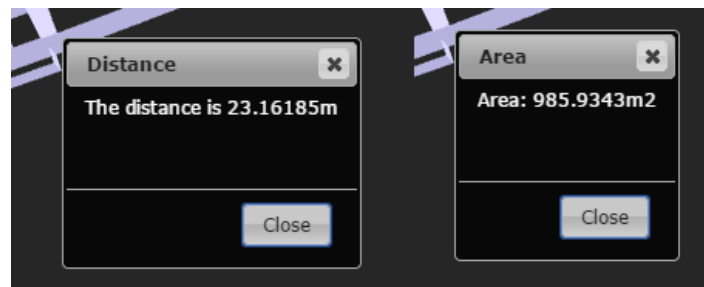
Figure 4 displays the design of the system, retaining familiarity from the user interface of popular desktop GISs (i.e. ArcScene, QGIS). It also shows the visualisation of 3D geometries into the scene.



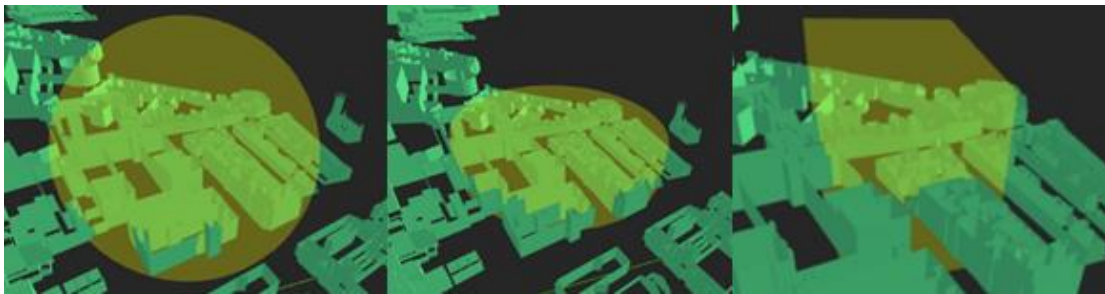
**Figure 5** Marquee selection and resulting selection

Figure 5 expresses the marquee selection tool, showing how multiple geometries can be

selected.

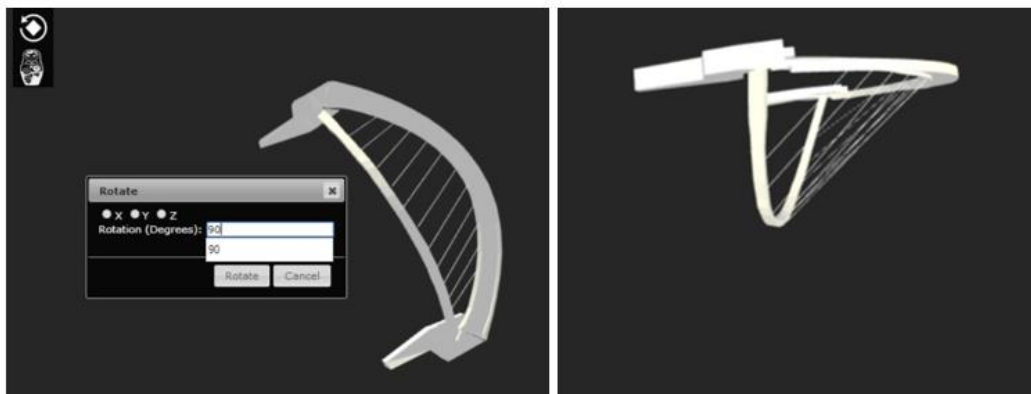


**Figure 6** Distance and Area operation results

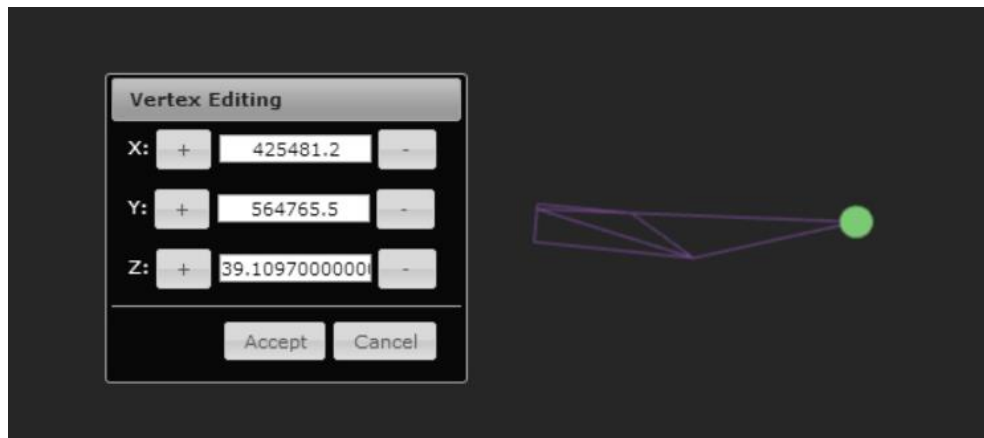


**Figure 7** A spherical, cylindrical and box buffer in scenes

Outputs from the calculations from the formulas presented in distance and area function can be seen in Figure 6. Results from all three buffer operations can be seen in Figure 7. An example transformation (rotation) can be seen in Figure 8 alongside the vertex level editing interface in Figure 9.

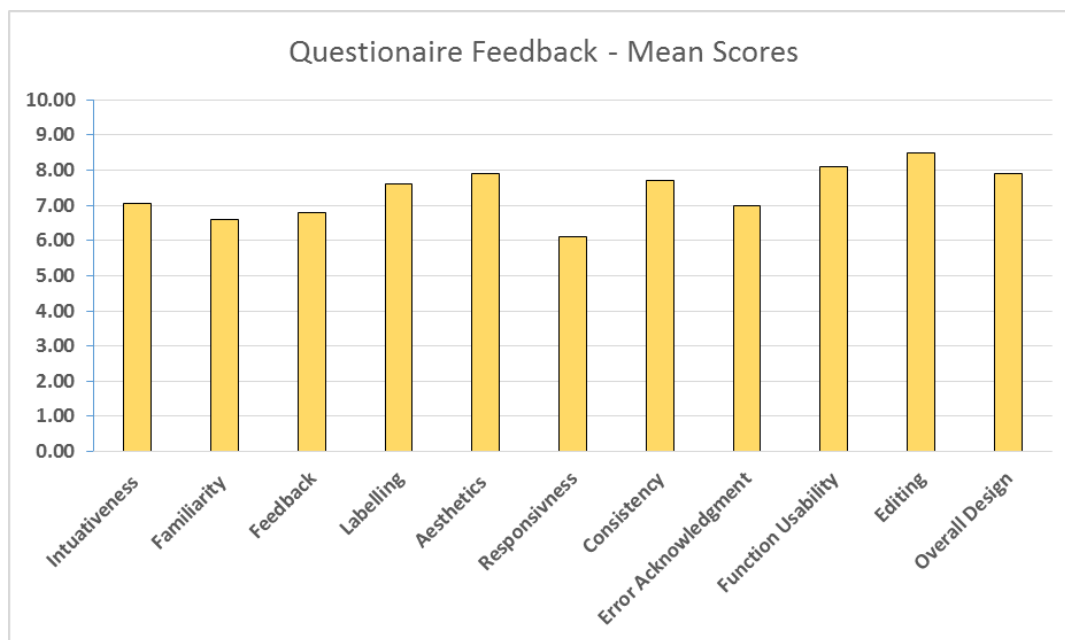


**Figure 8** Geometry rotation in the scene



**Figure 9** Vertex editing with interface

Figure 10 shows the user study results. Responsiveness and was the lowest (6.1) rated factor and editing was the highest rated (8.5).



**Figure 10** User study results

## 6. Discussion

The system developed has shown how it is possible to produce an elementary plugin-free 3D GIS, with editing tools in the web. The projects aims were achieved: implement and document visualisation, selection, attribute retrieval, measurement and 3D buffering, alongside more complex operations such as object level and vertex level editing.

The user study proved that 3D web GIS has promising potential, with favour for 3D GISs that follow similar designs to popular GIS desktop applications. Users showed an overall approval of the 3D system

implemented in both quantitative assessment (7.39/10 mean score) and think-out-loud protocol. Perhaps the most important contributions are in fields of selection and editing, although other elements of the GIS may be foundations for successive work. The project has also highlighted the applicability of WebGL/Three.js for the platform of a 3D web GIS.

The research highlights barriers for successful web GIS including a robust and efficient 3D triangulator for arbitrary 3D geometries and layer loading times for complex geometry sets (expressed in several think-out-loud protocol assessments). The latter is a critical issue on low bandwidth and low specification devices. Furthermore, the research exposes that although suitable for storage and retrieval, 3D support in PostGIS requires enhancement in order to be effective in the functionality of a 3D web GIS.

## 7. Conclusion and Further Work

This system shows that 3D Web GIS is possible directly in the browser using no additional plugins. The benefit of this approach is the potential for cross platform and device web GIS, built on accessible open technologies. The web GIS explorers initial steps, highlighting the potential for what could be possible with further work.

It is recommended that performance is an area of focus for future research. Use of PostgreSQL's *string\_agg(expression, delimiter)* and PostGIS's *ST\_AsGeoJSON* may yield improved results (as per Ellul and Altenbucher, 2014). Alternative databases with JSON support (e.g. MongoDB), client side string manipulation, and generalisation/simplification could all possibly improve transfer and rendering speeds.

Further editing capabilities such as face and edge splitting, and construction from base extrusion or primitives/constructive solid geometry might be further directions for research. The system could also benefit from integration with a topological model, allowing for topological querying and ensuring topological consistency by defined rules. Finally further spatial analysis tools such as flood analysis, 3D viewsheds and hotspot analysis would push the functionality of the 3D web GIS.

## 8. Acknowledgements

This project was funded and supported by the Ordnance Survey. The Ordnance Survey also made their 3D datasets available for use with this project.

## 9. Biography

James Milner studied a MSc. in Geographic Information Science at University College London. For his dissertation he researched the development of a 3D Web GIS built on WebGL. He currently works for Esri UK as a developer evangelist.

Kelvin Wong is an EngD research engineer at the UCL Centre for Virtual Environments, Interaction and Visualisation, Department of Computer Science, University College London. His research interests focuses on the challenges of deploying 3D geographic datasets at a national level with particular interests in usability, applications and data quality of 3D geographic information. Additional research relates to 3D visualisations and 3D requirements gathering.

Claire Ellul is a Lecturer in Geographical Information Science at University College London. Prior to starting her PhD, she spent 10 years as a GIS consultant in the UK and overseas, and now carried out research into the usability of 3D GIS and 3D GIS/BIM integration. She is the founder and current chair of the Association of Geographical Information's 3D Specialist Interest Group.

## References

- Abdul-Rahman, A. and Pillouk, M., 2008. *Spatial Data Modelling for 3D GIS*. New York: Springer
- Ahting, J., 2014. Polygon Near-Linear Triangulation in JavaScript [Online]. Accessible form: <https://github.com/jahting/pnltri.js> [Accessed: 20/08/2014]
- Ellul, C. and Altenbuchner, J., 2014. Investigating approaches to improving rendering performance of 3D city models on mobile devices, *Geo-spatial Information Science*, 17(2), 73-84.
- Ellul, C. and Haklay, M., 2006. Requirements for Topology in 3D GIS. *Transactions in GIS*, 10(2), 157–175.
- Gröger, G. and Plümer L., 2012. CityGML – Interoperable semantic 3D city models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 71, 12–33.
- Ming, W., 2008. A 3D Web GIS System Based on VRML and X3D. In: *Proceedings of 2nd International Conference of Genetic and Evolutionary Computing*, WGECC 2008, September 25–26, 2008, Hubei, 197–200.
- Seidel, R., 1991. A simple and fast incremental randomized algorithm for computing trapezoidal decompositions and for triangulating polygons. *Computational Geometry Theory & Applications*, 1(1), 51-64.
- Wang, G., Zhang, S., Yan, C., Song, Y., Qu, J., Zhu, Y., Li, D., 2014. 3D-GIS Analysis for Mineral Resources Exploration in Luanchuan, China. In: Pardo-Igúzquiza, E. Guardiola-Albert, C., Heredia, J., Moreno-Merino, L., Durán, J., Vargas-Guzmán J., eds. *Mathematics of Planet Earth: Proceedings of the 15th Annual Conference of the International Association for Mathematical Geosciences*. Berlin: Springer: 295-298.
- Yin, L. and Shiode, N., 2014. 3D spatial-temporal GIS modeling of urban environments to support design and planning processes. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 7(2), 152-169.

# Visualize and interactively design weight matrices

Angelos Mimis<sup>\*1</sup>

<sup>1</sup>Department of Economic and Regional Development,  
Panteion University of Athens, Greece  
Tel.: +30 6936670414

October 29, 2014

## Summary

A GIS tool that permits to visualize, explore and interactively modify weight matrices is described. Weight matrices, created in various formats, can be imported and the spatial relationship, by using polylines, can be visualized. Scripts are developed to explore the structure of the weight matrix by illustrating basic statistics, to illustrate the full matrix and to compare different matrices. The spatial relationship can then be modified (by deleting or adding polylines) and exported in order to further use it in computations. The extension is developed in Python, is based in PySAL and matplotlib libraries and is implemented in ArcGIS.

**KEYWORDS:** weight matrix, interactive design, ArcGIS extension.

## 1. Introduction

The use of weight matrices is central in spatial analysis. They are used in the definition of segregation indices (Wong, 1993), in spatial autocorrelation (Anselin, 1995), in spatial econometric models (Anselin, 2010) and in network analysis (Barthélemy, 2011). Over the years, the majority of the research have focused on the philosophy captured in the weight matrices (Harris et al., 2011), the different definitions (e.g. theoretical topological or empirical as described in Getis, 2009) and on the effect these have on the evaluated results (Stakhovych and Bijmolt, 2008). On the other hand, little effort has been put on visualizing and interactively design the weight matrices. By visualize, you can map and explore the relationship between neighbouring points or areas intuitively without having to employ complex coding schemes. In that direction, Bivand et al. (2008), create a graph of neighbours in order to illustrate the polygon contiguities.

My approach adopts that idea and extends it, to not only visualize but to explore and modify the spatial relationship or even design it from scratch. In order to demonstrate this approach, an extension in the commercial package ArcGIS has been developed (same code could be used in an open source platform e.g. QGIS) and it is based on two freely available libraries. The first one is the PySAL library (Rey and Anselin, 2010) of spatial analysis and the second is the matplotlib (Hunter, 2007) plotting library. So by using the scripts, one can import many of the formats created in the most popular spatial software (e.g. GeoDA, Matlab). Further, exploratory analysis can be performed by displaying basic statistics of the weight matrix, capture in an image the sparseness of the full matrix, compare different weight matrices and visualize the linkage between neighbouring areas or points. Having explored the given weight matrix, one can proceed by modifying (deleting or adding) the linkages (polylines) between neighbouring entities. Finally, the relationship produced (weight matrix) can be exported in any format supported by the script and consequently used in ArcGIS or any other

---

\* mimis@panteion.gr

spatial software using weight matrices.

This short paper starts by describing the ArcGIS extension, gives a realistic example of importing and altering the weight matrix of the Greek prefectures and concludes by discussing future improvements.

## **2. Program description**

The weight matrix tool is implemented as an extension to ArcGIS using Python programming language. It is based on the python's libraries PySAL and matplotlib. The toolkit is organized into a) importing and exploratory functions and into b) functions permitting to design and export the weight matrix.

### **2.1. Importing and exploratory analysis**

All the scripts are designed to import and export weight matrices created in the binary form of ArcGIS (swm), in contiguity (gal) and distance (gwt) based form of GeoDa and in Lesage's library form (dat) of MATLAB (Lesage and Pace, 2009). It should be noted that based on these formats one can import/export weight matrices in other software such as the R statistical software or Stata.

By importing a weight matrix, the elements of the matrix are visualized by creating links between the areal or point data that have got a connection. This creates an optical realization of the spatial relationship of the data and by using the capabilities of the GIS one can inspect that relationship in various scales. Further in order to be able to export it in one of the supported formats, the script ensures that the start and the end point of the link are within the relevant polygons. In the case of a multipart object, the link start or ends within the area of the polygon with the biggest area. Finally convex as well non-convex polygons can be treated.

As far as the exploratory part is concerned, one can generate basic statistics, graphically displays the non-zero elements of the full weight matrix and can compare two different weight matrices, as will be shown in the example application.

### **2.2. Design and export**

The exploratory procedures described above might be followed by changes in the relationship of the data and export it in one of the weight formats supported. The layer keeping the links between polygon or point data is a polyline layer and so it can be altered in the usual way done in a GIS environment. So one can easily add a new polyline segment, remove a segment and alter the weights in the attribute table. When the designing part is finished, the matrix can be exported by using the appropriate script which permits the user to enforce symmetry and standardization.

The toolkit requires a license of ArcGIS 10 and was tested in ArcGIS 10.2, by using PySAL 1.7 and matplotlib 1.3.

## **3. Example application**

An example use of the GIS tool will be given for the prefectures of Greece (NUTS 3 level). A distance based weight matrix is created in GeoDA, in gwt form (with a given threshold). Then the weight matrix is visualized in ArcGIS by using the toolbox (Figure 1). This results in a polyline shapefile (Figure 2) having an attribute table (Figure 3) with the actual weights and the ids of the corresponding neighbours. So for example the polygon with id 30 (part of island Crete, south in the map of Greece) is considered as a neighbour of the polygons 10 and 40.



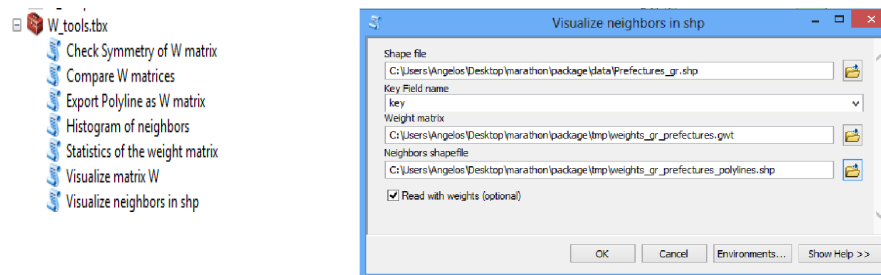


Figure 1. The weight matrix toolbox (left) and the “visualize neighbours in shp” menu script (right).

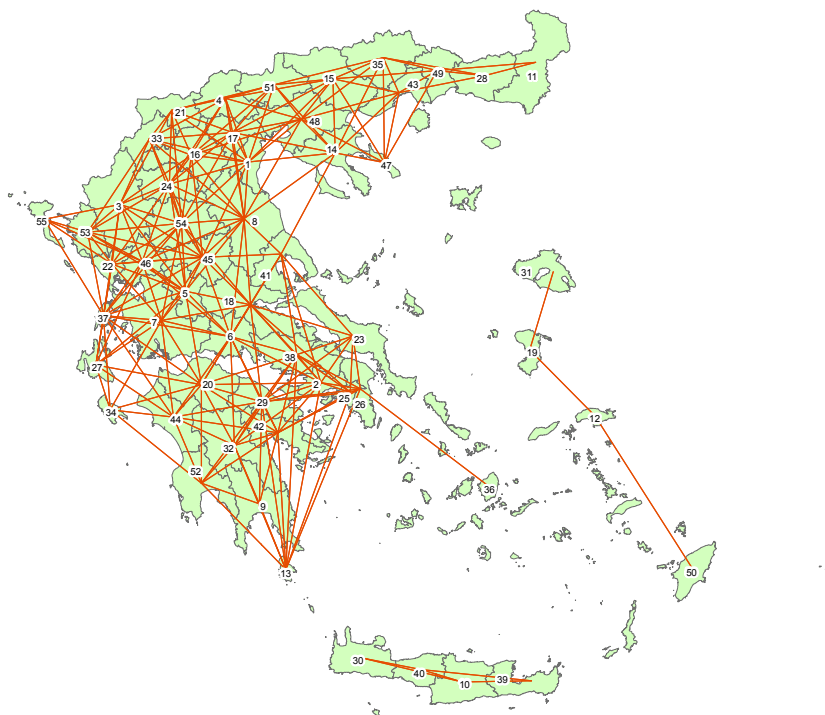


Figure 2. The weight matrix visualized.

id	Shape	n_id	n_id	w
0	Polyline	39	10	115589,838
1	Polyline	30	40	89528,957
2	Polyline	28	11	55176,349
3	Polyline	28	43	92116,8782
4	Polyline	28	35	112709,083
5	Polyline	28	49	55432,1299
6	Polyline	29	25	84402,0081
7	Polyline	29	13	82330,1325
8	Polyline	29	38	51815,1873
9	Polyline	29	20	84661,753
10	Polyline	29	23	122032,321
11	Polyline	29	18	100632,841
12	Polyline	29	44	109347,153
13	Polyline	29	42	45767,2742
14	Polyline	29	26	97380,908
15	Polyline	29	52	131311,792
16	Polyline	29	2	53509,3656
17	Polyline	29	6	78853,0893
18	Polyline	29	9	128568,551
19	Polyline	29	32	66295,7692
20	Polyline	35	15	59079,4380
21	Polyline	35	14	124261,724
22	Polyline	35	48	115252,107
23	Polyline	35	49	57495,9791
24	Polyline	35	47	111467,772
25	Polyline	35	51	123375,041
26	Polyline	35	43	52920,2318
27	Polyline	35	28	112789,983
28	Polyline	34	27	70444,8712
29	Polyline	34	20	100872,781
30	Polyline	34	44	67876,7098
31	Polyline	34	37	107835,663
32	Polyline	34	52	126285,424
33	Polyline	34	7	104500,357
34	Polyline	24	21	80700,2644
35	Polyline	24	17	89587,8533
36	Polyline	24	16	39704,785
37	Polyline	24	33	54121,843
38	Polyline	24	54	44812,1474
39	Polyline	24	45	87776,4553
40	Polyline	24	22	110480,345

Figure 3. The attribute table of the polyline shapefile created.

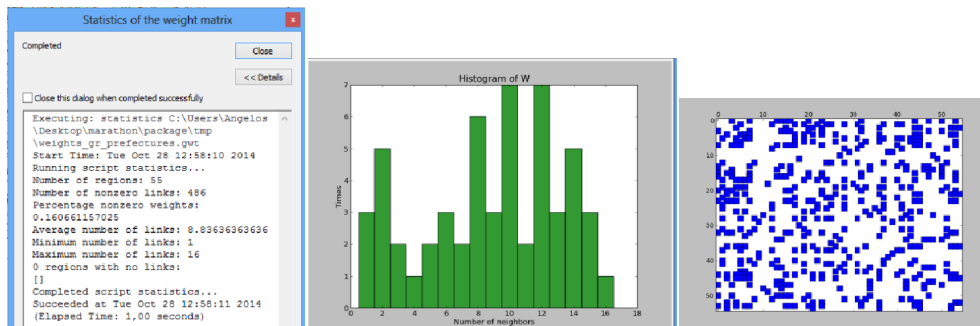


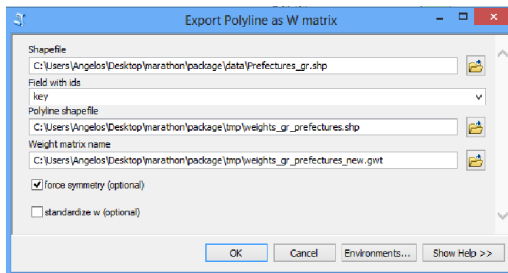
Figure 4. Basic statistics of the weight matrix.

So by using the functionality of the GIS, one can explore the spatial relationship of prefectures captured in the shapefile. Further, by using the statistics scripts, the percentage of nonzero weight (16%), the average number of links (8.8), the existence of islands etc as well as the histogram of the number of neighbours and the image of the nonzero elements of the full matrix is displayed (Figure 4).

Finally, one can edit the polyline shapefile and thus changing the neighbouring relationships. So for example one can delete the link 12-50 and add the new links 36-50, 36-12 and 36-30. Then you can export the polyline shapefile in one of the supported formats by keeping the spatial relationship and the relevant weights (Figures 5 and 6).



Figure 5. The modified weight matrix.



```

35 15 59679.4
35 14 124282
35 48 115252
35 49 57496
35 47 111458
35 51 123375
35 43 52920.2
35 28 112789
36 26 136332
36 12 139220
36 50 241363
36 30 233516
37 27 36631
37 20 118603
37 22 59919.4
37 55 127733
37 54 130876
37 44 125559
37 45 121507
37 53 104075
37 34 107836
37 46 58931.8
37 3 127107
37 5 87535.9
37 7 50704.6
37 6 133629

```

Figure 6. The export menu (left) and part of the gwt file (right).

#### 4. Conclusions

A GIS tool that permits to visualize, explore and interactively modify weight matrices has been illustrated. This can be used to create a weight matrix from scratch or modify an existing matrix created in a supported format. This toolbox can also be used as an educational interactive utility.

The tool is implemented as an extension in ArcGIS and several improvements can be made. Every time a script is used, the user should import the weight matrix. One can overcome this limitation by designing a separate tool incorporating the code and resulting in faster computations since the weight matrix will have to be read only once.

## 5. Biography

Angelos Mimis is an assistant professor of spatial analysis in Panteion University of Athens, Greece. His interests include GIS, spatial analysis, computational geometry and optimization. He teaches GIS and spatial analysis in undergraduate and postgraduate level. He is visiting the Geography department of Bristol University in the summer semester of 2015.

## References

- Anselin L (1995). Local indicators of spatial associations – LISA. *Geographical Analysis*, 27, 93-115.
- Anselin L (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1), 3–25.
- Barthélemy M (2011). Spatial Networks. *Physics Reports*, 499, 1–101.
- Bivand R S Pebesma E J and Gomez-Rubio V (2008). *Applied Spatial Data Analysis with R*. Springer, New York.
- Getis A (2009). Spatial weight matrices. *Geographical Analysis*, 41, 404-410.
- Harris R Moffst J and Kravtsova V (2011). In search of ‘W’. *Spatial Economic Analysis*, 6(3), 249-270.
- Hunter J D (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3), 90-95.
- Lesage J and Pace R K (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC, Boca Raton.
- Rey S J Anselin L (2010). PySAL: a python library of spatial analytical methods, In: Fischer, M. M., Getis, A. (Ed.), *Handbook of Applied Spatial Analysis*, Springer Berlin Heidelberg, pp. 175-193.
- Stakhovych S and Bijmolt T H A (2008). Specification of spatial models: A simulation study on weight matrices. *Papers in Regional Science*, 88(2), 389-408.
- Wong D W S (1993). Spatial Indices of Segregation. *Urban Studies*, 30, 559-572.

# Better Census statistics for Civil Parishes – When “best-fitting” just isn’t good enough

Bruce Mitchell<sup>1</sup>

Geography Research Unit

Geography Branch

Research Development and Infrastructure Directorate

Office for National Statistics

April-20<sup>th</sup> 2015

## Summary

The Government Statistical Service (GSS) [National Statistics Geography Policy](#) published by the Office for National Statistics (ONS) defines the way that statistics for any geography larger than Census Output Area should be generated. For Civil Parishes in England this ‘best fit’ method is not satisfactory.

ONS is therefore examining the relative merits of other methods of aggregating Census data to parishes, especially with reference to grid cells. The outcome of the research will inform a decision on which method to mandate for the 2021 Census.

**KEYWORDS:** Census, GSS, ONS, Civil Parish, Parish Council, Eurostat, Geostat, grid, best-fit, best-fitting

---

<sup>1</sup> [bruce.mitchell@ons.gov.uk](mailto:bruce.mitchell@ons.gov.uk)

## 1. Introduction

The Office for National Statistics (ONS) is the national statistical institute of the UK. ONS publishes statistics related to the economy, population and society of England and Wales at national, regional and local levels. ONS is independent of ministers, but data produced by the office, including a decennial Census (most recently in 2011), inform policies, priorities and allocation of resources.

ONS Geography is responsible for defining the geographical data referencing framework for the *Government Statistical Service's (GSS) Geography Policy* and providing the tools and data that underlie it via a dedicated [Open Geography Portal](#).

This paper details research into providing statistics for Civil Parishes in England using an INSPIRE-compliant<sup>2</sup> 1 km grid supplied by Eurostat. This would represent an exception to the Government Statistical Service (GSS) National Statistics Geography Policy (ONS, 2010, 2015) and a unique departure for the provision of Census data, which is supplied using the 'best fit' methodology based on Output Area population-weighted centroids (OA-PWC).

Civil Parishes are one of the most commonly-requested geographies on the Neighbourhood Statistics website<sup>3</sup>.

## 2. 'Exact fit' and 'best fit'

ONS, in common with its predecessors, is faced with a long-standing challenge. The UK has a large number of very different, cross-cutting and persistently changing geographies. Boundary change in the UK is endemic – there is more here than in the rest of the EU combined. This has long hampered the comparison of statistics over time. In an effort to solve this challenge, a hierarchical set of statistical geographies (Output Area (OA), with larger aggregate units LSOA and MSOA) was developed for the 2001 Census of England and Wales.

The OA is the smallest geographical area for which Census data are released. This geography was built from the Census data themselves. OAs were built from clusters of adjacent unit postcodes and designed to have similar population sizes. In the interests of disclosure control, each layer was regulated by population threshold sizes: the minimum population size for an OA was 100 usual residents or 40 households. Every OA was designed to be non-disclosive and this meant that any estimates derived from them or from LSOA and MSOA would also automatically be non-disclosive.

We use the statistical geographies as "building blocks" for estimating Census statistics for other – generally larger – geographies. We use OA as the building block for Census univariate data and LSOA for multivariate data due to the higher risk of disclosure.

The GSS Geography Policy defines the way that statistics for any 'higher' geography larger than OA, such as Local Authority Districts (LADs) should be generated. 'Exact estimates', derived directly from the Census households located within them, are calculated for OAs. These estimates are applied to the OA-PWC. Statistics for the individual instances of any higher geography are aggregated up from the OA-PWCs they contain.

---

<sup>2</sup> INSPIRE stems from an EU directive and aims to establish an infrastructure for spatial information in Europe to support Community environmental policies.

<sup>3</sup> <https://neighbourhood.statistics.gov.uk/dissemination/>

The procedure is known as ‘best-fit’, and this applies regardless of whether the higher geography comprises exact aggregations of OAs e.g. LSOA, MSOA or LADs, or crosscuts them as may be the case with parishes or postcode sectors.

The best fit procedure is reliable and simple. It eliminates the possibility of slivers, and consequently of disclosure by differencing. Statistics are guaranteed to be non-disclosive, and remain consistent even if the geography changes. Best-fitting works well for virtually every geography.

An exception had to be made of National Parks. These usually have very small populations, and OA-PWS associated with these populations are usually located beyond the Parks’ boundaries. It follows that National Parks are always under-counted if calculated by best-fit. We therefore always provide ‘exact-counts’ for National Parks, based on the Census Households resident within them.

Civil Parishes in England<sup>4</sup> are another geography not well served by best fitting.

Parishes originated as an ecclesiastical geography, but some of their functions were devolved into the civil realm at the time of Henry VIII’s dissolution of the monasteries. Over the succeeding centuries, two systems co-existed, but gradually diverged and were definitively divorced under the provisions of the Local Government Act, 1894.

Today the Civil Parish is a predominantly rural phenomenon, covering 91% of the population of England, but housing less than half of the country’s population. The majority of the population of England live in ‘unparished areas’.

Higher and lower population thresholds were embedded into the design algorithm for OAs Output Areas, and this meant that OAs had to be small where the population was dense and large where thinly spread. But there is no such correlation with parishes. Many parishes are very thinly inhabited – there are a few deserted medieval villages and military firing ranges, but there are also some large towns. Weston-super-Mare has the highest population (75,000). In consequence, some parishes are completely missed by the mesh of OA-PWCs. In fact, 1,140 (10%) English parishes do not contain an OA-PWC. It is therefore not possible to generate best-fit estimates for this group despite their combined Census population of nearly 120,000 residents. In some rural areas, clusters of such parishes without publishable data (e.g. Herefordshire, Yorkshire, Oxfordshire) present a challenge to the relevant Local Authority.

### **3. Investigating alternative options**

Because they are such a mix of old and new, populous and tenuous, small and large, it is difficult to develop a single methodology that can be applied successfully across the entire geography.

I was therefore set the challenge of finding an alternative approach which would permit publication of Census data for all parishes on a consistent basis and across the entire geography. Any solution had to be achievable and transferrable within ONS as well as being transparent to the user community.

A pilot project looked in depth at a variant of the best-fit procedure whereby the statistics for the OA-PWC closest to the orphan parish’s geometric centroid was applied to any ‘orphan’ parish (without an OA-PWC). Both the orphan parish and the OA-PWC had to be within the same LAD. But the external OA-PWC could represent an OA with a statistical character very different from the orphan parish. And this would result in wildly inaccurate estimates being generated for the orphan parish. Another

---

<sup>4</sup> Wales is entirely covered by an equivalent geography – Communities –which gets on well with best-fitting. The focus of this work is England.

serious problem was that the statistics for the external OAPWC would serve once for the parish within which it actually fell, and then again for any qualifying orphan parishes nearby.

Most of the 10,500 parishes in England contain one or more of the 170,000 OA-PWCs. But 1,140 parishes do not contain a single one. So, do we need more points? Perhaps so, but not necessarily of the same type. It turns out that this is a blind alley, best illustrated by the complete set of 26 million households. Even if we used this full resolution set, the populations of over 1,000 parishes would still fall below the thresholds (100 residents, 40 HH) decreed for OAs. We would therefore be faced with almost the same scale of problem.

So perhaps not more points: what we really need is a different way of allocating Census data to parishes. ONS has already provided one partial solution by providing a [lookup](#) of 2011 Census enumeration postcodes in England and Wales linked to 2011 parishes, but this is a palliative, supplying a strictly limited number of variables.

A further option which is being actively pursued is to base parish estimates on a grid. This is where the Eurostat 'Geostat' grid comes in. This is a pan-EU 1 x 1 km grid compiled according to INSPIRE principles. England is covered by ca. 130,000 grid-cells. While this may be fewer than the number of OAs across the country, their regular size and distribution compensate. Geostat has been revealed as a good test-bed for research.

There are three principal steps:

1. Associate the household Census data with each grid cell.
2. Associate each grid cell with each parish,
3. Calculate statistics for each parish based on these associations.

The association of Census data to grid cell proceeds through four possible routes:

- a. The grid cell's **geometric centroid**. 130k points;
- b. The grid cell's **population-weighted centroid**. 130k points;
- c. **Four Corners**: a simple smoothing method. One **quarter** of each cell's total data applied to each of the four corner points. Method produces 234k points;
- d. **Five Points**. Census data applied to **nearest** of geo centroid and four corner points. Method produces a finer grid at 45 degrees to the original, with 364k points.

The individual points produced by methods a-d are associated with the parish that they fall within, and statistics are then calculated for each parish, by summing the values for each associated point.

All of these methods produce pseudo-parishes whose boundaries are 'pixellated'. This contributes to disclosure control, by introducing a measure of give-and-take along parish boundaries. This amounts to a degree of smoothing, levelling peaks and filling in troughs in the dataset. In consequence, each will reduce the number of parishes for which ONS cannot publish data.

As an illustration, we will consider a pair of civil parishes (Figure 1: Staunton on Wye CP and Monnington on Wye CP) in the County of Herefordshire Unitary Authority. The two parishes (different colour polygons) are coterminous with a pair of Output Areas (red outlines), but both of the OA-PWCs are contained in one parish. Census households and inhabited Geostat grid squares are shown. Applying the best fit procedure to this pair would result in an exaggerated population for the one and no publishable data for the other.

The first grid method (Figure 2), calculating the parish statistics using the sum of HH data at the cells' geometric centroid, produces publishable population estimates of 418 for Staunton on Wye CP and 112 for Monnington on Wye CP. The grid square population-weighted centroid method on the other hand (Figure 3) produces an unpublishable 59 for Monnington on Wye CP.



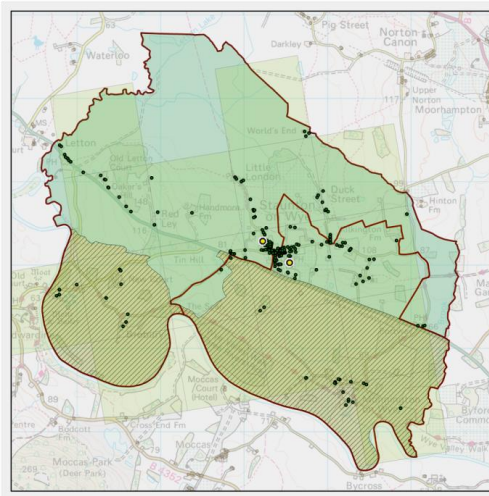


Figure 1: Two parishes, two OAs, HHs, gridcells

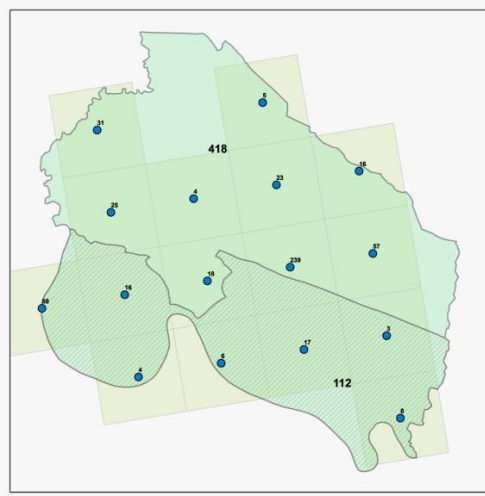


Figure 2: Cell geo-centroids

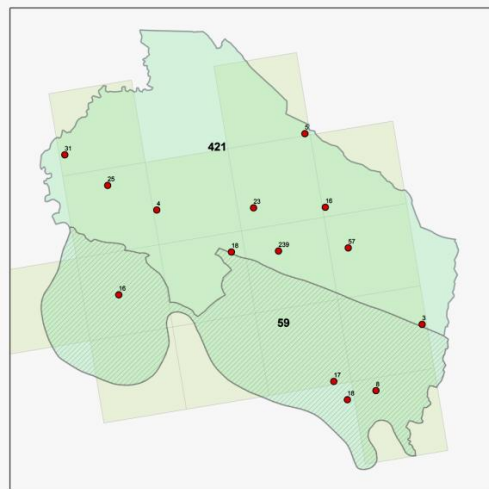


Figure 3: Cell pop-wighted centroids

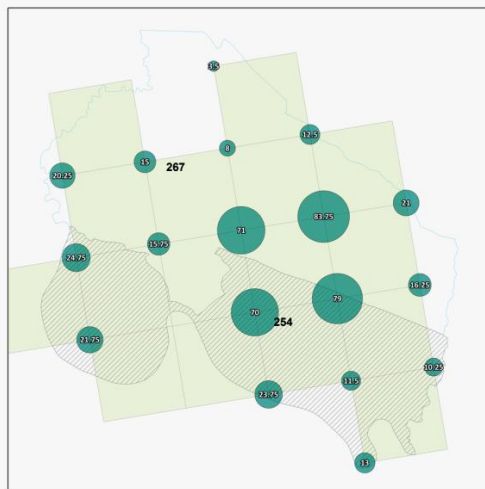


Figure 4 : Four-corner smoothing

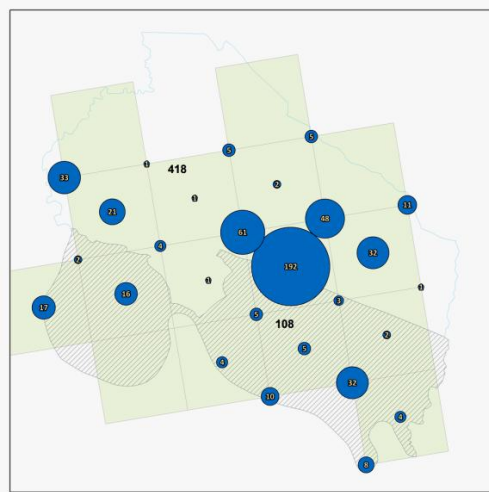


Figure 5 : Five-point nearest



Figure 6 : Joint parish council

With the four-corner method (Figure 4), there is a noticeable smoothing of the values across the space, resulting in similar values for both parishes. The results for the five-point method (Figure 5) are similar to those for the cell geometric centroid.

These results are typical for the entire country. In terms of reducing the number of parishes with disclosure levels of data, the most powerful method is four corner smoothing, with FivePoint and grid square geometric centroid methods some distance behind, and the grid square PWC method producing the least gain. It is no accident that the most successful method is also the most abstract. The least successful is also the one that is closest to the real settlement pattern. But whichever method is used, some low-population parishes will still have disclosure levels of data.

Results for all methods can be dramatically improved if, instead of publishing to individual Civil Parishes, we publish to the Parish Councils that run them.

The Civil Parish is a statutory geography – which legally has to be maintained by Ordnance Survey and ONS, but a civil parish is administered by a council, and it is this administrative unit, and not the Civil Parish, that people and Local Authorities relate most to.

The benefit of publishing to parish council level is that while the extent of the civil parish normally coincides precisely with the area of responsibility of the parish council, this is not always the case.

Under the Local Government Act, 1972, there is a provision for low-population parishes that are too small to run their own council, or to be otherwise viable, to come together in a working alliance under a common grouped parish council. Many counties (including Herefordshire, Figure 6) have taken advantage of this provision, grouping up to eight low-population civil parishes under a single common council. The majority of civil parishes without OA-PWCs are run by such joint councils, and this would allow us to publish data for the group.

There is unfortunately, no single definitive centrally-held list of parish councils across England. This information is held separately by around 200 Local Authorities. ONS and DCLG have recently begun collaborating on compiling such a list.

## **Summary**

ONS is continuing to assess the relative merits of the above methodologies, including the impact of grouped parish councils. We will publish a report on the findings and move on to a consultation with the user community.

## **Acknowledgements**

Andy Tait, Nick O'Rourke, Ian Coady and Dr Amy Fowler.

## **Biography**

Bruce Mitchell works in the Research Team of ONS' Geography Branch. He has an MSc in Geographical Information Science from Birkbeck College, University of London. His research interests extend beyond GIS to geography, history, languages and forestry.

## References

- Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2011) *Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales*. *Environment and Planning A* 43, (10), 2399-2418
- Martin, D. (2002) *Geography for the 2001 Census in England and Wales* *Population Trends* 108, 7-15
- ONS (2010) *GSS Geography Policy for National Statistics*, July 2010.
- ONS (2014): Ian Coady, Samantha Cockings, David Martin, Andrew Harfoot, Bruce Mitchell *Publishing statistical grids in the United Kingdom*. Publication pending.
- ONS (2015) *GSS Geography Policy for National Statistics*, forthcoming.
- Open Geospatial Laboratory Southampton (2015) *Opengrid: an open gridded population dataset for England and Wales. Product Manual*. Unpublished report. Open Geospatial Laboratory, University of Southampton.

# The Quality of Local Authority Spatial Data

Amy Mizen<sup>\*1</sup>, Sarah Rodgers<sup>1</sup> and Richard Fry<sup>1</sup>

<sup>1</sup>Farr Institute, College of Medicine, Swansea University

November 7, 2014

## 1. Introduction

The obesity epidemic is understood to be caused and influenced by multiple factors (Jones, 2014; Finegood, 2010). Amongst obesity research, the food environment is increasingly acknowledged as having a significant influence on obesity levels (Papass, 2007). Poor quality food environments, describe areas where there is an abundance of unhealthy food sources and poor availability of fresh and healthy foods. The availability of unhealthy food is believed to be promoting an unhealthy lifestyle and an over consumption of energy dense foods. This is hypothesised to contribute to rising rates of obesity.

This investigation will evaluate the quality of data, including addresses, provided by local authorities (LA). We will endeavour to collect a time series of data to be used in a natural experiment. We will model the food environment of child home-to-school routes at household level. This environmental data will then be linked to routine health data, which will allow the analysis of the relationships between child health and environmental exposures.

## 2. Background

The global obesity epidemic is a public health priority in countries around the world (Lobstein, 2004; Han, 2010). There is a serious concern for the issue of *childhood* obesity because of the health impacts later on in life. Many studies support the idea that obese children are at much higher risk of becoming obese adults than children and adolescents who have a healthy weight (Biro, 2010; Serdula, 1993). Of increased interest is the impact of the food environment on obesity. Poor food environments describe areas where there is poor quality of food available to residents (Dean, 2011). Poor food environments are usually densely populated with fast food outlets and lack in supermarkets or shops selling fresh fruits and vegetables. Such areas have also been referred to as obesogenic environments which do not just describe areas lacking healthy food outlets but also deficient in access to green spaces and walking opportunities (Egger, 1997; Williams, 2014).

The prevalence of obesogenic environments surrounding schools is of concern because of the threat to children's health (Austin, 2005). An increased density of fast food outlets near to schools encourages the increased consumption of fast food and also contributes to lifelong unhealthy attitudes towards food (Egger, 1997). Many studies that have previously investigated the impact of food environment and exposure have used "fixed" spatial units. However, to better understand causes of childhood obesity, there is a need to investigate the environment that children experience throughout their day, rather than simply where they live (Kestens, 2010). Harrison et al. (2014) highlighted the importance of further investigating the exposure environment that children experience on their journey to school.

In order to accurately model a spatial environment however, it is imperative that accurate data are imported into a GIS. Equally important is an awareness of the data limitations and implications this may have on proceeding analyses (Maynooth University, 2014; Devillers, 2007; Devillers, 2010). Initially, a data quality assessment will be carried out on data provided by a typical local council and

---

\* Amy@chi.swansea.ac.uk

the Food Standards Agency. These datasets will be cleaned and then compared in order to see whether local council datasets will be able to add valuable information to our model. If the local council data are found to be useful, this will add to a sophisticated model of the exposure environment that children experience on their way to school. There are currently no longitudinal investigations of the food exposure environment within the UK (CEDAR UK, 2014).

Spatial data are becoming increasingly important in policy research (Pirog, 2014). However, obesogenic environments are little researched and so local councils in the UK currently do not have the evidence they need to prevent additional fast food outlets from opening. In order to help planners and councillors, it is important to create high resolution spatial patterns of the exposure environment that children are subjected to on their way to school. Currently, council planners believe there is harm caused from an overabundance of food outlets in an area and deny planning applications. However, appeals against blocked planning applications are often successful due to a lack of evidence of the harmful effects.

We anticipate that this study will lead to an increase in data standards within local councils and standardised national datasets; through working with the Welsh Collaboration of Health and Environment, a group comprised of environmental health protectorate directors, local council health improvement officers, academics, and public health practitioners. To our knowledge this is the first study combining council-sourced GIS data and routinely collected health data to investigate obesogenic environments. To date, comparable studies have had to recruit participants and collect data prospectively.

### 3. Data & Methods

The main purpose of collecting the food outlet data is to develop a density model of food outlets for home-to-school routes. The densities will then be combined with health data held within the Secure Anonymised Information Linkage (SAIL) at Swansea University to evaluate the impact of the food environment on children's Body Mass Index (BMI).

Datasets of food outlets were obtained from Swansea County Council and Monmouthshire County Council. Food outlet data from the Food Standards Agency (FSA) were also obtained (FSA, 2014). The FSA dataset was taken to be the reference data, or "gold standard," as it is a national, routinely collected dataset. The council datasets were compared with the FSA data to see whether they are suitable datasets to use in building the density model.

The outlet data from Swansea and Monmouthshire Councils were received in different formats. In order to optimise comparability, the outlet data were formatted to csv files and where possible the same column names were allocated. Character columns were transformed in to lowercase, to minimise differences between datasets. The council data sets were matched with FSA data by the only common field which was outlet name.

Unmatched records were identified and sought to solve why they did not match any FSA outlets.

### 4. Results

Unitary Authority	Auto Matched (%)	Manual Matched (%)	Rural Matched (%)	Urban Matched (%)
Swansea	96	100	99	97
Monmouthshire	4	89	90	88

**Table 1.** Summary of matching progress

The rate of match is shown in Table 1. Auto matched describes the percentage of outlets that were the same as outlets contained in the FSA data. Manual matched refers to the percentage of outlets that match the FSA data after manual corrections have been made to the unmatched records. Swansea Council had a much greater rate of matching the FSA data compared to Monmouthshire. Of the 1361 food outlets in Swansea, 1307 were correctly automatically matched. Of the 47 unmatched outlets from Swansea, 42 records were found have spelling errors. The remaining five outlets had been closed down and the council dataset had not been updated. Monmouthshire had a far lower auto match rate. The manual match rate saw a significant increase in matching.

After cleaning the data, rural areas were more likely to be matched than urban areas.

For the wider investigation, food outlet data for all 22 Welsh LAs has been requested. Ten LAs are currently collating the datasets. Permission from the remaining ten LAs is still being sought. At the conference we aim to present the match rates of the data obtained to date and present a preliminary density model to give an idea of the potential of this study.

## **5. Discussion and Conclusions**

The LA data was susceptible to errors. Human error was the biggest contributor to error in the LA datasets i.e. spelling mistakes. Furthermore, the data providers from the LAs had to bring together datasets from various departments in order to collate these datasets. There is no uniform method of data collection between LAs, or even between departments within an LA. This complicates the comparison of data between LAs as it means that auto matching cannot be depended on to produce comparable results.

In the FSA data, the spatial locators are postcode centroids. Post code centroids have a spatial accuracy of approximately 100m. However, UPRNs which are being provided with the LA data gives the address level coordinates which will allow for higher resolution analyses when constructing the density model. UPRNs allow for the unique identification of a property which removes the possibility of error. As seen in our results, there can be many ways to format an address, but only one UPRN.

An advantage of using LA data is that this promotes engagement between LAs and academia; which not only promotes the importance of data that LAs record but also encourages better data collection and storage. Investigations such as this one encourages LAs to communicate with one another and share knowledge and data collection methodologies. Furthermore, LA data has proven to provide better typologies of food outlets stores compared to the FSA data. Along with opening and closing times of the outlets. The LA data was longitudinal whereas FSA is not which is very valuable for planning policy and public health research.

From the analysis of two local authority food outlet datasets, it can be concluded that despite potential flaws, LA data can provide researchers with valuable information that may not be captured in national datasets. Care should be taken when using such data sets but used correctly, LA data can add value to larger, uniformly collected national datasets. The value of local council data should be promoted so to encourage more stringent data capture and recording methods.

## **6. Acknowledgements**

Many thanks to Swansea County Council and Monmouthshire County Council for sharing their data.

## **7. References**

Biro FM, Wien M (2010) **Childhood obesity and adult morbidities** 91:1499–1505.

- Austin SB, Melly SJ, Sanchez BN, Patel A, Buka S, Gortmaker SL: **Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments.***Am J Public Health* 2005, 95:1575–81.
- Cetateanu A, Jones A (2014) **Understanding The Relationship Between Food Environments, Deprivation And Childhood Overweight And Obesity: Evidence From A Cross Sectional England-Wide Study.** *Health & Place* 27 : 67-76
- Dean WR, Sharkey JR (2011) **Rural and urban differences in the associations between characteristics of the community food environment and fruit and vegetable intake.***J Nutr Educ Behav* 43:426–33.
- Devilleers R, Stein A, Bédard Y, Chrisman N, Fisher P, Shi W (2010) **Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities.** *Trans GIS.* 14:387–400.
- Devilleers R, Bédard Y, Jeansoulin R, Moulin B (2007) **Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data.** *Int J Geogr Inf Sci* 21:261–282.
- Egger G, Swinburn B (1997) **An “ecological” approach to the obesity pandemic.***BMJ.* 315:477–80.
- Finegood DT, Merth TDN, Rutter H (2010) **Implications of the foresight obesity system map for solutions to childhood obesity.***Obesity (Silver Spring)* 18 (1) :13–6.
- Food Standards Agency (2014) **Search for food hygiene ratings** [<http://ratings.food.gov.uk/open-data/en-GB>] [Accessed October 12<sup>th</sup> 2014]
- Han JC, Lawlor D a, Kimm SYS (2010) **Childhood obesity.***Lancet.* 375:1737–48.
- Harrison F, Burgoine T, Corder K, van Sluijs EMF, Jones A (2014) **How well do modelled routes to school record the environments children are exposed to?: A cross-sectional comparison of GIS-modelled and GPS-measured routes to school.***Int J Health Geogr* 13:5.
- Kestens Y, Lebel A, Daniel M, Thériault M, Pampalon R (2010) **Using experienced activity spaces to measure foodscape exposure.***Health Place* 16:1094–103.
- Lobstein T, Baur L, Uauy R, Obesity I (2004) **Obesity in children and young people : a crisis in.** 5:4–85.
- Maynooth University (2014) Data Quality in GIS Available from <http://www.nuim.ie/staff/dpringle/gis/gis11.pdf> [Accessed October 22nd, 2014]
- Papas M a, Alberg AJ, Ewing R, Helzlsouer KJ, Gary TL, Klassen AC (2007) **The built environment and obesity.***Epidemiol Rev* 29:129–43.
- Pirog MA (2014) **Data will derive innovation in public policy and management research in the next decade.** *J Policy Anal Manag* 33:537–543.
- Serdula M, Ivery D, Coates R, Freedman D, Williamson D, Byers T (1993) **Do obese children become obese adults.** *Preventive* 22:167–177.

Williams J, Scarborough P, Matthews a, Cowburn G, Foster C, Roberts N, Rayner M (2014) **A systematic review of the influence of the retail food environment around schools on obesity-related outcomes.***Obes Rev* 15:359–74.

## **Biographies**

Amy Mizen is a DECIPHER PhD student at the Farr Institute @ CIPHER, Swansea University. Beginning in October 2014, her PhD project is investigating the impact of modelled school travel routes on child health using GIS and routine linked data.

Richard Fry is a Senior Research fellow in GIS at the Farr Institute @ CIPHER, Swansea University. His research interests include accessibility modelling, health geographies, data integration and linkage, OpenSource and WebGIS.

Sarah Rodgers is an Associate Professor in Spatial Epidemiology and an investigator in the new MRC e-health centre of excellence, CIPHER, at Swansea University. Her research is aided by anonymised individually-linked health, and demographic data, and aims to influence policy to improve environments and positively impact physical and mental health.



# Is VGI Big Data?

Peter Mooney and Adam C. Winstanley  
Department of Computer Science,  
Maynooth University,  
Co. Kildare, Ireland.

## Summary (100 words)

Volunteered Geographic Information (VGI) has become a popular source of geographic data for GIS practitioners in recent years. VGI datasets are characterised as being: large in volume, subject to dynamic changes and updates, collected through crowdsourcing architectures using a variety of devices and technologies and contain a mixture of structured and unstructured information. Can we call VGI a form of Big Data? Are VGI datasets developing characteristics that make processing them using traditional data processing applications and techniques difficult and unsatisfactory? We explore this question with reference to a number of sources of VGI.

**KEYWORDS: (5)** VGI, Big Data, GI Data Processing, Crowdsourcing

## INTRODUCTION

Volunteered Geographic Information (VGI) continues to gain research attention (Mooney and Corcoran, 2014). VGI is a special case of user-generated content (UGC), usually having an explicit or implicit embedded spatial component. The large quantities and diverse information generated as VGI presents a number of challenges for developing methodologies to use it in research, applications and for understanding its societal implications (Elwood et al, 2012). In this abstract we explore the question “Is VGI Big Data?” by presenting an overview of three popular sources of VGI. Does carrying out research with VGI require GIScientists and practitioners to equip themselves with a new set of tools, skills and methodologies capable of extracting knowledge from these very large dynamic datasets as defined as Big Data? Clear definitions of what exactly Big Data is are very difficult to find (Goodchild, 2013). Kitchin and Laurialt (2014) state that new forms of Big Data are produced predominantly through new information and communication technologies (ICTs). Prior to 2008 data were rarely considered in terms of being ‘small’ or ‘big’ (Kitchin and Laurialt, 2014). All data were, in effect, what is now sometimes referred to as ‘small data’ regardless of their volume. Big Data is associated with data from sensors and software that digitize and store a broad spectrum of social, economic, political, and environmental patterns and processes. Miller and Goodchild (2014) write that geographical Big Data is produced by in-situ sensors carried by individuals in phones, attached to vehicles, embedded in sensing infrastructure and georeferenced social media. Boyd and Crawford (2012:663) argue that “there is little doubt that the quantities of data now available are often quite large, but that is not the defining characteristic of this new data ecosystem”.

Using VGI for GIS research and application development has been growing in popularity over the past number of years. Cinnamon and Schuurman (2013) state three principal methods by which VGI is collected and generated: (1) by using geo-aware mobile devices, (2) annotating geographic features using geoweb mapping interfaces, and (3) by extracting or inferring location information from ambient geospatial data in social media (photos, videos, blog posts, tweets, etc) (Stefanidis et al, 2013). Ambient geospatial data, as opposed to VGI generated by (1) or (2) are often messy, consisting of data that are unstructured, collected with no quality control and frequently accompanied by no documentation or metadata (Miller and Goodchild, 2014).

### CHARACTERISING VGI AS BIG DATA

We use OpenStreetMap, geolocated Twitter ‘tweet’ datasets and Foursquare Venue data as our three examples of VGI. These three sources are available openly and for free and have been used widely by researchers over the past number of years. We shall apply characterisations from Kitchin (2014) and Boyd and Crawford (2012) to assess these sources of VGI.

Kitchin (2014) characterises Big Data as being:

- **Voluminous:** consisting of terabytes or petabytes of data
- **High Velocity:** being created in or near real-time
- **Varied:** Being structured and unstructured in nature
- **Exhaustive:** In scope - striving to capture entire populations or systems
- **High Resolution:** Fine-grained and aiming to be as detailed as possible
- **Relational:** Containing common fields enabling the joining of different datasets
- **Flexible:** Having traits of being easily extended (adding new fields) and scalability (expand in size rapidly)

Boyd and Crawford (2012) characterises Big Data as having

- **Technology Requirements:** maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
- **Analysis Possibilities:** drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
- **Mythology:** the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

K denote characterisation from Kitchin (2014) and BC denotes from Boyd and Crawford (2012).

Dataset Characteristic	OpenStreetMap	Foursquare	Twitter
<b>K:Voluminous</b>	Whole world 36GB compressed XML (~500 GB Uncompressed). City and regional areas much smaller (several GB). A binary compressed format is also available. Difficult	50 million users have generated 6 billion “checkins”. Approximately 60 million Foursquare Venues worldwide. Almost 2 million business listed	500 million tweets per day and around 200 billion tweets per year. Estimates vary from 1% to 5% on the quantity of tweets with explicit geographical information (geolocated). Estimates

	to estimate daily volume. Several hundreds of thousands polygons added per day		indicate about 12GB per day in tweet text (not including other message overheads)
<b>K:High Velocity</b>	Changes and edits are reflected quickly in OSM. Many services provide hourly updated downloads. Several APIs are available	Very high velocity. However to access 'checkin' data there are rate limits imposed which limit the number of API requests which can be processed in an hour by an application.	Very high velocity. Approximately 800 per second. Special access requirements for this stream of Tweets. There are rate limits imposed for free usage of API and Streaming
<b>K:Varied</b>	Variation in how tagging rules are implemented. Tags can contain structured and unstructured data	Data is returned from API calls in JSON format. This provides a robust data structure.	140 characters per tweet which includes multilingual free text, URLs, hashtags, twitter handles etc.
<b>K:Exhaustive</b>	Yes - data model flexibility allows any geographical feature to be included in OSM	Foursquare database of venues grows as businesses, venue owners and users add to the database.	Provides a communication medium for people. In this sense Twitter looks to connect a very large percentage of the world's population.
<b>K:High Resolution</b>	Resolution can vary over features, device types used to capture the VGI, etc	Fine grained locational information is attached to venues (geographical coordinates and addresses)	Not directly relevant to datasets of Tweets. Very high temporal resolution. Due to small percentage of geolocated Tweets spatial resolution is difficult to quantify
<b>K:Relational</b>	Not directly - but mapping to other datasets have been performed	Yes - properties of the JSON responses can be linked to other datasets	Dependent upon the contents of the Tweets themselves.
<b>K:Flexible</b>	OSM's flexibility, and also a cause of some QA/QC issues, is a flexible tagging/attribution model in combination with a reasonably simple data model.	Unclear. The data model appears to be fixed at present.	Twitter is open text limited to 140 characters.

<b>BC: Technology Requirements:</b>	Processing the entire OSM DB requires computing power and resources beyond that available on a desktop. Regions and subsets can be processed on standard desktop machines	Accessing Foursquare's API or Streaming processes requires programming and software knowledge. Storage of Foursquare data is not overly cumbersome. Rate limits means download of data may need to be spaced over a long time period.	Accessing Twitter's API or Streaming processes requires programming and software knowledge. Storage of Twitter data is not overly cumbersome. Analysis will require advanced string-based data mining algorithms.
<b>BC: Analysis Possibilities</b>	Analysis possibilities are beginning to emerge. Recent interest amongst research in the social construction of OSM on a regional and global basis.	Very wide range of possibilities as the Foursquare data can combine user movement patterns between venues over time. Venue data contains metadata about the venue itself. This offers great analysis possibilities	Large number of analysis studies have been produced. VGI type analysis is restricted by the low rate of geolocation in Tweets. The ability to link Tweets to user-profile and location offers significant analysis possibilities for researchers
<b>BC: Mythology:</b>	There is a belief amongst many OSM users that this VGI dataset could yield some very interesting social patterns and knowledge about the digital divide, socio-demographics online and spatial cognition	Some researchers suggest that Foursquare users use the service to document their social movement history and find venue information. This could provide a vast history of human movement and social patterns.	The entire dataset of Tweets has been called the largest dataset on human interaction ever created. A lucrative industry has emerged as being Twitter Content Partners involving the reselling, curation and analysis and business insight extraction of Twitter data for commercial partners.

## CONCLUSIONS

In this abstract we have addressed the question 'Is VGI Big Data?'. Our analysis only considers VGI which can be accessed freely and openly. The three examples of VGI presented exhibit many of the characteristics of Kitchin and Boyd and Crawford's Big Data. These characteristics will exert different influences depending on the types of analysis or applications these VGI data are being used for. For example collecting one month of Twitter or Foursquare data for London will not present storage problems but significant computational resources may be required if highly complex spatial data mining algorithms are applied. Similarly this applies to OSM data which can be accommodated in any standard spatial

database. However the type of analysis performed will greatly influence the resource requirements. For example Gao et al (2014) build a high-performance cloud-computing Hadoop-based geoprocessing MapReduce platform to facilitate gazetteer development using OSM and other georeferenced social media. This platform is implemented not because of characteristics of the datasets used but rather to speed-up the computation being performed.

Researchers accessing these sources of VGI are not necessarily working with 'Big Data'. The original creation of these data (in particular Foursquare and Twitter) and the potentially highly complex tools and skills required to analyse them exhibit Big Data characteristics. Many papers published on VGI, to date, have used VGI datasets for a specific set of locations over a specific time period. Few, if any, researchers are analysing VGI as it is produced in real-time. Instead the VGI is collected then analysed in the same way as the 'small-scale' studies of the past (Kitchin and Laurialt, 2014). We are presently witnessing a fast changing landscape with respect to geographical data. The types of data flows we are seeing from VGI and UGC are part of this changing landscape. Ubiquitous, ongoing, data flows are important because they allow us to capture spatio-temporal dynamics directly and at multiple scales (Mooney and Corcoran, 2014; Miller and Goodchild, 2014). Graham and Shelton (2013:255) write that while there has been significant discourse surrounding Big Data "there has yet to be a significant, sustained effort to understand its geographic relevance".

## BIOGRAPHY

Peter Mooney is a Research Fellow with the Irish Environmental Protection Agency and the Department of Computer Science at Maynooth University. Adam Winstanley is a senior lecturer and Head of the Department of Computer Science at Maynooth University.

## REFERENCES

- Boyd, D. and Crawford, K. (2012) Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society*. 15(5), 662-679. DOI: 10.1080/1369118X.2012.678878
- Graham, Mark, and Taylor Shelton. 2013. "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography* 3 (3): 255–61.
- Cinnamon, Jonathan, and Nadine Schuurman. 2013. "Confronting the Data-Divide in a Time of Spatial Turns and Volunteered Geographic Information." *GeoJournal* 78 (4): 657–74.
- Elwood, Sarah, Michael F. Goodchild, and Daniel Z. Sui. 2012. "Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice." *Annals of the Association of American Geographers* 102 (3): 571–90.
- Gao, Song, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. "Constructing Gazetteers from Volunteered Big Geo-Data Based on Hadoop." *Computers, Environment and Urban Systems*, forthcoming.

- Goodchild, M. F. 2013. "Quality of Big (Geo)Data." *Dialogues Human Geography* 3(3): 280–84.
- Graham, Mark, and Taylor Shelton. 2013. "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography* 3 (3): 255–61.
- Kitchin, Rob. 2013. "Big Data and Human Geography Opportunities, Challenges and Risks." *Dialogues in Human Geography* 3 (3): 262–67.
- Kitchin, Rob. (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society*. 1(1), April 1, 2014. <http://bds.sagepub.com/content/1/1/2053951714528481.abstract>.
- Kitchin, Rob and Lauriault, TraceyP. (2014) Small data in the era of big data. *GeoJournal*. 2014/10/11, 1-13. <http://dx.doi.org/10.1007/s10708-014-9601-7>.
- Miller, HarveyJ and Goodchild, MichaelF. (2014) Data-driven geography. *GeoJournal*. 2014/10/10, 1-13. <http://dx.doi.org/10.1007/s10708-014-9602-6>.
- Mooney, Peter, and Pdraig Corcoran. 2014. "Analysis of Interaction and Co-Editing Patterns amongst OpenStreetMap Contributors." *Transactions in GIS* 18 (5): 633–59.
- Stefanidis, Anthony, Crooks, Andrew and Radzikowski, Jacek. (2013) Harvesting ambient geospatial information from social media feeds. *GeoJournal*. 78(2), 2013/04/01, 319-338. <http://dx.doi.org/10.1007/s10708-011-9438-2..>

# Exploring the role of consumer data for food in national survey reporting

Morris MA<sup>\*1</sup>, Clarke GP<sup>†1</sup> and Birkin MH<sup>‡1</sup>

<sup>1</sup>School of Geography, University of Leeds

January 9, 2014

## Summary

National data collected to aid understanding of spending patterns associated with food consumption and nutrition in the UK are reported in the Family Food module of the Living Costs and Food Survey. This survey data is used to better understand our society, including the geographies of food spending and informs the Consumer Price Indices. Such surveys require a nationally representative sample of volunteers to report their food spending and consumption patterns. This paper explores the role in which big data relating to food purchases could supplement such surveys and how reports using consumer data compare to survey data geographically.

## KEYWORDS:

Consumer data  
National survey  
Living Costs and Food  
Food Consumption  
Nutrition

## 1. Extended Abstract

### 1.1 Background

Collecting data from households using survey methods is a costly and time consuming process. The Office for National Statistics (ONS) carry out this process for the Living Costs and Food Survey (LCFS) on an annual basis, recruiting around 6000 households in the UK (around 150 of these being in Northern Ireland) (Office for National Statistics 2012). The European Standard Classification of Individual Consumption by Purpose (COICOP) (United Nation Statistics Division 2015) is used within the LCFS to categorise spending. In conjunction with DEFRA the Family Food module of the survey is produced, focussing on household spending on food and drink (Department for Environment Food and Rural Affairs 2012). The nutrient content of family diets is derived from the food spending records.

Food price inflation has been higher than general inflation since 2007 meaning that spending on food is a higher proportion of household budgets than previously (Office for National Statistics 2012). During the recent recessionary times, cost savings have been made in food purchases (United States Department of Agriculture 2011, Crossley, Low et al. 2012). Diet is the leading modifiable risk factor in non-communicable chronic diseases such as Obesity, Type 2 diabetes, Cardiovascular disease and some cancers (Institute for Health Metrics and Evaluation 2013, US Burden of Disease Collaborators

---

\* m.morris@leeds.ac.uk

† g.p.clarke@leeds.ac.uk

‡ m.h.birkin@leeds.ac.uk

2013). It is widely published that a healthy diet is more expensive than a less healthy one (Rehm, Monsivais et al. 2011, Morris, Hulme et al. 2014) so it is wholly possible that cost savings in diet, at the expense of diet quality, could impact health in the longer term. Consideration of the opportunity cost of making such choices is unlikely to be considered by the consumer, but reports such as the Family Food Survey allow the government and researchers to monitor consumer behaviour in relation to food.

Consumer transaction data from UK retailers for food and drink purchases are collected from millions of households, often with the ability to link to loyalty card information containing demographic characteristics (Felgate, Fearne et al. 2012). The Consumer Data Research Centre (CDRC), an ESRC data investment, has been established to broker such data between retailers and researchers in order to tackle global challenges over the coming years (Consumer Data Research Centre 2015). Better understanding of consumer behaviour in respect to food purchases could contribute to improved public and preventative health.

Geographies of consumption are typically reported at a large geographical unit such as Government Office Region in the UK. This ensures anonymity but aggregates to such a scale that wide generalisation occurs and pockets of certain dietary behaviours are lost. Some surveys present results for geodemographic group which allows specific groups to be more easily pinpointed spatially, but the generalisation occurs according to demographics through segmentation of the whole country into a defined number of groups. Using Big Data which includes supermarket loyalty card records results at a small spatial scale can be generated, but care regarding the level which these are reported needs to be taken to ensure anonymity is maintained.

Using consumer data as a measure of diet is not a new concept. Much research has been carried out using a 'basket analysis' approach whereby a typical basket of food is defined, often with the components required to meet the national dietary guidelines (Mooney 1990, Sooman, Macintyre et al. 1993, Larsen and Gilliland 2009, Cummins, Smith et al. 2010, Drewnowski, Aggarwal et al. 2012). Important concerns relating to the cost of the basket, and hence the cost of meeting dietary guidelines can be explored. The concept of Big Data provides an opportunity to build on existing research on an amplified scale.

## **1.2 Methods**

While consumer data is 'Big Data' and presents many exciting research opportunities, there are also challenges to overcome. Building relationships with retailers is essential. It is likely that data will be available only from certain retailers, especially in the beginning. This data will likely over-represent certain socioeconomics groups within a population. The data will then require weighting such that it becomes representative. While the numbers will be much higher than those in national survey samples, with wide geographic reach, they may not be evenly spatially distributed and again this will need to be accounted for in any reporting and analysis.

This paper will discuss the methods employed for manipulation of consumer data into shape for research purposes as a first step for researchers; a process which will be contributed to from previous lessons learned and continue to be refined over time. Nutrient costs will be presented in the format of the Family Food Survey. Utilising the same classification will allow for comparison of the reports generated from different data sources (survey verses consumer data). It will be possible to perform post hoc power calculations for the ability to detect nutrient differences which are clinically important for health for each data source.

Research into the effect of nutrient consumption and health outcomes typically uses dietary data from self-reported food records, be it a 4 day weighed food diary, a series of 24 hour dietary recalls or a Food Frequency Questionnaire which aims to record typical diet over defined period of time. Dietary self-reporting is subject to a range of bias, such as the under-reporting by those who are overweight or



obese. Using consumer data provides real data on food which has been purchased, but without dietary consumption records. Assumptions need to be made regarding the distribution of food consumption within a household and issues such as how much of food is wasted, or remains in the store cupboard until another week or month. Methods to account for these will be discussed.

Geographies of consumption will be explored and results compared to those of the Family Food Survey. If possible inter retailer differences will be explored spatially, according to Government Office Region and by Geodemographic category. The retailer generated report will be cross referenced to a synthetic population generated by spatial microsimulation techniques using small area microdata from the 2011 census. This nationally representations synthetic population will enable us to better understand the bias resulting from regional variation in retailer distribution, shown in figure 1 and also sociodemographic variations in customers to these stores, shown by education level in figure 2. Spatial interaction modelling methods will be used to explain such variations.

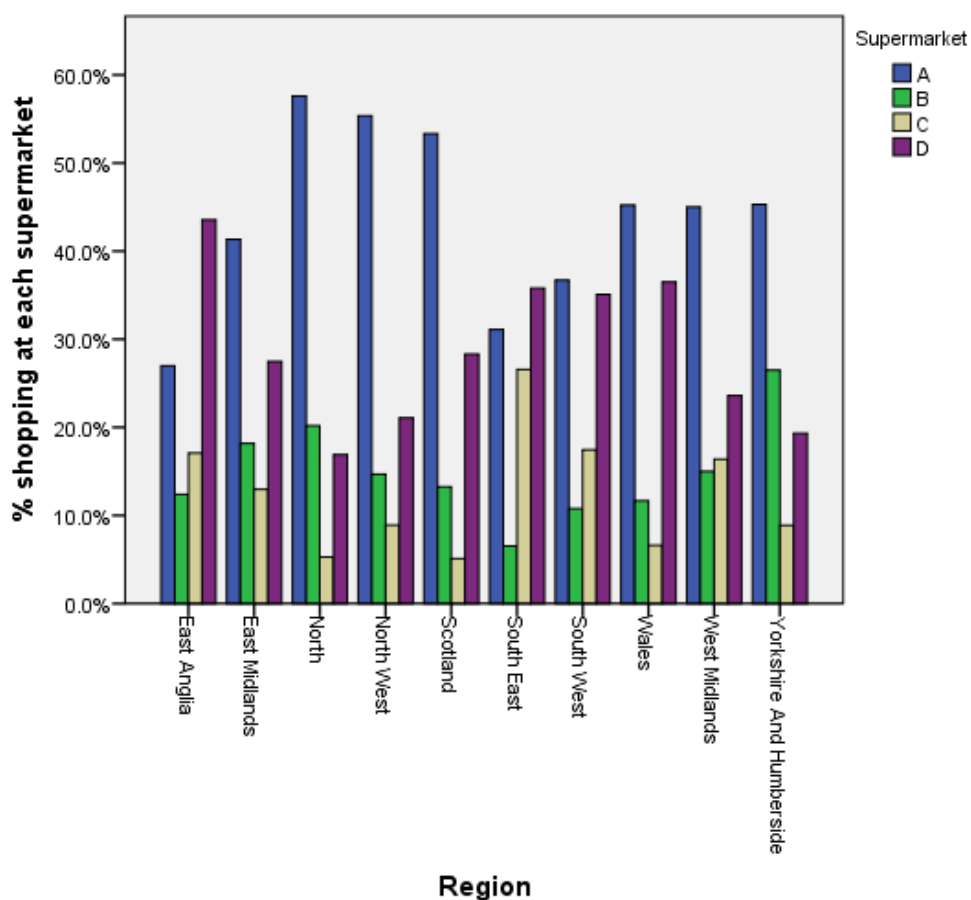


Figure 1 – Use of different supermarkets by region, derived from Acxiom survey data 2007.

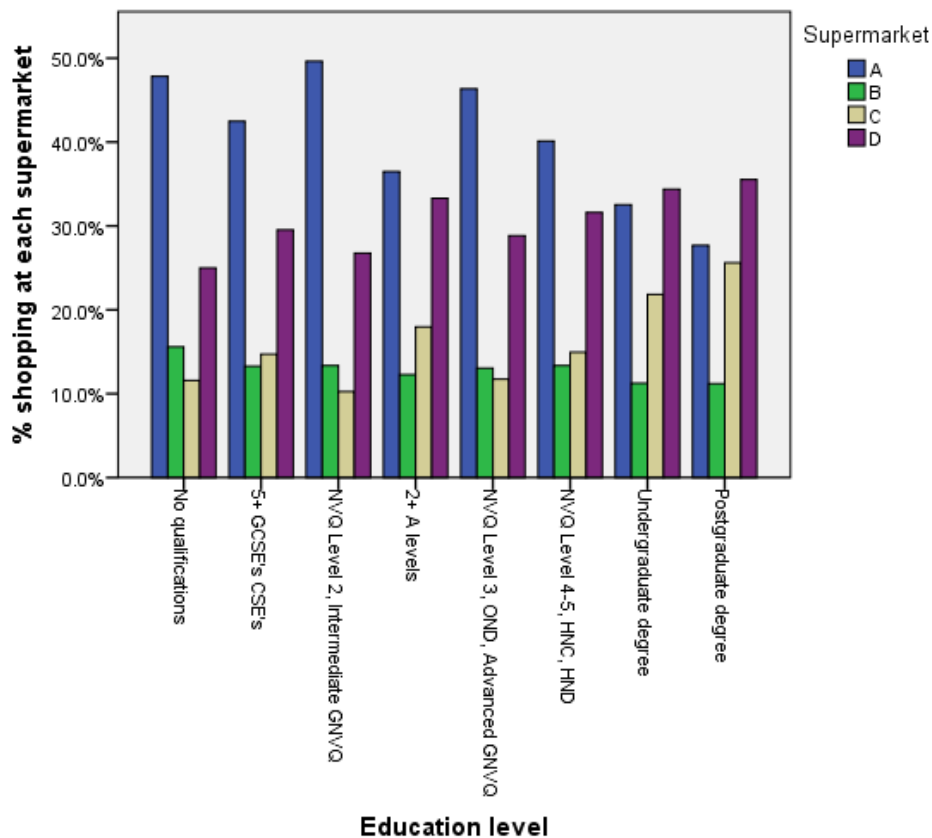


Figure 2 – Percentage of shoppers at different supermarkets by educational level, derived from Acxiom survey data 2007.

### 1.3 Implications

Results could open up new channels of thought regarding how to report and analyse behaviour such as food purchasing and consumption. It is possible that results from Big Data could provide greater reliability in survey estimates from increased sample size power and a decrease in reporting bias, in a more timely manner, as retail data is collected daily, without a survey data collection time lag. It could provide policy makers with greater opportunity and confidence in implementing new initiatives to influence diet and subsequent health. Better understanding of the geographies of spending and consumption of food are important to policy makers, health professionals and retailers.

### 1.4 Future work

Real time experiments could be carried out in conjunction with retailers to pilot important interventions such as: the impact of taxing certain foods; changing product prices; altering availability of products; simulated effects of change in quality; impact of guidance from national campaigns; or the effect of manipulating local social norms. This provides an exciting alternative to virtual supermarket experiments and advanced modelling techniques.

## Acknowledgements

This work has been funded by the Consumer Data Research Centre – an ESRC data investment.

## Biographies

### Michelle Morris

Michelle is a postdoctoral research fellow in the Consumer Data Research Centre at the University of Leeds. Her primary research interests are in spatial variations in diet and health. Michelle is an interdisciplinary researcher with a background spanning, spatial analysis and policy, nutritional epidemiology and health economics.

### Graham Clarke

Graham is professor of business geography at the Centre for Spatial Analysis and Policy at the University of Leeds. Graham's research interests include GIS, urban services, retail and business geography, geographies of health, urban modelling and continued professional education.

### Mark Birkin

Mark is Director of the Consumer Data Research Centre, based at the University of Leeds where he is Professor of Spatial Analysis. His major research interests are in simulating social and demographic change within cities and regions and in understanding the impact of these changes on the need for services.

## References

Consumer Data Research Centre. (2015). "Consumer Data Research Centre." from <http://cdrc.ac.uk/>.

Crossley, T., H. Low and C. O'Dea (2012). Household Consumption Through Recent Recessions. Institute for fiscal studies - working paper W11/18.

Cummins, S., D. M. Smith, Z. Aitken, J. Dawson, D. Marshall, L. Sparks and A. S. Anderson (2010). "Neighbourhood deprivation and the price and availability of fruit and vegetables in Scotland." *Journal of Human Nutrition & Dietetics* **23**(5): 494-501.

Department for Environment Food and Rural Affairs (2012). Family Food 2012. D. f. E. F. a. R. Affairs.

Drewnowski, A., A. Aggarwal, P. M. Hurvitz, P. Monsivais and A. V. Moudon (2012). "Obesity and supermarket access: proximity or price?" *Am J Public Health* **102**(8): e74-80.

Felgate, M., A. Fearne, S. DiFalco and M. G. Martinez (2012). "Using supermarket loyalty card data to analyse the impact of promotions." *International Journal of Market Research* **54**(2): 221-240.

Institute for Health Metrics and Evaluation (2013). The Global Burden of Disease: Generating Evidence, Guiding Policy – European Union and Free Trade Association Regional Edition.

Larsen, K. and J. Gilliland (2009). "A farmers' market in a food desert: Evaluating impacts on the price and availability of healthy food." *Health & Place* **15**(4): 1158-1162.

Mooney, C. (1990). "Cost and availability of healthy food choices in a London health district." *Journal of Human Nutrition and Dietetics* **3**: 111-120.

Morris, M. A., C. Hulme, G. P. Clarke, K. L. Edwards and J. E. Cade (2014). "What is the cost of a healthy diet? Using diet data from the UK Women's Cohort Study." *J Epidemiol Community Health*

68(11): 1043-1049.

Office for National Statistics. (2012). "Living Costs and Food Survey, 2012." Living Costs and Food Survey (Expenditure and Food Survey) Retrieved January, 2015, from <http://discover.ukdataservice.ac.uk/catalogue/?sn=7472&type=Data%20catalogue>.

Rehm, C. D., P. Monsivais and A. Drewnowski (2011). "The quality and monetary value of diets consumed by adults in the United States." American Journal of Clinical Nutrition **94**(5): 1333-1339.

Sooman, A., S. Macintyre and A. Anderson (1993). "Scotland's health--a more difficult challenge for some? The price and availability of healthy foods in socially contrasting localities in the west of Scotland." Health Bull (Edinb) **51**(5): 276-284.

United States Department of Agriculture. (2011). "Food Spending Adjustments During Recessionary Times." from <http://www.ers.usda.gov/amber-waves/2011-september/food-spending.aspx>.

United Nation Statistics Division. (2015). "COICOP (Classification of Individual Consumption According to Purpose)." Retrieved January, 2015, from <http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5>.

US Burden of Disease Collaborators (2013). "The state of US health, 1990-2010: burden of diseases, injuries, and risk factors." The Journal of the American Medical Association **310**(6): 591-608.

# Retail Modelling in Tourist Resorts: A case study of Looe, Cornwall

Newing, Andy<sup>\*1</sup>, Clarke, Graham<sup>†1</sup> and Clarke, Martin<sup>‡1</sup>

<sup>1</sup>School of Geography, University of Leeds

January 6th, 2015

## Summary

We demonstrate that applied retail modelling can be used to support retail planning and location based decision making within highly seasonal tourist resorts. Using small area spatiotemporal demand estimates and a custom built Spatial Interaction Model (SIM), we evaluate a ‘live’ retail development scheme. Our modelling approach can be used to estimate store revenue and to identify the impact of supply side changes on consumer flows, store and retailer market shares and network performance in order to support the retail planning process.

**KEYWORDS:** Seasonal consumer demand, Grocery retail, Retail planning, Spatial Interaction Modelling, Tourism

## 1. Introduction

Seasonal visitor-induced demand fluctuations within tourist resorts present considerable challenges for the retail planning process. ‘Traditional’ approaches to predict store revenues, market shares and local economic impacts of proposed store developments often fail to account for the spatial and temporal characteristics of visitor demand. We report on applied research carried out in conjunction with a major UK retailer and consider sales of groceries (food and drink) in highly seasonal coastal tourist resorts. Drawing on a live retail development scheme from the Cornish resort of Looe (UK), we combine small area seasonal visitor demand estimates with a custom built Spatial Interaction Model (SIM) in order to evaluate the impact of new store development on consumer choice, consumer trip-making behaviours and on overall store and network performance. We demonstrate that our modelling approach can be used to evaluate new grocery store development in tourist resorts, identifying implications for the retail planning process.

## 2. Estimating seasonal demand fluctuations and store revenue in tourist resorts

Modelling local demand fluctuations driven by tourism is an under-researched area and little is known about the volume or seasonal distribution of visitors or their expenditure at the level of individual store catchments. We developed small area spatiotemporal estimates of visitor numbers and expenditures for the county of Cornwall (South West England), reported fully in Newing *et al.* (2013), utilising local and national surveys and insight from store loyalty cards. Our SIM utilises these demand estimates to model interactions (expenditure flows) between demand origins (census Output Areas) and grocery stores. Interactions are driven by store characteristics (size and brand) and accessibility (road travel time), with the model disaggregated by retail band and consumer characteristics in order to generate realistic flows. The development and calibration of the model is illustrated in detail in Newing *et al.* (2014), drawing upon empirical data from a major retailer.

We use the SIM to model visitor grocery expenditure flows within tourist resorts, accounting for visitor expenditure alongside existing residential spend to generate accurate seasonal store revenue

---

<sup>\*</sup> [a.newing@leeds.ac.uk](mailto:a.newing@leeds.ac.uk)

<sup>†</sup> [g.p.clarke@leeds.ac.uk](mailto:g.p.clarke@leeds.ac.uk)

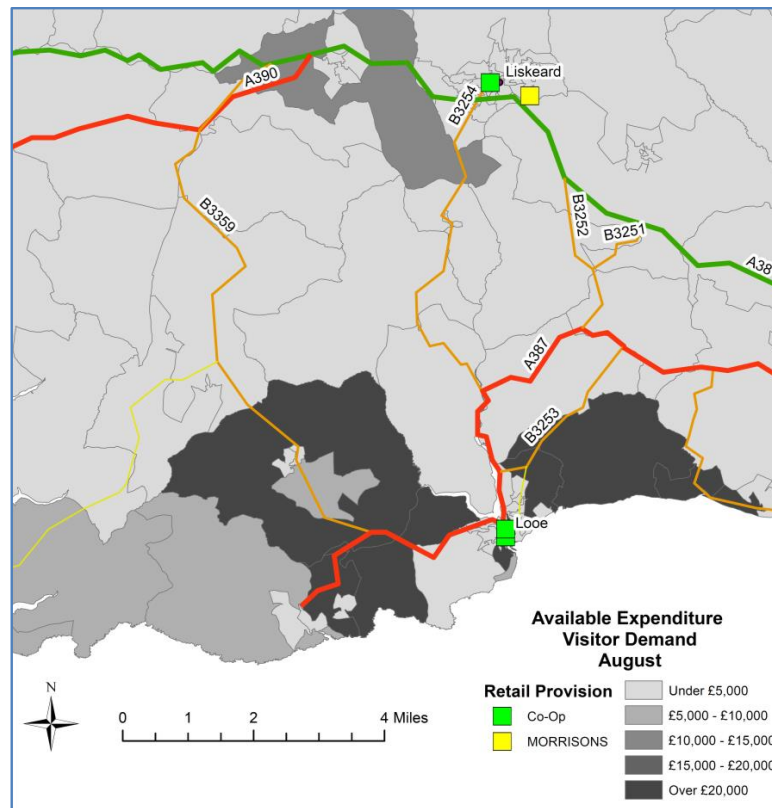
<sup>‡</sup> [martin.c.clarke@btinternet.com](mailto:martin.c.clarke@btinternet.com)

estimates. We have previously reported that this model, calibrated with reference to known consumer flows and store sales data, is able to estimate store revenues to an impressive level of accuracy (Newing *et al.* 2014). In the following sections we assess existing retail supply and demand within the tourist resort of Looe, before evaluating a new store proposal to serve the resort. All demand estimation, market shares, store revenues and other values reported refer to the year 2010 and are derived from our modelling.

### 3. Modelling consumer flows within the resort of Looe

Looe is a popular waterfront tourist destination located on the south coast of Cornwall. Our demand estimation suggests that 4,104 residents live within a 15 minute off-peak drive time of Looe, with an average total weekly spend on food and drink estimated at around £266,000. Expenditure derived from overnight visitors staying within the catchment generates additional demand of £312,000 per week during the August peak tourist season. Peak season visitor demand (Figure 1) is spatially clustered towards the coastline adjacent to Looe.

Grocery provision in the resort is limited to two small stores under the ‘Co-Op’ brand, suitable primarily for top-up shopping. Modelling suggests that residents and visitors within this catchment are dependent on these stores, which have a combined market share of around 27% of all food and drink expenditure, well in excess of modelled averages for this brand and store format. The resort lacks provision for residents or visitors to carry out a main food shop and many travel beyond the town to access a large format store in Liskeard (approximately 20 minutes’ drive from Looe) and another in Bodmin (around 30 minutes’ drive) (Figure 1), which exhibit a combined market share of 44% of all food and drink expenditure originating within this catchment. Visitors staying within over 1,500 accommodation units to the west of Looe face journey times in excess of 30 minutes to reach larger food stores due to the rural nature of this catchment and its poor road network.

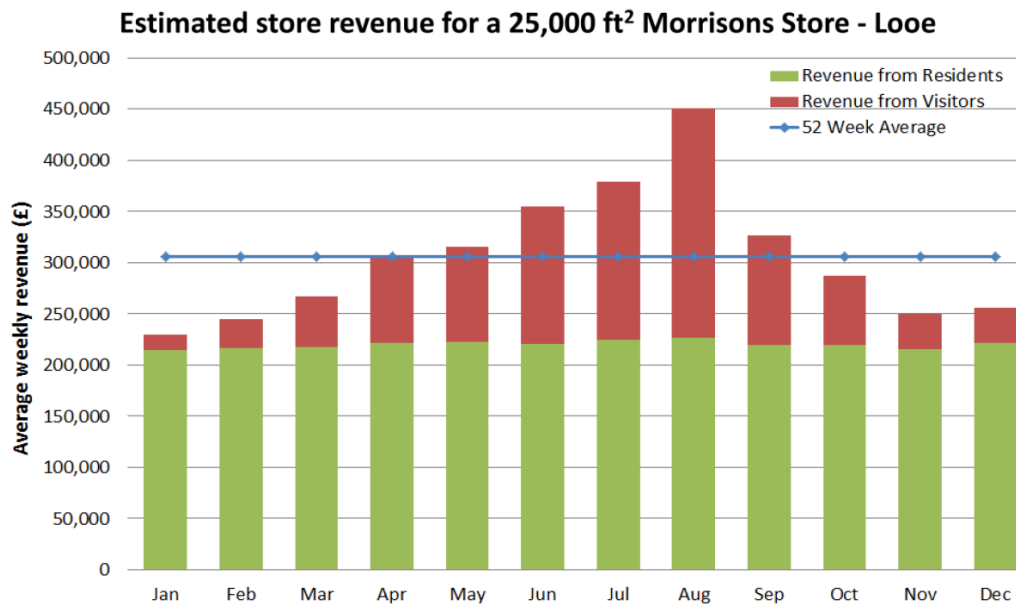


**Figure 1** Visitor grocery expenditure estimates (£ per week) by Census Output Area

#### 4. Modelling new store development in Looe

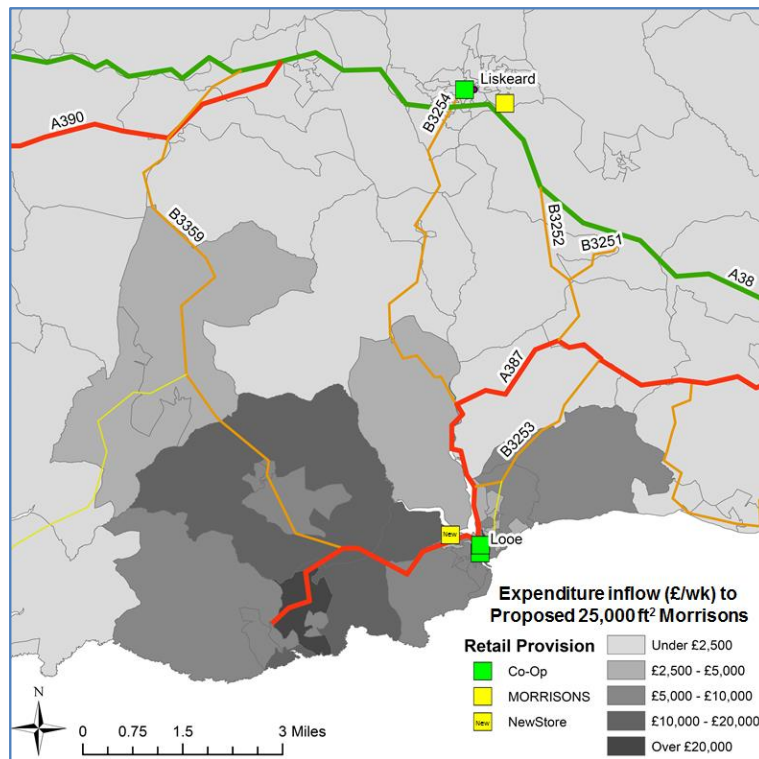
Morrisons have outlined interest in opening a new supermarket of around 25,000 ft<sup>2</sup> to serve the town via development of a brownfield edge-of-centre site at Polean in West Looe (Langford 2013) (shown on Figure 3). We use this development to demonstrate that our modelling approach can be used to assess proposed foodstore development in tourist resorts. Given word-limit constraints, we are not able to elaborate fully on the full range of modelled outputs or dwell in detail on the full impacts of this proposed store. Instead, we highlight the type of insight this modelling approach could provide, and consider broader implications in section 5.

Modelling suggests that the proposed store would generate average weekly revenue of around £300,000, with considerable seasonal sales fluctuations evident, driven by visitor demand (Figure 2). The store is modelled to trade at an average sales density of £12.23 ft<sup>2</sup>/week, well below the modelled company average (£17.83 ft<sup>2</sup>/week) for Morrisons' Cornish stores. However, sales densities are identified to increase to over £18 ft<sup>2</sup>/week during the August peak-season. As such, a store of this size is well-placed to cope with the summer seasonal influx of visitors and any population growth within Looe, but must address operational challenges driven by a very low sales density at times during the low-season, well below the usual levels experienced by grocery retailers. Seasonal visitor demand thus improves the viability of this store which provides much needed facilities for local residents, but where residential demand alone may not be sufficient to support this level of floorspace provision.



**Figure 2** Modelled store revenue (£/week) for proposed new Looe foodstore

The modelled expenditure inflow (Figure 3) indicates that the store would draw trade primarily from the town itself, alongside the rural and coastal catchment to the west of the town, offering considerable access benefits to residents and visitors to the town, reducing the average trip distance across the Looe catchment by 1.5km. The store would also help retain food and drink expenditure modelled at £369,869 per week (in August) within Looe which is currently attracted to stores elsewhere. 'Claw-back' of this form of expenditure would be likely to generate additional non-food spend in stores and services in Looe town centre, via linked trips, given the proposed foodstore edge-of-centre location.



**Figure 3** Modelled expenditure inflow (£/week by census Output Area) to proposed new Looe foodstore

Whilst offering much needed retail facilities and opportunities for linked-trips with other town centre stores and services, the modelled impact on the existing Co-Op stores in Looe suggests these stores will face a 65.6% sales reduction (52-Week average). Impacts would also be felt by the existing Morrisons' store in Liskeard, where 52-Week average sales are predicted to fall by 11.4% as a result of this investment. Nonetheless, following this investment, Morrisons' market share within the Looe catchment area increases by 30.7%, generating an overall net sales increase to the company. Our approach enables detailed impact assessment of this nature, incorporating underlying spatiotemporal demand estimates within a robust spatial model which considers consumer interactions with the supply side, offering considerable benefits to the retail planning process.

## 5. Implication for retail planning within tourist resorts

The incorporation of seasonal visitor demand within a modelling framework such as this enables complex location-based decision making and impact assessment to be undertaken. Using a live retail development scheme in a major tourist resort, we have demonstrated that our modelling approach can be used to assess the impact of proposed retail developments in tourist resorts, quantifying changes in consumer flows, store revenues and retailer market shares following new store development. Incorporation of visitor demand throughout the modelling process in this fashion allows location planners, developers and local planning authorities a more complete evidence base for store development and assessment of local economic impacts in tourist resorts. The approach could be applied to other retail or service sectors in highly seasonal tourist destinations where accurate estimation of the impact of demand uplift driven by tourism could help optimise service provision.

## 6. Acknowledgements

The research reported in this abstract was supported by an ESRC doctoral CASE Award (2010-2013)



as part of the Retail Industry Business Engagement Network.

## **7. Biography**

Andy Newing is a lecturer in Retail Geography at the University of Leeds. His research interests include applied spatial and quantitative analysis for retail location planning and analytics, consumer data analysis, health service delivery, census/neighbourhood analysis and geodemographics.

Graham Clarke is Professor of Geography at the University of Leeds. His research interests include GIS, urban services, retail and business geography.

Martin Clarke is Professor of Geographic Modelling in the School of Geography at the University of Leeds. From 1990 to 2004 he was Chief Executive of GMAP Ltd. Martin is Deputy Director of the ESRC Consumer Data Research Centre (CDRC).

## **References**

- Langford, R. 2013. *Letter from WM Morrison Supermarkets Plc to Cornwall Council dated 4th February 2013 Re: Polean, West Looe, Cornwall*. At: Bradford: Morrisons.
- Newing, A., Clarke, G. and Clarke, M. 2013. Identifying seasonal variations in store-level visitor grocery demand. *International Journal of Retail & Distribution Management*, **41**(6), pp.477-492.
- Newing, A., Clarke, G. P. and Clarke, M. 2014. Developing and applying a disaggregated retail location model with extended retail demand estimations. *Geographical Analysis*, Early view article DOI: 10.1111/gean.12052.

# The changing geography of deprivation in Britain: exploiting small area census data 1971 to 2011

Paul Norman School of Geography, University of Leeds

6<sup>th</sup> November 2014 (submission), 19<sup>th</sup> April 2015 (revised)

## Summary

This paper will describe the method being used to devise a time-series of area deprivation 1971 to 2011 using census data for all years harmonised to contemporary definitions of LSOAs / Datazones in GB. This involves identifying appropriate deprivation indicator variables from each census, converting data between boundary systems since these change over time and calculating deprivation such that change over time, rather than just cross-sectional deprivation, can be measured. Comparisons between censuses will be presented to show the degree of consistency or change in situation.

**KEYWORDS:** Deprivation; Area measures; Time-series; Census.

## 1. Introduction

Townsend (1987), defines deprivation as, “a state of observable and demonstrable disadvantage relative to the local community or the wider society or nation to which an individual, family or group belongs.” To identify small area deprivation, a wide variety of indexes have been devised which provide a single score which summarises information from several variables that each indicate something relating to deprivation. Various deprivation schemes / indexes exist including: Jarman Underprivileged Area, Townsend, Carstairs, Breadline Britain, Index of Multiple Deprivation (IMD) (Norman, 2010).

Many policy-related and academic studies use deprivation scores calculated cross-sectionally. A recent example in the health literature is by Maguire et al. (2015) in ‘Health & Place’ who investigate the relationship between area deprivation and the food environment over time in a repeated cross-sectional study on takeaway outlet density and supermarket presence in Norfolk, UK, 1990–2008. To do this, they link food outlet locations to wards but, “due to changing electoral ward boundaries, we were only able to use 2001 deprivation rather than capturing deprivation at multiple time points across the study period” (p. 143). Stratifying a time-series of an outcome across a single time point of deprivation is a common and pragmatic approach (see Norman and Fraser 2014, for example). However, as noted by Maguire et al. (2015, p. 146), “this approach may have introduced some error into our estimates of outlet density, and so future studies should utilise data where this information has been captured at multiple time points.”

Thus, it is useful to identify whether small areas have changed their level of deprivation over time and be able to assess the impact of area-based planning initiatives or determine whether a change in the level of deprivation leads to a change in health. However, the changing relationship with an outcome cannot be judged if the ‘before’ and ‘after’ situations are based on deprivation measures which use time point specific variables, methods and geographies. Changing deprivation is both a cause and a consequence of demographic change. As such, in areas with improving deprivation over time: infant mortality improves more than for other areas (Norman et al. 2008) and cancer survival improves more (Basto et al. 2014). In areas of persistent (dis-)advantage over time have the (worst) best self-reported health & mortality (Boyle et al. 2009; Norman et al. 2010; Exeter et al. 2011).

Identifying deprivation change presents various challenges. UK deprivation indexes (e.g. Townsend 1987; Carstairs & Morris 1989) have traditionally been at ward scale and predominantly based on census variables as indicators of relative conditions between areas. In recent Indexes of Multiple Deprivation (IMD) alternative geographies and input variables have been used (Nobel et al. 2006). A drawback with the IMDs is they should only be used for individual countries in the UK. A drawback with ward schemes is the uneven population sizes compared with the Lower Super Output Areas (LSOAs) and equivalents

used in the IMDs. Any cross-sectional scores are not comparable over time and geographical boundaries are liable to change.

This paper will describe the method being used to devise a time-series of area deprivation 1971 to 2011 using census data for all years harmonised to contemporary definitions of LSOAs / Datazones in GB. This involves identifying appropriate deprivation indicator variables from each census, converting data between boundary systems since these change over time and calculating deprivation such that change over time, rather than just cross-sectional deprivation, can be measured. Comparisons between censuses will be presented to show the degree of consistency or change in situation.

## 2. Methodological background

The development of methods to analyse demographic change in the face of small boundary change is relatively recent. Norman (2002) and in subsequent publications estimated a set of population related resources for GB and the UK through the development of methods: for geographical harmonisation when small area boundaries change (Norman et al. 2003; Norman 2006); of populations by age and sex, the estimation of the past (Rees et al. 2004; 2005; Norman et al. 2008) and projection of the future (Norman et al. 2010; Rees et al. 2010, 2011 & 2012); of the calculation of changing area deprivation (Norman, 2010a); and of the analysis of demographic change (Tromans et al. 2008; Norman 2010b; Norman 2011). The resources relate to the period 1981 to 2001 with very full detail (relevant to the purposes) though with less detail from 1971 to 1981 and after 2001. Various datasets have been deposited at the UK Data Archive (study numbers 5850, 6045 & 6777).

In applied work, these resources were used for health related research of; infant mortality (Norman et al. 2008), all cause mortality (Rees et al. 2003; Norman et al. 2011), cause specific mortality (Exeter et al. 2011); limiting long-term illness and incapacity benefit (Bambra and Norman 2006; Norman & Bambra 2007) and of children with life limiting conditions (Fraser et al. 2012; Norman & Fraser 2014). Further topics include small area analyses of local democracy (Norman et al. 2007), environmental equity (Mitchell & Norman 2012), traffic accidents (Lyons et al. 2009) and fire risk (Corcoran et al. 2007).

The examples above are *area* based; about whether aspects for small populations vary over space and time. Parallel to this, has been research which seeks to determine whether for *individuals*, there are different experiences for people who live in different kinds of places over time. As above, the focus is on health, particularly for persons: who move between levels of deprivation (Boyle et al. 2002; Norman et al. 2005) at different ages (Norman & Boyle 2014) or between urban and rural areas (Riva et al. 2011); who do not move residence (Boyle et al. 2004); or who are social mobile (Boyle et al. 2009); and where linkages to residential areas need estimation when specific locations are unclear or names of places have changed (Norman & Riva 2012).

The resources and methods have been applied in studies of general cancer (e.g. van Laar et al. 2010, 2012 & 2013); specific cancers (Basta et al. 2014; Blakey et al. 2014; McNally et al. 2012 & 2014), coronary heart disease (Bajekal et al. 2013a & b; Scholes et al. 2013); diabetes (Harron et al. 2010 & 2011) asthma (Hoskins et al. 2011 & 2012) and sensory impairment (Dawes et al. 2014; Dawes et al. 2015).

The resources and applicability of previous decisions on which boundaries to use have become dated. More recent data are now available (both census and demographic births and deaths events) but with the inevitable boundary and data definitional changes which were resolved in the previous work. There is a need then to update, to redefine and to ensure the resources are fit-for-purpose for long run time-series analysis from 1971 to 2011 and by contemporary geographies (2011 definitions). The latter ensures that interpretations are relevant to current applications. Thus there is an overall aim to produce for small area subnational areas in England, Wales and Scotland various datasets which comprise:

- 1971-2001 annual time-series of populations by five years age-groups and sex;
- Population density for the census years, 1971, 1981, 1991, 2001 and 2011;
- Deprivation scores and quantiles for the census years, 1971, 1981, 1991, 2001 and 2011;

- Sociodemographic variables (the inputs to deprivation measures and others).

To create the above requires data to be converted from their original geographies of dissemination (different at least once per decade) to the small area geographies for which the 2011 Census data were released: i.e. Super Output Areas in England and Wales and Datazones in Scotland. (Unfortunately, a lack of data over this time-frame precludes widening out the geographical coverage to include Northern Ireland but *some* data will be made available for 1991, 2001 and 2011.) Given that the National Statistician stated that there should be greater use of administrative data, comparisons will be made with schemes which characterise area deprivation using administrative data (Abejon & Norman, 2015; D'Silva & Norman, 2015)

This paper provides interim results on progress to date and reports on the calculation of changes in deprivation in England from 1971 to 2011 by the 2011 definition of the Lower Super Output Areas.

### 3. Deprivation by LSOAs in England, 1971 to 2011

The specification of the work here relates to a time frame of the census years: 1971, 1981, 1991, 2001 and 2011. The coverage of this initial study is England and the small areas used are the 2011 definitions of the LSOAs. The variable inputs obtained from the decennial censuses are the inputs to the Townsend deprivation scheme which relate to unemployment, non-home ownership, lack of car access and household overcrowding. Deprivation is then calculated which is comparable over time, rather than being just for a census year cross-section.

*Boundary change.* The small areas for which census data are disseminated change each decade by both terminology and by where they are placed. Older terminologies for the smallest areas are 'Enumeration Districts' (EDs) and newer are 'Output Areas' (OAs). At each census, these nest into electoral wards and / or Super Output Areas. The former have been the geography of choice for many deprivation schemes and the latter (particularly LSOAs) for official schemes and the release of administrative data since 2001. Even though they are designed to be frozen over time, even the LSOA boundaries have had some changes between 2001 and 2011.

The need here is to take original data as released for a geography relevant to a previous census (the 'source' geography) and convert to the contemporary geography of interest (the 'target' geography). Figure 1 shows the 1991 EDs (left) in Birmingham and the 2011 LSOAs (right). Whilst there are a few lines which correspond, the source and target geographies are very different.

**Figure 1: 1991 ED and 2011 LSOAs in Birmingham**



Since postcode distribution is a proxy for population distribution, the conversion between source and target geographies can be achieved for associating postcode points with the different geographies (Norman et al., 2003). The count of postcodes in the intersections of the boundary systems can be used to redistribute counts for the source geography into the target units. Figure 2 shows Sutton Coldfield in the northern part of Birmingham with the 1991 EDs and the 2011 LSOAs along with the postcode distribution (sized by address count). The less populous parts (including Sutton Park) have fewer postcodes than more densely populated areas.

**Figure 2: 1991 ED, postcode distribution and 2011 LSOAs in Birmingham**

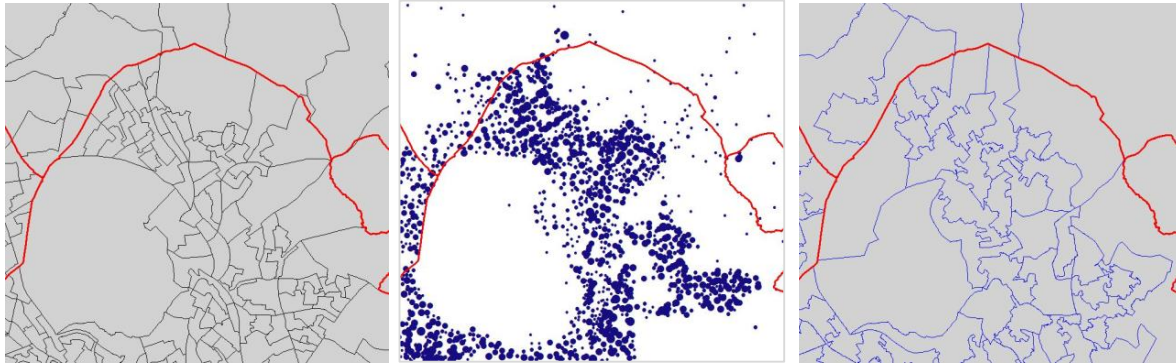


Table 1 shows the conversion weights between the 1991 EDs and the 2011 LSOAs. The weight is used to apportion the population counts, for 07CNGL05, for example, into the (four) LSOAs the area overlaps. The weights sum to one for each source area. The data are then aggregated across the target zones.

**Table 1: Conversion weights between source and target geographies**

ED91	LSOA11-Code	LSOA11-Name	Add	Source	Weight
07CNGL05	E01009415	Birmingham 002A	126	223	0.5650
07CNGL05	E01009416	Birmingham 003A	27	223	0.1211
07CNGL05	E01009418	Birmingham 001B	30	223	0.1345
07CNGL05	E01009423	Birmingham 002C	40	223	0.1794
07CNGL06	E01009417	Birmingham 001A	37	247	0.1498
07CNGL06	E01009421	Birmingham 004A	210	247	0.8502
07CNGL07	E01009417	Birmingham 001A	112	264	0.4242
07CNGL07	E01009419	Birmingham 001C	152	264	0.5758
07CNGL08	E01009419	Birmingham 001C	260	260	1.0000

*Input variables.* For the census years: 1971, 1981, 1991, 2001 & 2011, the numerators and denominators of unemployment, non-home ownership, no car access, household overcrowding and persons have been obtained at ED and OA level as appropriate and converted to LSOAs for 2011.

*Calculating comparable deprivation.* The conventional way to calculate the cross-sectional (census time point) Townsend deprivation scores is to transform variable proportions to near normal distributions as necessary, to standardise as z scores and then to sum the four z scores, unweighted into a single score. Even when converted to the same geography, a change in deprivation score cannot be interpreted as an improvement or not. To calculate time comparable deprivation, stack the data for successive census years and calculate the z scores relative to the indicator and national level for both / all years. In Figure 3, area 1 based on just this one variable, would change (9% to 10%). In this way, the data for all censuses 1971 to

2011 have been stacked with deprivation calculated on that basis. Population weighted quintiles (20% population in each) over time have also been calculated.

### Figure 3: Cross-sectional and comparable deprivation calculation

Cross-section: data for one year

Area	Variable	Z-score
1	9%	
2	11%	
3	11%	
4	14%	
...		
...	9%	
...	9%	
...	12%	
n	16%	

$$zscore = \frac{(Obs - Mean)}{SD}$$

Time comparable: data for more than one year

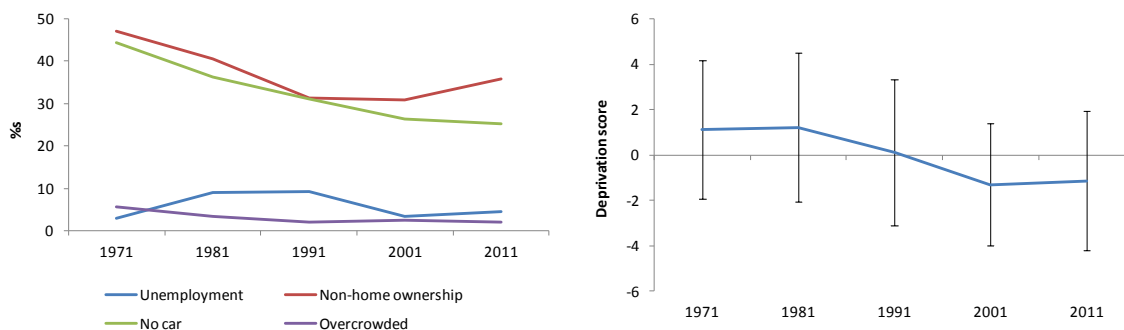
Area	Variable	Z-score
1	9%	
2	11%	
3	11%	
4	14%	
...		
...	9%	
...	9%	
...	12%	
n	16%	
1	10%	
2	12%	
3	9%	
4	16%	
...		
...	8%	
...	10%	
...	11%	
n	14%	

$$zscore = \frac{(Obs - Mean)}{SD}$$

### 4. LSOA deprivation change: England 1971 to 2011

Figure 4 shows on the left the average level of each of the four indicator variables over time. Lack of car access steadily decreases over time. Overcrowded households are rarer, reduce over time a little and seem to bottom out by 1991 through to 2011. Unemployment increases to 1981, is steady to 1991 and then reduces, staying at a similar level though a slight rise to 2011 (there is more change in the inter-censal periods, of course). Non home ownership is more intriguing with large reductions from 1971 to 1991, steady to 2001 and then a sharp increase to 2011. The average level of deprivation (figure 4 on the right) is higher in 1971 and 1981 and then reduces to 1991 and then 2001. To 2011 there is a small increase in deprivation driven by increases in non-home ownership and slight rise in unemployment.

### Figure 4: Deprivation change: England 1971 to 2011



The correlations reported in Table 2 show strong consistencies between successive censuses and over time. Essentially, this suggests that there are not major changes in the geography of deprivation over time.

**Table 2: Correlations between deprivation at each census: LSOAs in England 1971-2011**

	towns81	towns91	towns01	towns11
towns71	0.88	0.84	0.79	0.75
towns81		0.93	0.87	0.84
towns91			0.94	0.92
towns01				0.95

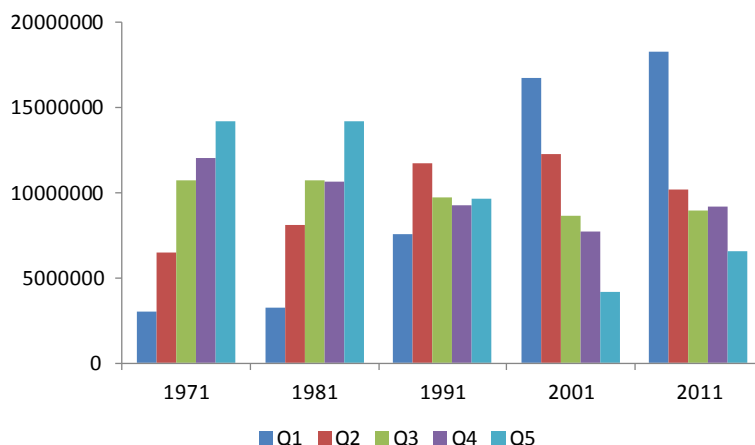
Using the quintiles of deprivation in 1971 and 2011, Table 3 shows that 28% of LSOAs are classified into the same quintile over time (the leading diagonal). 67% of LSOAs improve their categorisation of deprivation over time and 6% have a worse situation. These subsets can readily be used to analyse phenomena which might also have changed.

**Table 3: Cross-tabulation of deprivation quintiles in 1971 and 2011**

	Q1	Q2	Q3	Q4	Q5	Total
Q1	2047	271	86	37	8	2449
Q2	3582	967	339	142	20	5050
Q3	4133	2432	1104	403	53	8125
Q4	1515	2225	2556	1749	470	8515
Q5	272	490	1435	3272	3236	8705
Total	11549	6385	5520	5603	3787	32844
	Same	28	Better	67	Worse	6

Population change and deprivation change are linked by population migration and by people changing their attributes. Figure 5 shows how population is differently distributed across deprivation at each census. In 1971, the gradient shows fewer people living in less deprived areas through to more people living in more deprived areas. The extremes change little by 1981 but the mid deprived areas have changed. In 1991 there is little difference in population distribution across deprivation and by 2001 and 2011, population is redistributed such that more people are living in less, than more deprived locations.

**Figure 5: Population distribution across deprivation quintiles: England 1971 to 2011**





## 5. Further work

This work for England is soon to be extended to all of GB. A lack of data for Scotland until recently has held progress back. There will also be calculations of deprivation at different levels including MSOA and LA which will enable data relevant to those geographies to be analysed. Various health analyses will be carried out at LSOA and other levels including: infant mortality, all cause and cause specific mortality; being permanently sick / disabled. The variable inputs from censuses will also be broadened to enable tenure to be investigated as well as social class and education achievement. Population density will provide another area attribute.

## 6. Acknowledgements

This work used census data obtained via MIMAS' CASWEB facility and GIS boundary data obtained via EDINA's UKBORDERS facility services supported by ESRC and JISC. The census data for England have been provided by the Office for National Statistics and the digital boundary data by OSGB and OSNI. These data are Crown copyright and are reproduced with permission of OPSI.

## 7. References

- Ajebon M & Norman P (2015) Beyond the Census: A Spatial Analysis of Health and Deprivation in England. *GeoJournal* DOI:10.1007/s10708-015-9624-8
- Bajekal M, Scholes S, O'Flaherty M, Raine R, Norman P & Capewell S (2013a) Implications of using a fixed IMD quintile allocation for small areas in England from 1981 to 2007. *PLOS ONE* 8(3): e59608. doi:10.1371/journal.pone.0059608.S003
- Bajekal M, Scholes S, O'Flaherty M, Raine R, Norman P & Capewell S (2013b) Unequal trends in coronary heart disease mortality by socioeconomic circumstances, England 1982-2006: analytical study. *PLOS ONE* 8(3): e59608. doi:10.1371/journal.pone.0059608
- Bambra C & Norman P (2006) What is the association between sickness absence, morbidity and mortality? *Health & Place* 12: 728-733
- Basta NO, James PW, Gomez-Pozo B, Craft AW, Norman PD, McNally RJQ (2014) Survival from teenage and young adult cancer in northern England, 1968-2008. *Pediatric Blood & Cancer* DOI 10.1002/pbc.24939
- Blakey K, Feltbower RG, Parslow RC, James PW, Pozo BG, Stiller C, Vincent TJ, Norman PD, McKinney PA, Murphy MF, Craft AW, McNally RJQ (2014) Is fluoride a risk factor for bone cancer? Small area analysis of osteosarcoma and Ewing sarcoma diagnosed among 0-49 year olds in Great Britain, 1980-2005. *International Journal of Epidemiology* 43(1): 224-234 doi: 10.1093/ije/dyt259
- Boyle P, Norman P & Popham F (2009) Social mobility: evidence that it can widen health inequalities. *Social Science & Medicine* 68(10): 1835-1842
- Boyle P, Norman P & Rees P (2002) Does migration exaggerate the relationship between deprivation and limiting long-term illness? A Scottish analysis. *Social Science & Medicine* 55: 21-31
- Boyle P, Norman P & Rees P (2004) Changing places: do changes in the relative deprivation of areas influence limiting long-term illness and mortality among non-migrant people living in non-deprived households? *Social Science & Medicine* 58: 2459-2471
- Carstairs, V. & Morris, R. (1989). Deprivation: explaining differences in mortality between Scotland, England and Wales. *British Medical Journal* 299: 886-889.



- Corcoran J, Higgs G, Brunsdon C, Ware A & Norman P (2007) The use of spatial analytical techniques to explore patterns of fire incidence: a South Wales case study. *Computers, Environment and Urban Systems* 31: 623–647
- D'Silva S & Norman P (2015) Impacts of mine closure in Doncaster: an index of social stress. *Radical Statistics* (in press)
- Dawes P, Dickinson C, Emsley R, Bishop P, Cruickshanks K, Edmondson-Jones M, McCormack A, Fortnum H, Moore DR, Norman P & Munro K (2015) Understanding visual impairment in UK Biobank – Authors' Reply. *Ophthalmic and Physiological Optics* 35(1):107-108  
DOI:10.1111/opo.12178
- Dawes P, Fortnum H, Moore DR, Emsley R, Norman P, Cruickshanks K, Davis A, Edmondson-Jones M, McCormack A, Lutman M, Munro K (2014) Hearing in middle age: a population snapshot of 40- to 69-year olds in the United Kingdom. *Ear and Hearing* doi: 10.1097/AUD.000000000000010
- Exeter D J, Boyle P J & Norman P (2011) Deprivation (im)mobility and cause-specific premature mortality in Scotland. *Social Science & Medicine* 72: 389-397
- Fraser L K, Miller M, Hain R, Norman P, Aldridge J, McKinney P A & Parslow R C (2012) Rising national prevalence of Life Limiting Conditions in Children in England. *Paediatrics* DOI: 10.1542/peds.2011-2846
- Harron K, McKinney PA, Feltbower RG, Stephenson CR, Bodansky HJ, Norman PD, Chhokar G & Parslow RC (2011) Incidence rate trends in childhood Type 1 diabetes in Yorkshire, UK 1978-2007: effects of deprivation and age at diagnosis in the south Asian and non-south Asian populations. *Diabetic Medicine* 28, 1508–1513 doi:10.1111/j.1464-5491.2011.03413.x
- Harron K, McKinney PA, Feltbower RG, Stephenson CR, Norman PD, Bodansky HJ, Chhokar G & Parslow RC (2010) Ethnic differences in incidence rates of childhood Type 1 diabetes in Yorkshire 1978-2007. *Diabetologia* 53: S143-S143
- Hoskins G, Williams B, Jackson C, Norman P D & Donnan P T (2011) Assessing Asthma Control in UK Primary Care: Use of routinely collected prospective observational consultation data to determine appropriateness of a variety of control assessment models. *BMC Family Practice*, 12: 105  
doi:10.1186/1471-2296-12-105
- Hoskins G, Williams B, Jackson C, Norman P D & Donnan P T (2012) Patient, practice and organizational influences on asthma control. Observational data from a national study on primary care in the United Kingdom. *International Journal of Nursing Studies* 49(5) 596-609  
doi:10.1016/j.ijnurstu.2011.10.017
- Lyons R A, Ward H, Christie N, Macey S, Norman P & Griffiths S (2009) Road traffic injury and disadvantage: people and areas. In *Behavioural Research in Road Safety 2007*. Department for Transport. www.dft.gov.uk/pgr/roadsafety/research/behavioural/ 77-93
- Maguire, E. R., Burgoine, T., & Monsivais, P. (2015). Area deprivation and the food environment over time: A repeated cross-sectional study on takeaway outlet density and supermarket presence in Norfolk, UK, 1990–2008. *Health & Place* 33: 142-147
- McNally RJQ, Blakey K, Parslow RC, James PW, Pozo BG, Stiller C, Vincent TJ, Norman P, McKinney PA, Murphy MF, Craft AW & Feltbower RG (2012) Small area analyses of bone cancer diagnosed in Great Britain provide clues to aetiology. *BMC Cancer* 12: 270 doi:10.1186/1471-2407-12-270
- McNally RJQ, James PW, Ducker S, Norman PD, James OFW (2014) No rise in incidence but geographical heterogeneity in the occurrence of Primary Biliary Cirrhosis in northeast England. *American Journal of Epidemiology* 179(4):492-498 DOI: 10.1093/aje/kwt308
- Mitchell G & Norman P (2012) Longitudinal environmental justice analysis: Co-evolution of environmental quality and deprivation in England, 1960-2007. *Geoforum* 43: 44-57  
doi:10.1016/j.geoforum.2011.08.005

- Noble, M., Wright, G., Smith, G. & Dibben, C. (2006). Measuring multiple deprivation at the small area level. *Environment & Planning A* 38: 168-185.
- Norman P & Bambra C (2007) Unemployment or incapacity? The utility of medically certified sickness absence data as an updatable indicator of population health. *Population, Space & Place* 13(5): 333-352
- Norman P & Boyle P (2014) Are health inequalities between differently deprived areas evident at different ages? A longitudinal study of census records in England & Wales, 1991-2001. *Health & Place* 26: 88-93 <http://dx.doi.org/10.1016/j.healthplace.2013.12.010>
- Norman P & Fraser L (2014) Prevalence of life-limiting and life-threatening illness in children and young people in England: time trends by area type. *Health & Place* 26: 171-179 <http://dx.doi.org/10.1016/j.healthplace.2014.01.002>
- Norman P & Riva M (2012) Population health across space and time: the geographical harmonisation of the ONS Longitudinal Study for England and Wales. *Population, Space & Place* 18: 483-502 DOI: 10.1002/psp.1705
- Norman P (2006) Sociodemographic spatial change in the UK: data and computational issues and solutions. *GIS Development* 10(12): 30-34
- Norman P (2010) Identifying change over time in small area socio-economic deprivation. *Applied Spatial Analysis and Policy* 3(2-3) 107-138
- Norman P (2010b) Demographic and deprivation change in the UK, 1991-2001. In *Understanding Population Trends and Processes Volume 2: Spatial and Social Disparities* (eds.) John Stillwell, Paul Norman, Claudia Thomas & Paula Surridge. Springer: Dordrecht: 17-35
- Norman P (2011) Relationships between UK subnational trends in infant mortality and fertility. In *Population Dynamics and Projection Methods*, UPTAP Volume 4, (eds.) John Stillwell & Martin Clarke. Springer: Dordrecht: 99-114
- Norman P, Boyle P & Rees P (2005) Selective migration, health and deprivation: a longitudinal analysis. *Social Science & Medicine* 60(12): 2755-2771
- Norman P, Boyle P, Exeter D, Feng Z & Popham F (2011) Rising premature mortality in the UK's persistently deprived areas: Only a Scottish phenomenon? *Social Science & Medicine* 73 1575-1584 doi:10.1016/j.socscimed.2011.09.034
- Norman P, Gregory I, Dorling D & Baker A (2008) Geographical trends in infant mortality: England and Wales, 1970–2006. *Health Statistics Quarterly* 40: 18-29 <http://www.ons.gov.uk/ons/rel/hsq/health-statistics-quarterly/no--40--winter-2008/index.html>
- Norman P, Purdam K, Tajar, A & Simpson S (2007) Representation and local democracy: geographical variations in elector to councillor ratios. *Political Geography* 26 57-77
- Norman P, Rees P & Boyle P (2003) Achieving data compatibility over space and time: creating consistent geographical zones. *International Journal of Population Geography* 9(5): 365-386
- Norman P, Rees P, Wohland P & Boden P (2010) Ethnic group populations: the components for projection, demographic rates and trends. Chapter 14 in Stillwell, J. and van Ham, M. (eds.) *Ethnicity and Integration*. Series: Understanding Population Trends and Processes. Springer: Dordrecht: 289-315
- Norman P, Simpson L and Sabater A (2008) 'Estimating with Confidence' and hindsight: new UK small area population estimates for 1991. *Population, Space and Place* 14(5): 449-472
- Rees P, Brown D, Norman P & Dorling D (2003) Are socioeconomic inequalities in mortality decreasing or increasing within some British regions? An observational study, 1990-98. *Journal of Public Health Medicine*. 25(3): 208-214
- Rees P, Norman P & Brown D (2004) A framework for progressively improving small area population estimates. *Journal of the Royal Statistical Society A* . Vol. 167 Part 1: 5-36

- Rees P, Parsons J & Norman P (2005) Making an estimate of the number of people & households for Output Areas in the 2001 Census. *Population Trends* 122: 27-34
- Rees P, Wohland P & Norman P (2013) The demographic drivers of future ethnic group populations for UK local areas 2001-2051. *Geographical Journal* 179(1): 44-60 doi: 10.1111/j.1475-4959.2012.00471.x
- Rees P, Wohland P, Norman P & Boden P (2011) A local analysis of ethnic group population trends and projections for the UK. *Journal of Population Research* 28(2): 129-148 doi: 10.1007/s12546-011-9047-4
- Rees P, Wohland P, Norman P & Boden P (2012) Ethnic population projections for the UK, 2001-2051. *Journal of Population Research* 29(1): 45-89 DOI 10.1007/s12546-011-9076-z
- Riva M, Curtis S & Norman P (2011) Residential mobility within England and urban-rural inequalities in mortality. *Social Science & Medicine* doi:10.1016/j.socscimed.2011.09.030
- Scholes S, Bajekal M, Norman P, O'Flaherty M, Hawkins N, Capewell S, Raine R (2013) Quantifying Policy Options for Reducing Future Coronary Heart Disease Mortality in England: A Modelling Study. *PLOS ONE* 8(7): e69935. doi:10.1371/journal.pone.0069935
- Townsend, P. (1987). Deprivation. *Journal of Social Policy* 16: 125-46.
- Tromans N, Natamba E, Jefferies J & Norman P (2008) Have national trends in fertility between 1986 and 2006 occurred evenly across England and Wales? *Population Trends* 133: 7-19
- van Laar M, McKinney PA, Parslow RC, Glaser A, Kinsey SE, Lewis IJ, Picton SV, Richards M, Shenton G, Stark D, Norman P, Feltbower RG (2010) Cancer incidence among the south Asian and non-south Asian population under 30 years of age in Yorkshire, UK. *British Journal of Cancer* 103(9):1448-1452
- van Laar M, McKinney PA, Parslow RC, Glaser A, Kinsey SE, Lewis IJ, Picton SV, Richards M, Shenton G, Stark D, Norman P, Feltbower RG (2013) Cancer incidence among the south Asian and non-south Asian population under 30 years of age in Yorkshire, UK [Corrigendum]. *British Journal of Cancer* 108, 1223–1224 | doi: 10.1038/bjc.2013.67
- van Laar M, McKinney PA, Stark DP, Glaser A, Kinsey SE, Lewis IJ, Picton SV, Richards M, Norman P, Feltbower RG (2012) Survival trends of cancer among the south Asian and non-south Asian population under 30 years of age in Yorkshire, UK. *Cancer Epidemiology* 36(1): e13–e18

## 8. Biography

Paul Norman is a population & health geographer interested in time-series analysis of area and individual data from census, survey and administrative records. Paul did an MA GIS and PhD at the School of Geography, University of Leeds, was research fellow at CCSR before returning to Leeds as a Lecturer.

# Data Exploration with GIS Viewsheds and Social Network Analysis

Giles Oatley<sup>\*1</sup>, Tom Crick<sup>†1</sup> and Ray Howell<sup>‡2</sup>

<sup>1</sup>Department of Computing, Cardiff Metropolitan University, UK

<sup>2</sup>Faculty of Business and Society, University of South Wales, UK

## Summary

We present a novel exploratory method combining line of sight visibility (viewshed analysis) and techniques from social network analysis to investigate archaeological data. At increasing distances different nodes are connected creating a set of networks, which are subsequently described using centrality measures and clustering coefficients. Networks with significant properties are examined in more detail. We use this method to investigate the placement of hillforts (nodes) in the Gwent region of south-east Wales, UK. We are able to determine distances that support significant transitions in network structure that could have significant archaeological validity.

**KEYWORDS:** Geographic networks, archaeological nodes, viewshed analysis, data mining, social network analysis

## Extended Abstract

We present a novel exploratory method that combines line of sight visibility (viewshed analysis) with techniques from social network analysis to investigate archaeological data. Within data mining exist the fields of graph-based and spatial-based data mining. Graph-based data mining (Cook and Holder, 2006) has a close cousin in the long established field of social network analysis, a set of metrics that operates over graphs (networks) created from links (Wasserman and Faust, 1995). Metrics include those to find clusters within networks, to find points that have significant properties, for instance how central a point is. Spatial data mining likewise has an extensive history (Lu et al., 1993), and is the discovery of interesting patterns from spatial datasets.

At increasing distances different nodes are connected creating a set of networks, which are subsequently described using centrality measures and clustering coefficients. Networks with significant properties are examined in more detail. We use this method to investigate the placement of hillforts (nodes) in the Gwent region of south-east Wales, UK. Our methodology is applied to the area of

---

<sup>\*</sup>goatley@cardiffmet.ac.uk

<sup>†</sup>tcrick@cardiffmet.ac.uk

<sup>‡</sup>ray.howell@southwales.ac.uk

the Iron Age tribe known as the Silures, described as a ‘resilient and sophisticated clan based tribal confederation’ (Howell, 2009). Our preliminary investigation focuses on the Gwent region with a study area which roughly approximates the county as constituted between 1974 and 1996. Figure 1 shows the placement of 30 hillforts in this region. We are able to determine distances that support significant transitions in net-work structure that could have archaeological validity. Our study uses both geographical and graph/network structures, and presents an exploratory methodology within which to discover significant distances underlying network creation. While based on archaeological informatics, the approach has a more general use, for instance neural architectures, transportation networks, and other forms of geographical networks.

This research lies in the intersection of spatial and graph-based data. Related work includes that of the physics literature on geographical networks (ben Avraham et al., 2003), architectural analysis and the isovist literature including visibility graphs (Steadman, 1973; Llobera, 1996; Turner et al., 2001), and the authors’ recent work incorporating kernel density estimation into the betweenness social network metric (Oatley and Crick, 2014b,a). The data used includes the Iron Age hillfort data, provided from the Historic Environment Records<sup>1</sup>, and a Digital Elevation Model based on the Shuttle Radar Topography Mission data (UK SRTM DEM)<sup>2</sup> with 90m horizontal resolution.

We develop connectivity between Iron Age hillforts based on viewsheds and an increasing distance threshold. A viewshed is the area of land that is within line of sight from a fixed viewing position. We analyse the generated set of networks of connected hillforts using social network analysis, and use the metrics to inform theories of possible use and communication between hillforts. Degree centrality is simplest and is a count of the number of links to other nodes in the network. Closeness however is a measure of how close a node is to all other nodes in a network (Sabidussi, 1966). It is the mean of the shortest paths between a node and all other nodes reachable from it. Betweenness is the extent to which a node lies between other nodes in the network and is equal to the number of shortest paths from all nodes to all others that pass through that node (Freeman, 1977). This measure takes into account the connectivity of the node’s neighbours, giving a higher value for nodes which bridge clusters.

We explore using a local clustering coefficient (Watts and Strogatz, 1998) quantifying how close a networks nodes neighbours are to being a clique (fully connected). Viewsheds are generated for each hillfort, in order to determine intervisibility between every hillfort. We are then able to determine which hillforts are intervisible at any given distance threshold. In this way we investigate networks of hillforts at different distance values examining the clustering coefficient and betweenness measures.

This reveals several interesting transition points (see Figure 2) in connectivity, including localised clusters being evident, connectivity between larger regions, and connectivity along key geographic features such as along a shoreline and up waterways. In previous studies ‘significant’ distances and decay values have been determined a priori. We, however, examine the centrality of individual

---

<sup>1</sup>Archwilio, the Historical Environment Records of the Welsh Archaeological Trusts: <http://www.archwilio.org.uk/>

<sup>2</sup>UK SRTM DEM created by Addy Pope. Spatial Reference System–Great Britain National Grid: <http://edina.ac.uk/projects/sharegeo/>

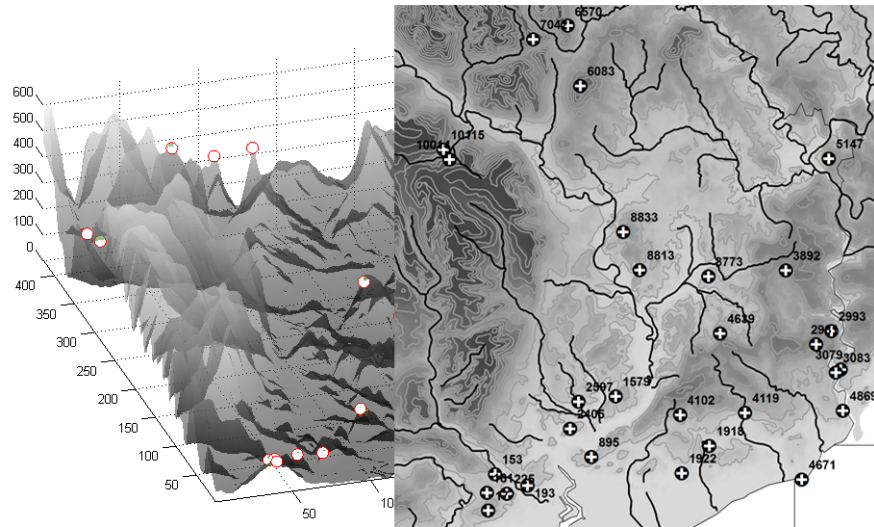


Figure 1: Hillforts in south-east Wales. Hillforts are displayed as white crosses on the front contour display. The same terrain and hillforts (white circles) are displayed behind on a Digital Elevation Model (DEM). The DEM display shows that there are many other sites that could have been used for placement of hillforts.

nodes (hillforts) in these networks with the most significant values. We are interested in discovering interesting patterns and clusters and then investigating them *a posteriori* for (archaeological) validity. Among preliminary conclusions arising from this first phase of investigation is that the methodology employed can effectively inform our understanding of Iron Age social structures. For example, viewshed analysis confirms hypothesised clan-based clustering of hillforts in the region with extensive line of sight communication, not only within clusters, but also with other hillfort groupings. The model of a clan-based confederation with regional emphasis, and possibly variation, but with wider connectivity sufficient to allow the cohesion necessary to have resisted the Roman advance so effectively seems wholly appropriate. Future work will utilise fuzzy viewsheds instead of the standard binary viewshed, with distance decay functions based on the limits of normal human vision and such features as the size of people, livestock, distances that smoke plumes can be seen and so on. We will also consider the integration of least-cost paths in landscapes.

## Biography

**Dr Giles Oatley** is a Reader in Intelligent Systems at Cardiff Metropolitan University. He has developed decision support systems based on behavioural models from data mining, primarily for UK police forces, supported by the EPSRC, Home Office, HEFCE, Nuffield Foundation, and DTI. He has a broad interest in anthropology and psychology, especially mindfulness and psychoanalysis.

**Dr Tom Crick** is a Senior Lecturer in Computing Science at Cardiff Metropolitan University. His

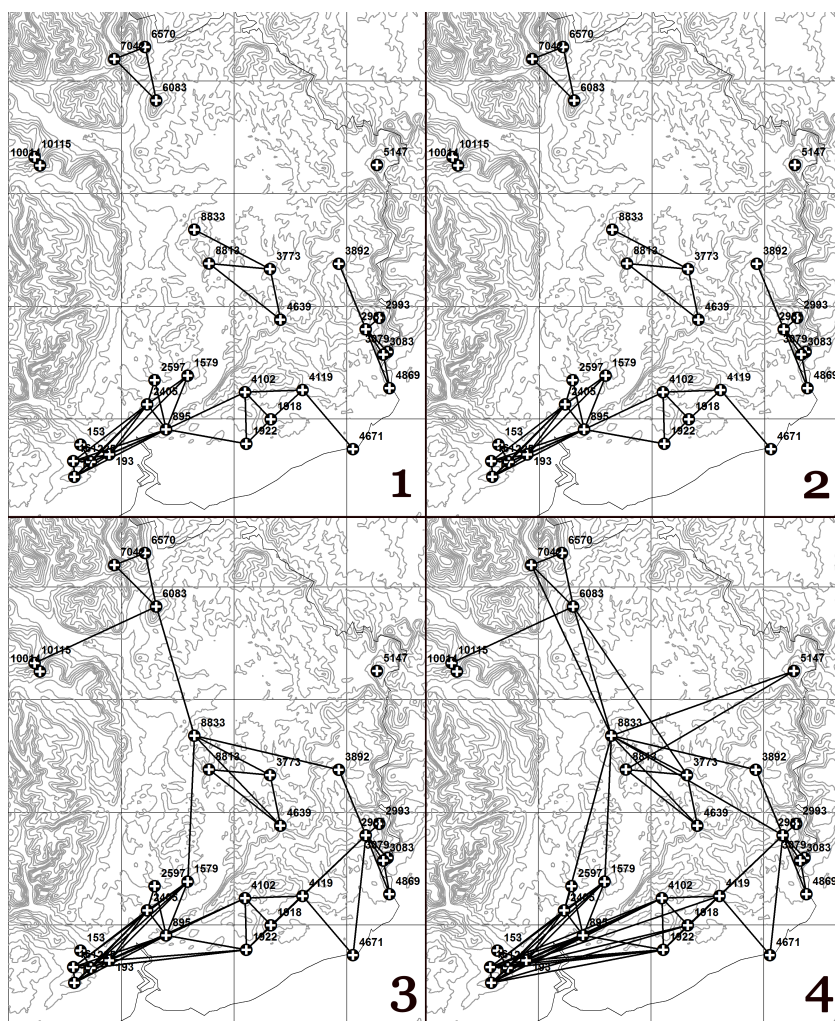


Figure 2: Interesting networks. 1: 5km. 2: 10km. 3: 15km. 4: 20km.

research is naturally interdisciplinary: optimisation, intelligent systems, data science and analytics, high performance computing and reproducibility. He is the NESTA Data Science Fellow, a 2014 Fellow of the Software Sustainability Institute and a member of *HiPEAC*, the European FP7 Network of Excellence on High Performance and Embedded Architecture and Compilation.

**Professor Ray Howell** is Professor of Welsh Antiquity and Director of the South Wales Centre for Historical and Interdisciplinary Research at the University of South Wales. He is a Fellow of the Society of Antiquaries of London. He is also Chairman of the Glamorgan Gwent Archaeological Trust and the Glamorgan Gwent Historic Environment Record Charitable Trust.

## References

- ben Avraham, D., Rozenfeld, A. F., Cohen, R., and Havlin, S. (2003). Geographical embedding of scale-free networks. *Physica A*, 330(1-2):107–116.
- Cook, D. J. and Holder, L. B. (2006). *Mining Graph Data*. Wiley.
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.
- Howell, R. (2009). *Searching for the Silures: The Iron Age in South-East Wales*. The History Press.
- Llobera, M. (1996). Exploring the topography of mind: Gis, social space and archaeology. *Antiquity*, 70(269):612–622.
- Lu, W., Han, J., and Ooi, B. (1993). Discovery of General Knowledge in Large Spatial Databases. In *Proceedings of the Far East Workshop on GIS (IEGIS'93)*, pages 275–289.
- Oatley, G. and Crick, T. (2014a). Exploring UK Crime Networks. In *2014 International Symposium on Foundations of Open Source Intelligence and Security Informatics (FOSINT-SI 2014)*. IEEE Press.
- Oatley, G. and Crick, T. (2014b). Measuring UK Crime Gangs. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE Press.
- Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.
- Steadman, P. (1973). Graph-theoretic representation of architectural arrangement. *Architectural Research and Teaching*, 2(3):161–172.
- Turner, A., Doxa, M., O’Sullivan, D., and Penn, A. (2001). From isovists to visibility graphs: A methodology for the analysis of architectural space. *Environment and Planning B: Planning and Design*, 28(1):103–121.
- Wasserman, S. and Faust, K. (1995). *Social Network Analysis: Methods and Applications*. Cambridge University Press.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442.



# A Framework for Big Data in studies of Urban Mobility and Movement

Eusebio Odiari<sup>1</sup>, Mark Birkin<sup>3</sup>, Susan Grant-Muller<sup>2</sup>, Nick Malleson<sup>1</sup>

Corresponding author: Eusebio Odiari ([gyeao@leeds.ac.uk](mailto:gyeao@leeds.ac.uk))

<sup>1</sup>School of Geography, University of Leeds, Leeds LS2 9JT, UK

<sup>2</sup>Institute for Transport studies, University of Leeds, Leeds LS2 9JT, UK

<sup>3</sup>Consumer Data Research Centre, Leeds Institute for Data Analytics, LS2 9JT, UK

## Summary

Various papers and strategies studying urban mobility and movement patterns are reviewed and, categorised according to the mathematical discipline of the models employed. This has enabled the development of a structured framework for studying urban mobility and movement patterns, while taking advantage of the opportunities that big data presents.

The choice of strategy for investigating urban mobility and movement is dependent on the type of data and, the particular problem, questions or situation that needs answers. The various types of Big Data come with volume, velocity, variety and veracity, and corresponding ever widening range of solution strategies, calling for a systematic approach based on an objective framework.

The particular opportunity that today's data analyst has is the freedom to introduce a new strategy better describing the situation being modelled, and then exploiting the flexibility in today's simulation tools, not limited or constrained as in the past, by purely mathematical concerns. That ability or skill to revise simulation code to suite a particular modelling scenario is a main strength.

For illustration purposes, an Agent Based Model (ABM) is used to simulate a railway infrastructure and its populated environs, and data is measured at various points to simulate types of consumer Big Data typically available within economic sector operators. This data is used to review solution techniques adopted in urban mobility papers, while developing solution concepts for illustration to sector operators, with a view to garnering support for the release of real-life data for future research.

**KEYWORDS:** Urban, Mobility, Movement, Patters, Framework, Big-Data, Mathematical discipline, Modelling, Model, Agent Based

# A national-scale application of the Huff gravity model for the estimation of town centre retail catchment area

Michail Pavlis<sup>1</sup>, Les Dolega<sup>1</sup>, Alex Singleton<sup>1</sup>

<sup>1</sup>Department of Geography and Planning, School of Environmental Sciences, University of Liverpool

November 7, 2014

## Summary

This work presents the application of an unconstrained retail gravity model based on the Huff algorithm. The main objective of the analysis was to estimate the retail catchment areas for town centres in England as a function of relative attractiveness and distance from potential customers, while taking into account the effects of competition among the retail destinations. This was achieved by relaxing the areal constraint for spatial interactions, allowing thus the relative size and attractiveness of the town centres to define the scale at which competition might occur.

**KEYWORDS:** Retail gravity model, Huff model, National scale

## 1. Introduction

This study is concerned with the problem of estimating the extent and volume of potential customer patronage flows between retail centres, that is, forming an estimate of the catchment area of a retail centre. Perhaps the most commonly used techniques to tackle such a problem belong to a family of models known as gravity models. The fundamental hypothesis of gravity models is that the likelihood of a person from an origin location  $i$  patronising a retail centre in destination  $j$  is inversely proportional to distance and influenced by some measure of attractiveness of the retail centre (Wilson, 1974). In other words the closer a retail centre and the more attractive it is, the more likely it will be patronised by customers from a given location. Historically, the development of gravity models derived analogies from Newtonian physics.

The Huff gravity model is one of the most commonly used models for catchment estimation and can be defined as follows (Huff, 1964):

$$P_{ij} = A_j^a D_{ij}^{-b} / \sum (A_j^a D_{ij}^{-b}) \quad (1)$$

Where  $P_{ij}$  the probability of a consumer from a location  $i$  patronising a retail centre  $j$ ,  $A_j$  a measure of attractiveness of the retail centre  $j$ ,  $D_{ij}$  a measure of distance between  $i$  and  $j$ ,  $a$  (alpha) the exponent of attractiveness and  $b$  (beta) the exponent of distance. Thus, the numerator of the Huff algorithm is obtained by calculating the product between the store attractiveness raised to a power indicating the degree of attractiveness of that store, with the inverse distance raised to a power indicating how fast the attractiveness of that store decays, divided by the sum of all products between a given point of origin and every retail centre within the study area. Essentially, the denominator of the Huff model provides a way of standardising the numerator so that the sum of probabilities for each point of origin add up to 1 and, hence, allow consideration of the effect of competition between retail centres. Another important attribute of the Huff probabilities is that the probability of a location patronising a given destination is independent of the probability of another location patronising the same destination, and hence they could be calculated separately if required.

For the development of the distance decay parameter ( $D_{ij}$ ) the time and money costs associated with travel is the most useful proxy, followed by the shortest road distance and the Euclidean (straight) distance (Wilson, 1974). Concerning the exponent of distance ( $b$ ) it is more realistic to be disaggregated by some characteristic of the destination (e.g. retail centre size) (Drezner and Drezner, 2002) or by some characteristic of the customer (e.g. whether or not is a car owner) (Birkin et al., 2010). It has also been suggested that the distance decay parameter could be better approximated as an exponential function (rather than the power function presented in equation 1), with the advantage of the exponential function argued as providing more realistic results (Drezner and Drezner, 1996). The retail centre attractiveness score ( $A_j$ ) is often derived from some measure of the relative size of a retail

destination (Moryadas and Lowe, 1975), nevertheless additional data have also been used (when available) (De Beule et al., 2014) and could offer a more accurate indicator of attractiveness. Concerning the attractiveness exponent ( $\alpha$ ) it is not often used (set to a default value of 1), nevertheless, it could offer a way of weighting the attractiveness score either based on a characteristic of the retail centre or of the points of origin.

## 2. Developing the Huff model

A Huff gravity model was developed to estimate the retail catchment area of 1192 retail centres in England. The boundaries of the retail centres relate to those DCLG definitions defined in 2004. The centroids of 32,843 Lower Super Output Areas (LSOA) were used as the origin locations of customers patronising the retail centres. In addition, the UK national road network was collated from the Meridian 2 dataset provided by Ordnance Survey.

A composite indicator for the attractiveness of the retail centres was created using the following data:

- 1) Retail centre size (number of units),
- 2) Number of comparison retailers,
- 3) Number of leisure units,
- 4) Number of anchor stores,
- 5) Number of vacant outlets

Each variable was first standardised to the range between 1 and 100. Consequently the composite indicator was calculated by summing the values of the first four variables and subtracting from the result the vacant outlets units. Following this the retail centres were divided into four groups based on the attractiveness score and these groups formed the basis for the development of the beta coefficient. More specifically, town centres with score above 100 are all metropolitan retail centres (e.g. Manchester, Birmingham), while town centres with score between 50 and 100 are regionally important retail areas such as Bolton. Town centres with score between 20 and 50 serve as district centres such as Boston and Warwick. Finally, half of the town centres had a score below 20 and represent locations that have small and local retail catchments. The beta exponent was disaggregated into four different values, one for each group of town centres. These values were 1.2, 1.4, 1.6 and 2.0, in decreasing order of town centre size.

The shortest road distance from each LSOA centroid to the boundary of each retail centre was employed as measure of distance in this study. Given the variable shape and extent of retail centres, this was found to produce more realistic catchments than when simply using a centroid of a retail centre. This was calculated by extracting the coordinates of the points which defined the retail area boundaries and, consequently, applying the Dijkstra algorithm in order to calculate the shortest road distance. The analysis was all performed in R using the rgraph library. Following this, the minimum distance of each unique pair of LSOA – retail centres was obtained. This estimated distance also formed the basis for the development of an attractiveness exponent ( $\alpha$ ). More specifically, any retail centre within 0.5 kilometres from the centroid of a LSOA was assumed to be the primary retail destination and hence the attractiveness score for that pair was raised to the power of two. For all other distances, the default alpha value equal to 1 was used.

## 3. Results

The presentation of the results of the analysis focuses on the area and population of the primary and secondary catchments of the retail centres. As primary catchment was considered in this analysis the area that had a Huff probability equal to or larger than 0.5 of patronising a retail centre. A secondary catchment was considered to be the area that had a Huff probability equal to or greater than 0.2 (note that primary catchments are nested within secondary catchments). The results from the application of the Huff model are presented in Table 1 as the absolute and the percentage value of the area and population share for each of the four types of retail centres. Out of 1192 retail centres 1163 were assigned a primary and/or secondary catchment area while 29 were not (all were smaller retail centres). The effect of the exponents on the output of the model was investigated first by decreasing the beta values by 0.1 (Model 2 in Table 1) and then increasing them by 0.1 (Model 3 in Table 1). In

addition, the effect of using only the default alpha coefficient (equal to 1) is also shown in Table 1 as Model 4. Finally, the output of a model with the same values as the Base Model when the distance is measured to the centroids of the retail centres is also shown in Table 1 as Model 5.

Table 1. Comparison of the output of 5 Huff models in terms of area and population assigned to the primary and secondary catchment, aggregated based on the size of the retail centres.

		Area		Population	
Total		130.537 Km <sup>2</sup>		53.010.253	
Model	Rank	Primary (%)	Secondary (%)	Primary (%)	Secondary (%)
Base Model beta = 1.2, 1.4, 1.6, 2.0 alpha = 1, 2	1	1.233 (0.94)	11.357 (8.7)	4.229.864 (7.98)	11.571.711 (21.83)
	2	1.128 (0.86)	5.704 (4.37)	3.651.869 (6.88)	9.194.608 (17.34)
	3	1.308 (1.00)	4.259 (3.26)	3.752.775 (7.08)	8.347.951 (15.75)
	4	745 (0.57)	1.731 (1.32)	2.377.897 (4.48)	3.992.822 (7.53)
Model 2 beta = 1.1, 1.3, 1.5, 1.9	1	826 (0.63)	8.767 (6.72)	3.361.106 (6.34)	10.831.968 (20.43)
	2	724 (0.55)	4.087 (3.13)	2.820.000 (5.32)	7.833.338 (14.78)
	3	940 (0.72)	3.116 (2.39)	3.046.818 (5.75)	6.924.254 (13.06)
	4	608 (0.46)	1.362 (1.04)	2.143.857 (4.04)	3.422.134 (6.45)
Model 3 beta = 1.3, 1.5, 1.7, 2.1	1	1.793 (1.37)	14.025 (10.74)	5.118.533 (9.65)	12.050.600 (22.73)
	2	1.595 (1.22)	8.347 (6.39)	4.556.380 (8.59)	10.335.884 (19.50)
	3	1.792 (1.37)	5.905 (4.52)	4.683.244 (8.83)	9.611.359 (18.13)
	4	940 (0.72)	2.238 (1.71)	2.722.559 (5.13)	4.619.869 (8.71)
Model 4 alpha = 1	1	1.242 (0.95)	11.380 (8.72)	4.287.560 (8.09)	11.727.393 (22.12)
	2	1.125 (0.86)	5.710 (4.37)	3.610.125 (6.81)	9.243.030 (17.44)
	3	1.267 (0.97)	4.259 (3.26)	3.432.624 (6.47)	8.336.511 (15.73)
	4	642 (0.49)	1.695 (1.3)	1.768.890 (3.34)	3.724.097 (7.02)
Model 5 Distance to Centroids	1	767 (0.59)	10.207 (7.82)	2.841.208 (5.36)	11.297.676 (21.31)
	2	744 (0.57)	4.924 (3.77)	2.032.737 (3.83)	7.911.645 (14.92)
	3	754 (0.58)	3.285 (2.51)	1.814.316 (3.42)	6.941.614 (13.09)
	4	392 (0.30)	1.289 (0.99)	1.231.337 (2.32)	2.802.417 (5.29)

Based on the comparison of the five different models it is obvious that when decreasing the beta values of the Huff model, the area and population of the secondary and primary catchments, on average, are also decreasing across all groups of retail centres. This might be due to greater competition and cannibalisation of the Huff probabilities among the retail centres as the distance decay parameter gets smaller for all of the retail centres. On the other hand, Model 3 shows that when the beta values are increasing this results in greater market share, most likely due to reduced competition among the retail centres. As it was noted in Section 1, the alpha parameter is often overlooked, however, its use could be useful to model certain scenarios that could be quite difficult to model otherwise. For example, by increasing the alpha exponent for distances smaller than 0.5 km, it is possible to model the behaviour of customers preferring a retail centre simply due to proximity and convenience to reach. The comparison between the Base Model and Model 4 shows that it is possible to model such a scenario, given that it is mostly the primary catchment area of small retail centres which increases when the former model is applied. Finally, it can be seen from Table 1 that there is a striking difference between the Base Model that uses the distance to the boundary of the retail centres and Model 5 that uses the distance to the centroid of the retail centres. Across all retail centres the Base Model predicts a greater primary and secondary catchment area, which might indicate that using

the centroids of the destinations could result in underestimating a catchment extent. In addition, it should be noted that the estimated catchment areas appear potentially more realistic for the Base Model as they follow the physical extent of the retail centres (Figure 1 A), compared to Model 5 which produces catchment areas that have more circular shapes (Figure 1 B).

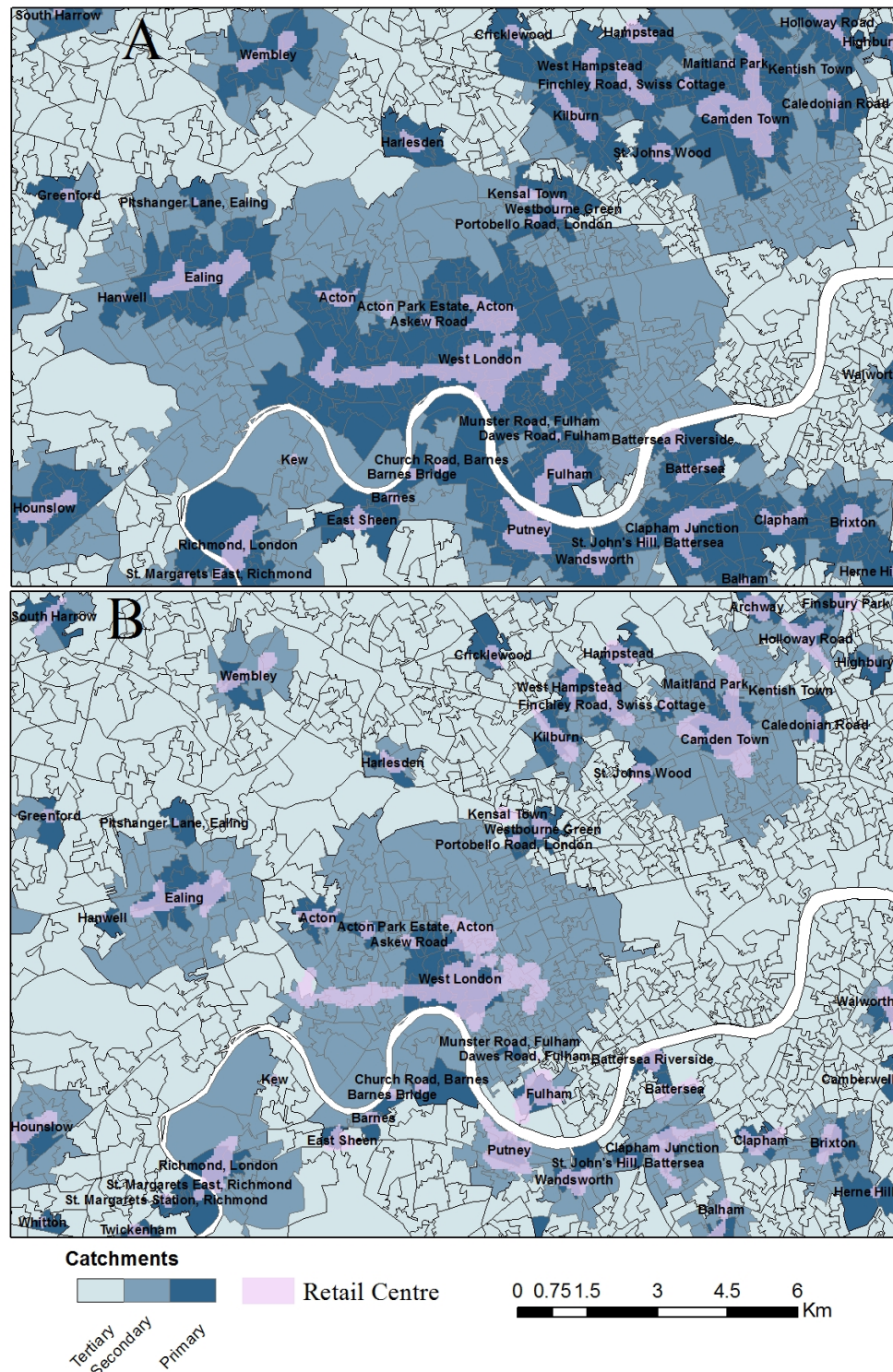


Figure 1. Comparison between the Base Model (distance to boundary) and Model 5 (distance to centroid).

## Conclusion

The analysis presented here has shown that by removing the areal constraint of the Huff model it is possible to develop a national-scale model of catchment areas for the retail centres in England. The analysis also investigated the effect of the Huff model parameters on the estimated catchment area and suggested that a more accurate representation of the catchment areas might be obtained when the distance to the boundaries of the retail catchments is used instead of the distance to their centroids. The next step in the analysis would be to calibrate and test the model using data of customer behaviour.

## References

- Birkin M, Clarke G and Clarke M (2010). Refining and operationalizing Entropy-Maximizing model for business applications. *Geographical Analysis*, 42(4), 422-445.
- De Beule M, Van den Poel D, Van de Weghe N (2014). An extended Huff-model for robustly benchmarking and predicting retail network performance. *Applied Geography*, 46, 80-89.
- Drezner T and Drezner Z (1996). Competitive facilities: Market share and location with random utility. *Journal of Regional Science*, 36(1), 1-15.
- Drezner T and Drezner Z (2002). Validating the gravity-based competitive location model using inferred attractiveness. *Annals of Operations Research*, 111, 227-237.
- Huff DL (1964). Defining and estimating a trading area. *Journal of Marketing*, 28, p. 34-38.
- Moryadas S and Lowe JC (1975). *The geography of movement*. Houghton Mifflin Company, p. 333.
- Wilson AG (1974). *Urban and regional models in geography and planning*. John Wiley & Sons, p. 418.

# Development and application of a two stage hybrid spatial microsimulation technique to provide inputs to a model of capacity to walk and cycle.

Ian Philips\*

Institute for Transport Studies  
University of Leeds

November 7, 2014

## Summary

This paper demonstrates the development and application of a two stage hybrid static spatial microsimulation technique. The first stage makes best use of Simulated Annealing with available micro-data, and the second uses Synthetic Reconstruction to add attributes not available in a single micro-data source. The new technique is applied to Leeds UK to generate a synthetic population which can be used as an input to a model of capacity to commute using only walking and cycling.

## KEYWORDS:

Hybrid spatial microsimulation method, Simulated Annealing, Synthetic Reconstruction, walking and cycling, transport planning model.

## 1. Introduction

In transport applications spatial microsimulation is often used to generate a synthetic population of individuals as a start point for a transport planning model (e.g. Beckman et al., 1996; Frick and Axhausen, 2004; Müller and Axhausen, 2010). A transport planning model of individuals' capacity to make journeys by walking and cycling has been constructed (the details of this model is reported elsewhere and is not the focus of this paper). It required a synthetic population of individuals as inputs. This paper focuses on the production of that synthetic population.

Existing populations were not suitable as these populations did not contain all of the attributes required to estimate capacity to walk or cycle used in the model. Table 1 shows the attributes required by the model. In addition to this correlated constraint attributes were selected. The synthetic population had to represent the variation in the physical capacity of individuals to walk and cycle. For example, Parkin (2008) expressed the importance of considering human power output in models of cycling. It should also consider constraints on walking and cycling such as bicycle availability and the need to escort children as part of a commute.

The population had to be available at a fine spatial resolution (Output Areas for UK applications). This is because journey origins were based on zone centroids. Typical trip length for walking and cycling is short (generally under 8km). This means that using the centroid of a large zone would introduce considerable error into estimates of distances (Iacono et al., 2010). Additionally being able to report results at a fine spatial resolution allows aggregation to coarser resolutions within a hierarchy and can be used to demonstrate heterogeneity to decision makers. Using the results in this way reduces the dangers of making poor decisions affected by the ecological fallacy.

---

\* Gy09ip@leeds.ac.uk

Table 1: Attributes required in the synthetic population

$VO_{2max}$ a measure of fitness in terms of the body's ability to make use of oxygen for exercise Physical activity Body Mass Index Age Gender Weight	<i>These attributes are used to derive pedalling power and the model uses this to estimate an individual's cycling speed and a corresponding value for walking speed.</i>
Bicycle availability  The need to escort children on the way to or from work Current commute distance	

There are several existing static spatial micro-simulation techniques (see Hermes and Poulsen, 2012 for an introduction). The suitability of existing spatial microsimulation techniques was examined. Simulated Annealing based combinatorial optimisation was reported as the best performing technique, particularly at Output Area resolution (Williamson, 2012; Harland et al., 2012). Synthetic Reconstruction (using Monte-Carlo sampling) is useful when a micro-data sample is not available (Barthelemy and Toint, 2012). The practical issues of constructing a population for Output Areas in the UK city of Leeds were examined. Not all of the required attributes were available in a single micro-data sample however the majority were available in the 2008 Health Survey for England.

The principal problems of using an existing technique are: Firstly, to use an existing technique would require sacrificing either spatial microsimulation performance (if Synthetic Reconstruction were used) or limiting the inputs to the indicator (if Simulated Annealing were used). Secondly, some attributes cannot be assigned to an individual until that individual has been allocated a location. For example, commute distance is strongly associated with location as well as individual socio-demographic attributes. Commute distance is collected in some micro-data surveys, but, it would not be appropriate to allocate this value out of its original spatial context. This is because individual survey data has geographical detail removed. This means that some individuals will not be allocated to areas to where they actually live. These problems led to development of a new hybrid technique.

## 2. A hybrid 2 stage spatial microsimulation technique

The two stage hybrid method works follows: In the first stage, a single synthetic population is constructed using Simulated Annealing In the application the open source FMF software was used (see Harland, 2013). The available micro-data is used as a sample population and constraint tables are taken from the census. In the second stage, Monte-Carlo sampling (Synthetic Reconstruction) is used to add attributes which are not available in the micro-data or which are geographically dependent. Monte-Carlo sampling can then be used to draw multiple synthetic populations.

This approach makes progress towards addressing the problems above: The performance benefits of Simulated Annealing are used with available data. Though Monte-Carlo sampling introduces increased computing time and data storage requirements, this is less of a drawback than it once was. A greater gain is made because it allows the full range of desired attributes to be modelled rather than having a model constrained by limited data sources. Introducing Monte-Carlo sampling also introduces a source of stochastic variation.

This is not a problem if a suitable number of draws is made; the standard error of the mean should not be excessive. Because only a minority of attributes are being added using Monte-Carlo sampling, the



stochastic variation between draws should be less than if the entire population was built using Synthetic Reconstruction. This will give an overall advantage in terms of performance. The process is outlined in Figure 1.

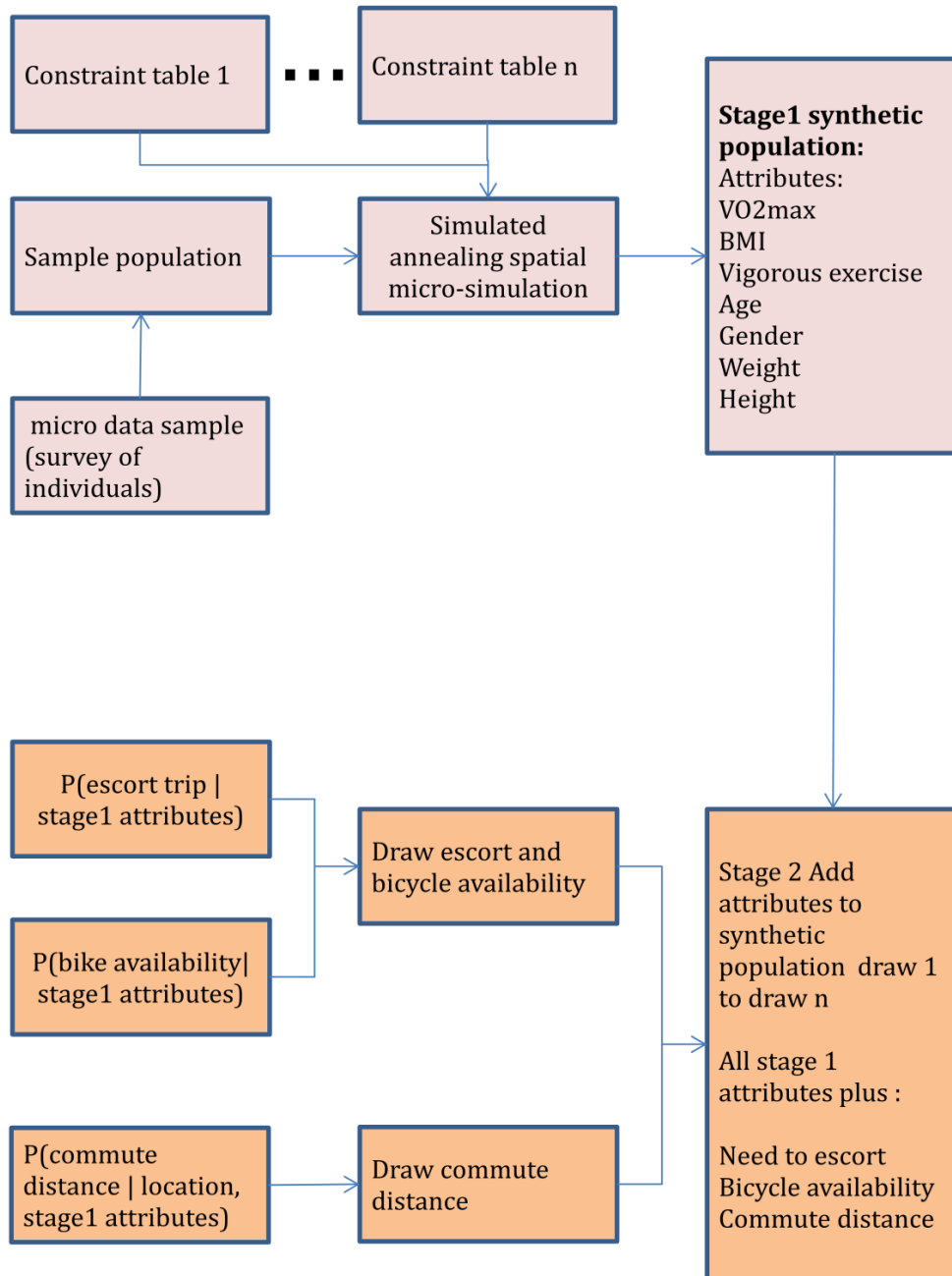


Figure 1: Outline diagram showing the two stage hybrid spatial microsimulation method.

### 3. Validation

Internal and external validation tests were performed using established validation techniques, TAE based measures and Z-scores (See Edwards and Tanton, 2012 for discussion of these techniques). The sensitivity of the final model result resulting from the stochastic variation and the different construction of constraints was estimated as less than  $\pm 5.1\%$  in 95% of Output Areas. This was acceptable for the application of the model.

### 4. Results

The spatial pattern of both the model output and the attributes contributing to it were mapped. The distribution of contributing attributes was internally consistent. Mapping individual attributes aids analysis of where specific attributes exert influence over the model. For example age affects pedalling power. The age structure in Figure 2 clearly shows an influence on the spatial distribution of pedalling power in Figure 3.

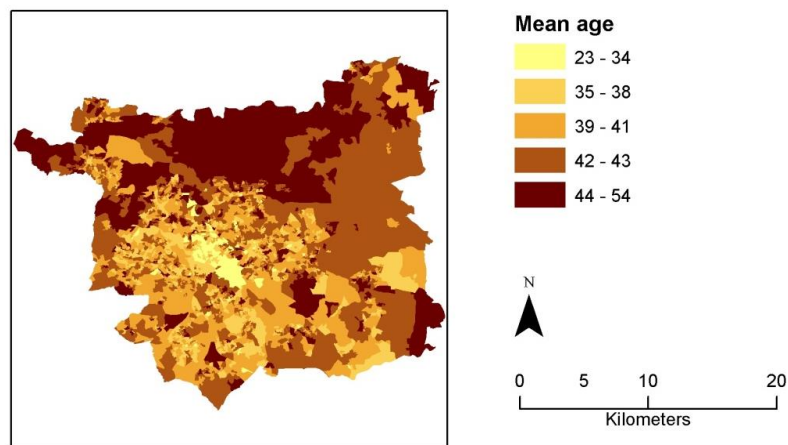


Figure 2 Mean age of working population in Leeds Output Areas

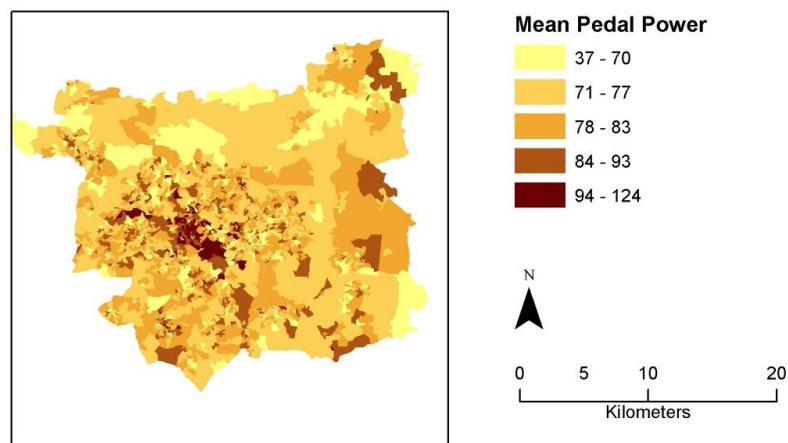


Figure 3: Mean Pedal Power of working population by Leeds Output Areas

Further analysis was conducted mapping other attributes output from the spatial microsimulation which contribute to the final model output shown in Figure 4.

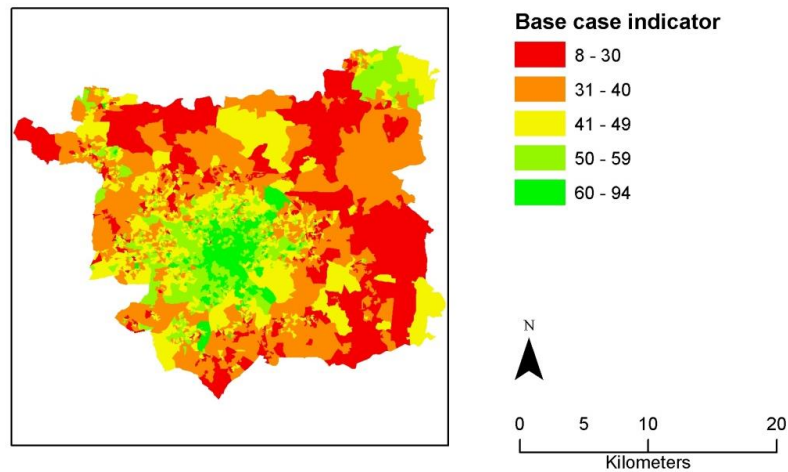


Figure 4. Model output based on the outputs of the hybrid spatial microsimulation method: Percentage of working population with capacity to walk or cycle to their current place of work in a network with no motorised traffic. Leeds Output Areas

## 5. Conclusion

The conclusion of the work is that this method can be usefully applied and is particularly useful where a synthetic population with many attributes is required as an input to a further modelling process.

## 6. References

- Barthelemy, J., Toint, P.L., 2012. Synthetic Population Generation Without a Sample. *Transportation Science*. doi:10.1287/trsc.1120.0408
- Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice* 30, 415–429. doi:10.1016/0965-8564(96)00004-3
- Edwards, K.L., Tanton, R., 2012. Validation of Spatial Microsimulation Models, in: Tanton, R., Edwards, K. (Eds.), *Spatial Microsimulation: A Reference Guide for Users*. Springer Netherlands, Dordrecht, pp. 249–258.
- Frick, M., Axhausen, K.W., 2004. Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. Presented at the Swiss Transport Research Conference, Ascona.
- Harland, K., 2013. Microsimulation model user guide Flexible Modelling Framework.
- Harland, K., Heppenstall, A., Smith, D., Birkin, M., 2012. Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *JASSS* 15, 1.
- Hermes, K., Poulsen, M., 2012. A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems* 36, 281–290. doi:10.1016/j.compenvurbsys.2012.03.005
- Iacono, M., Krizek, K.J., El-Geneidy, A., 2010. Measuring non-motorized accessibility: issues, alternatives, and execution. *Journal of Transport Geography* 18, 133–140. doi:10.1016/j.jtrangeo.2009.02.002
- Müller, K., Axhausen, K.W., 2010. Population synthesis for microsimulation: State of the art. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).
- Parkin, J., 2008. NECTAR Workshop Abstract The importance of human effort in planning networks. Presented at the NECTAR workshop.
- Williamson, P., 2012. An Evaluation of Two Synthetic Small-Area Microdata Simulation

Methodologies: Synthetic Reconstruction and Combinatorial Optimisation, in: Tanton, R., Edwards, K. (Eds.), *Spatial Microsimulation: A Reference Guide for Users*. Springer Netherlands, Dordrecht, pp. 19–47.

## **7. Acknowledgements**

This work was carried out as part of an ESPRC CASE PhD studentship partnered with Sustrans. I would like to acknowledge the help and support of my PhD supervisors David Watling and Paul Timms.

## **8. Biography**

Ian Philips has recently completed his PhD on resilience to fuel shocks. This involved creating a spatially explicit indicator of who could get to work by walking and cycling if there was no fuel for motorised transport. Ian's interests include spatially explicit modelling to inform sustainable transport policy.

# Combining Statistics and Texts Using GIS: Nineteenth Century Health Reports

Catherine Porter<sup>1</sup>, Paul Atkinson<sup>1</sup> and Ian Gregory<sup>1</sup>

<sup>1</sup>Department of History, Lancaster University, Lancaster LA1 4YT  
<http://www.lancaster.ac.uk/spatialhum/>

Nov 5, 2014

## Summary

This paper combines Geographic Information Systems (GIS) and Natural Language Processing (NLP) to explore how statistics and textual information may be compared. Combined with known mortality figures, for the first time, this research provides a spatial picture of the relationship between the Registrar-General's discussion of disease and deaths in England and Wales during the nineteenth and early twentieth centuries. A variety of techniques are employed to provide a new view on whether government published texts were directly related to changing mortality patterns during this time.

**KEYWORDS:** GIS; Natural Language Processing; Spatial Analysis; Infant mortality; Registrar General

## 1. Introduction

From the early nineteenth century the General Register Office for England and Wales (GRO) has been tasked with the registration and collation of records on births, deaths and marriages (Higgs, 2004). The Registrar-General's reports on mortality including cause of death are a key source for demographic history. The large and at times controversial literature on nineteenth century changes in mortality was well summarised by Woods (2000). However, hitherto no project has attempted to investigate the relationship between the Registrar-General's discussion of mortality and actual population mortality figures.

With the advent of applied technologies in the humanities such as Geographic Information Systems (GIS) and Natural Language Processing (NLP) such research is now possible. In this paper, said technologies have been utilised to extract disease related keywords from the Registrar-General's Decennial Supplements, and combined with known mortality figures, for the first time, provide a spatial picture of the relationship between the GRO's discussion of disease and actual deaths in England and Wales in the nineteenth and early twentieth centuries. This combination of corpus and GIS analysis provides a new view on the nineteenth and early twentieth century world according to the Registrar-General and as such, insight into whether his published texts were directly related to changing mortality patterns during this time.

## 2. Project data

In this study the primary dataset used to explore the relationship between the Registrar-General's writings and mortality is the Registrar-General's reports and the accompanying Decennial Supplements, available online at [www.histpop.org](http://www.histpop.org). The period 1850-1911 is selected here because of scholarly interest in the major decline in mortality which it witnessed: it is also the earliest covered by reporting on cause of death by local areas. These published documents, in addition to statistical tables,

---

<sup>1</sup> c.porter2@lancaster.ac.uk

<sup>1</sup> p.atkinson3@lancaster.ac.uk

<sup>1</sup> i.gregory@lancaster.ac.uk

include a discussion of the types of diseases and related places of interest to the Registrar-General during these decades. From these data it is therefore possible to gain not only the actual mortality figures for the time period, but also a discussion of the related diseases and places in which these diseases largely occurred or indeed were less prevalent.

For the GIS portion of the research the Registration Districts in vector polygon format were utilised as the basis for the study, the Registration Districts being the basis on which the GRO collected and collated population data. As well as this, the Hierarchical Regional Settlement matrix (HRS); a method based on that used by Gregory (2008), was employed. This matrix contains a group of numbered cells, 1-64, each of which describes a set of Registration Districts in terms of their distance from London, urban or rural, and core or peripheral. The most urban and core places are those Registration Districts that make up the London area, the most rural and peripheral signifying those Registration Districts in places such as Anglesey, Cornwall, the Lake District and Northumberland, and those between representing a variety of Registration Districts of varying distance from the capital. The importance of utilising the HRS matrix is therefore founded on the basis that in simplifying the data, in this case the Registration Districts and associated disease mentions and mortality data, it provides an overview of historical, geographical and statistical patterns for England and Wales.

### 3. Data preparation

The data preparation for this paper is threefold. The first, develops and establishes three main categories of disease for analysis, the second, concentrates on the extraction of the Registrar-General's mention of disease related to place, and the third, focuses on the actual mortality statistics derived from the Registrar-General's reports.

#### 3.1 Devising disease categories

Researchers are familiar with the difficulties created by the Registrar-General's changing classification of cause of death (Hardy, 1994). Woods and Shelton's *An Atlas of Victorian Mortality* (Woods, 1997) discusses how far causes may be linked in equivalent groups from one Decennial Supplement to the next, and proposes three analytical groups of disease: diseases of crowding, those related to food and water borne disease, and respiratory diseases (excluding tuberculosis, examined separately). These categories (Table 1) provide the basis for the analysis that follows.

**Table 1** The three disease categories used in the analysis

<b>Crowding</b>	<b>Food and Waterborne</b>	<b>Respiratory</b>
Diphtheria	Cholera	Bronchitis
Measles	Diarrhoea and Dysentery	Diseases of Lungs
Scarlatina	Enteric Fever	Disease of Respiratory System
Scarlet Fever	Simple Continued Fever	Influenza
Small-pox	Typhoid	Pneumonia
Typhus	Disease of the Digestive System	
Whooping cough		

#### 3.2 Extraction of disease related text and the georeferencing of disease data

Using the *Geographical Collocates Tool* (GCT), a front end application developed at Lancaster University for the extraction of disease terms that collocate with place, the diseases related to the chosen categories were obtained from Histpop to include the corresponding coordinate data for each disease collocating with place (georeferencing of the text was completed in Edinburgh by Grover et al., 2010). These coordinate data allow for the spatial analysis which follows by providing the addition of point data of disease mentions in the GIS software. As well as disease, place and coordinate data, the tool also provides a snapshot of the text in which the disease related words and places are mentioned and as such, provides context to the extracted information (Table 2).

**Table 2** Key columns derived from the GCT including the disease under investigation and the associated text, as well as the collocated place and related coordinates. The columns Left\_text and Right\_text refer to the text either side of the key word.

Place	Latitude	Longitude	Left_text	Disease	Right_Text
Epsom	51.33030891	-0.27019611	6 in the place of 8400 . The Epsom district suffered from scarlatina ; Guild ford from small-pox and	measles	; Farnham from fever , measles , hooping cough , and diarrhoea . The deaths for the first time exce
Farnham	51.21092987	-0.790143132	6 in the place of 8400 . The Epsom district suffered from scarlatina ; Guild ford from small-pox and	measles	; Farnham from fever , measles , hooping cough , and diarrhoea . The deaths for the first time exce
Epsom	51.33030891	-0.27019611	som district suffered from scarlatina ; Guild ford from small-pox and measles ; Farnham from fever ,	measles	, hooping cough , and diarrhoea . The deaths for the first time exceed the births in Farnham . In t
Farnham	51.21092987	-0.790143132	som district suffered from scarlatina ; Guild ford from small-pox and measles ; Farnham from fever ,	measles	, hooping cough , and diarrhoea . The deaths for the first time exceed the births in Farnham . In t
Gloucester	51.86437225	-2.239719987	ewhat less than the counties of the previous Division . The mortality was high in , Hereford , where	measles	was epidemic ; and somewhat above the average in Gloucester , Shrewsbury , Stafford , Worcester , a
Shrewsbury	52.70746422	-2.747530341	ewhat less than the counties of the previous Division . The mortality was high in , Hereford , where	measles	was epidemic ; and somewhat above the average in Gloucester , Shrewsbury , Stafford , Worcester , a
Stafford	52.80866623	-2.111278772	ewhat less than the counties of the previous Division . The mortality was high in , Hereford , where	measles	was epidemic ; and somewhat above the average in Gloucester , Shrewsbury , Stafford , Worcester , a
Worcester	52.19711876	-2.212242126	ewhat less than the counties of the previous Division . The mortality was high in , Hereford , where	measles	was epidemic ; and somewhat above the average in Gloucester , Shrewsbury , Stafford , Worcester , a

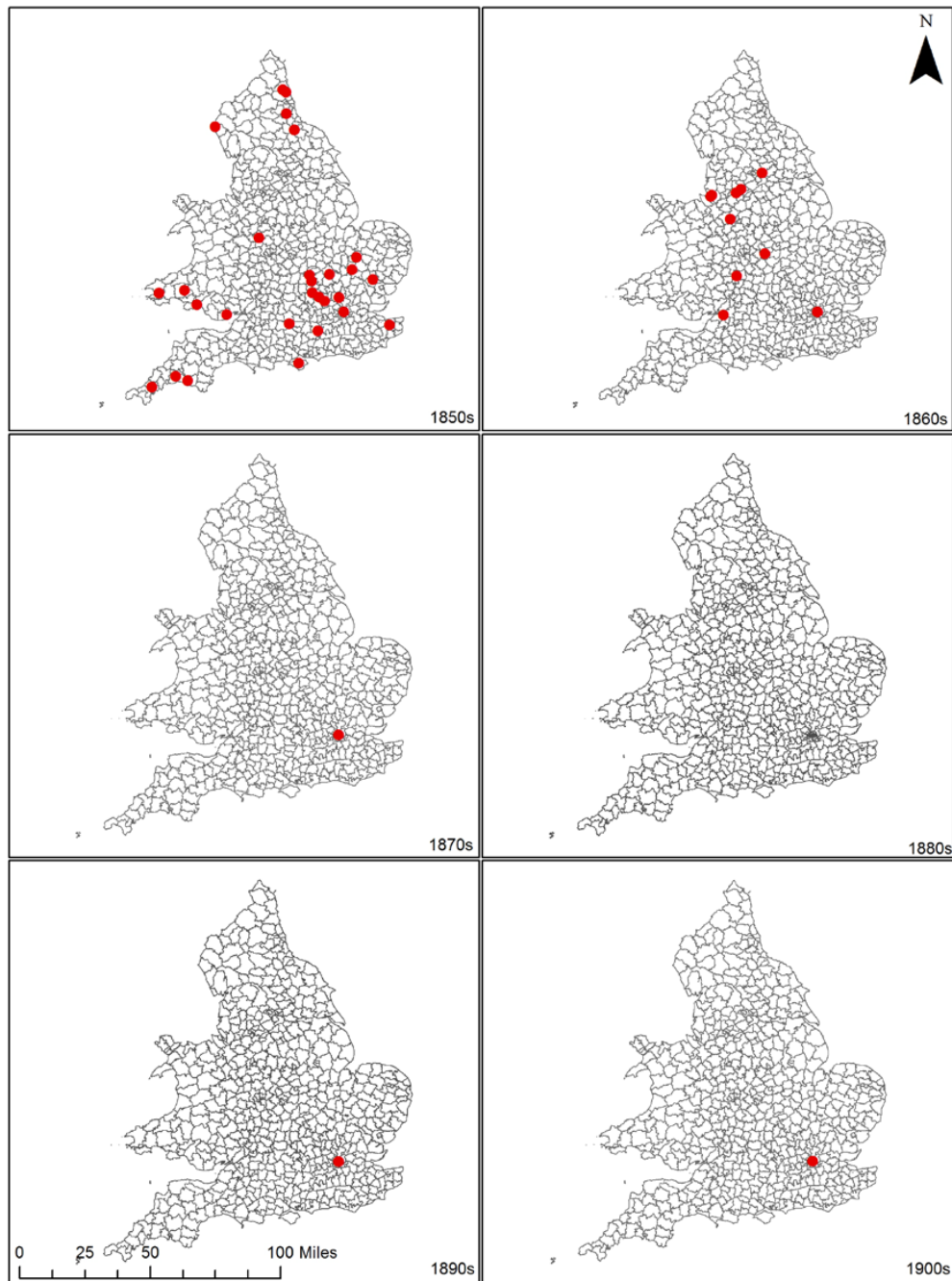
### 3.3 Deriving the mortality statistics

Decennial mortality figures were extracted from the Great British Historical GIS (GBHGIS) database (Gregory et al., 2002). Infant mortality (age under one year) was chosen as a focus because of its salience as an indicator of overall health conditions (Titmuss, 1943). The data have been organised according to the Registration Districts allowing for spatial analysis based on these polygon data and presenting the possibility of a relationship assessment between this and the Registrar-General's discursive work also under analysis.

## 4. The analysis process

Within the three core disease categories the data were subjected to a number of analytical processes based firstly on the Registrar-General's discussion of disease, and secondly, the mortality figures derived from GBHGIS. The Registrar-General's mentions of disease were first mapped spatially and temporally in the GIS as point data, the frequencies of the data based on disease category and decade as shown in Figure 1, this example being discussion of respiratory diseases. This provides a first glimpse of the varying temporal and spatial degree with which the Registrar General discussed disease throughout England and Wales, the greatest proportion of mentions appearing to relate to large settlements.

## Temporal Mentions of Diseases Related to Respiration



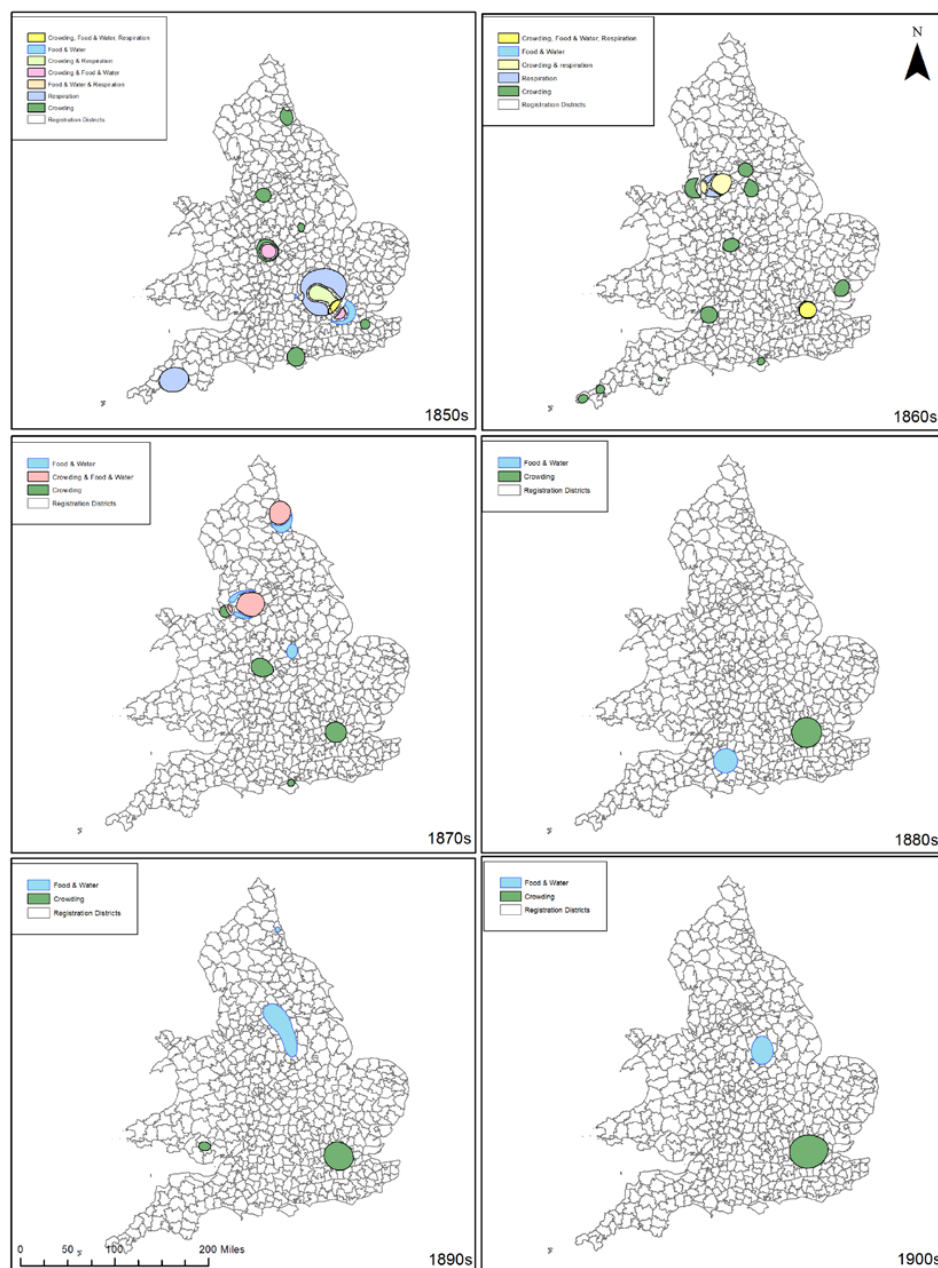
**Figure 1** The Registrar-General mentions related to respiratory diseases mapped on the Registration Districts for England and Wales and displayed for each decade of the research time frame

To expand on this, next, a density smoothing process was run on these data, again for each decade and for each of the three core disease categories, producing a set of polygons that pinpointed the regions with the greatest density of disease mentions. These polygons were then ‘unioned’ to determine where instances of the three core disease mentions overlapped (Figure 2). This amalgamation of disease data



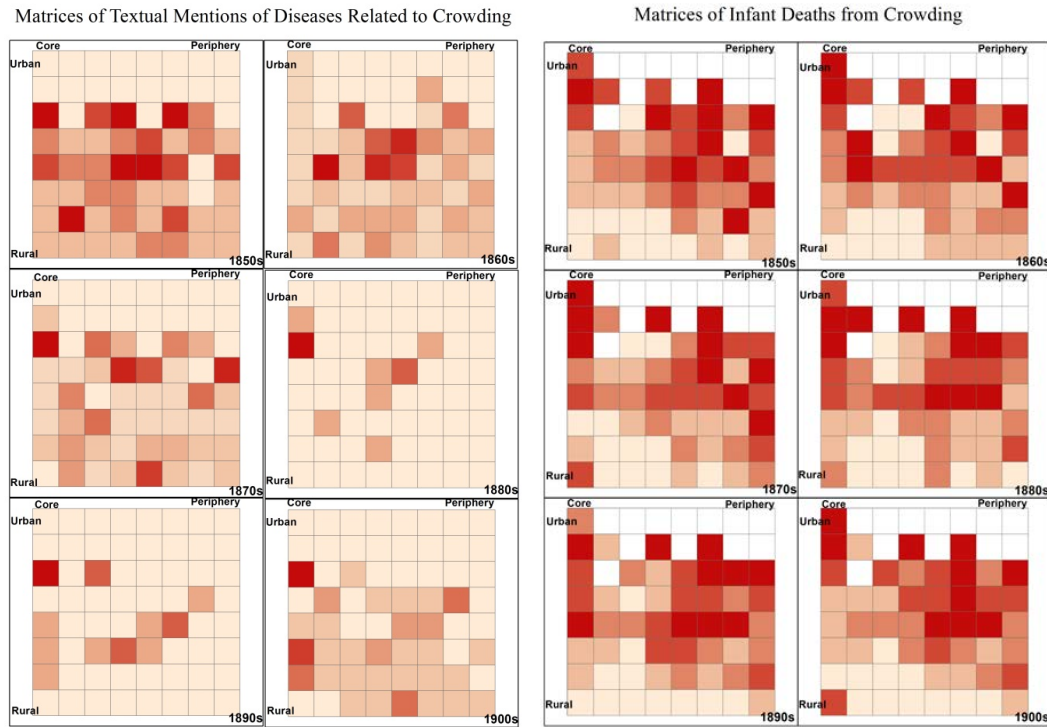
provides a combined picture of the Registrar General's discussion of the core diseases. For instance, a concentration on large and increasingly industrialised settlements such as, London, Manchester, Liverpool and Newcastle Upon Tyne is noteworthy, but only in the 1850s and 1860s did all three disease categories mentioned by the Registrar-General overlap in the same place, London (Figure 2).

### The Intersection of Statistically Significant Clusters for the 3 Disease Classes



**Figure 2** The 'unioned' density polygons of disease mentions for each of the three disease classifications and decades of the research time frame. The areas in yellow show where all three classifications coincide.

Thirdly, the Registrar-General's mentions were mapped to the HRS matrices to afford a more generalised geographical picture of the temporal and spatial patterns of the Registrar-General's discursive work (Figure 3, left). As before, the outputs present some tendency of the Registrar-General to concentrate on larger settlements, particularly London, however, the results also highlight that the focus of his interest often lay in those Registration Districts classed as core-rural and rural-peripheral.



**Figure 3** The Hierarchical Regional Settlement matrices for crowding diseases, the mentions by the Registrar General (left) and the mortality figures (right).

Lastly, and for comparative purposes, the mortality data were also incorporated into the HRS matrices allowing the two sets of data to be compared both visually and through statistical analysis. The mortality figures (Figure 3, right), showed a lack of coincidence with the disease mentions (Figure 3, left), the majority of deaths being located in the core-urban and core-peripheral regions. This comparison was further tested by employing regression analysis between each set of disease mentions and mortality data, confirming the HRS outputs by resulting in little, or in some cases, no significant correlation between the two sets of data ( $P$ -values  $> 0.05$ ).

## 5. Conclusions

The conclusions of this paper are twofold. Firstly, they confirm the importance of such mixed method approaches in the digital humanities, how the incorporation of text with GIS can help gain new understanding of textual and statistical sources. Secondly, the outcomes of the analyses show the spatial relationship between the Registrar-General's focus of attention and mortality. London and the larger industrial places were often the focal point of the Registrar-General's discussion of disease, but noticeably the places his texts explored did not always correlate with the highest death rates.

## 6. Acknowledgements

This research has been funded by the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant 'Spatial Humanities: Texts, GIS, places' (agreement number 283850).

## References

Gregory I N (2008). Different places, different stories: Infant mortality decline in England & Wales. 1851-1911. *Annals of the Association of American Geographers*, 98, 773-794.

Gregory I N, Bennett C, Gilham V L and Southall H R (2002). "The Great Britain Historical GIS: From maps to changing human geography" *The Cartographic Journal*, 39, 37-49.

Grover C, Tobin R, Woollard M, Reid J, Dunn S and Ball J (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A*, 368, 3875-3889.

Hardy A (1994). 'Death is the cure of all diseases': using the General Register Office cause of death statistics for 1837-1920. *Social history of medicine* 7(3), 472-92

Higgs E (2004). *Life, death and statistics: civil registration, censuses and the work of the General Register Office, 1836-1952*. Local Population Studies, Hatfield.

Histpop (2014) – The Online Historical Population Reports Project, [online], Available: <http://www.histpop.org> [October 2014].

Titmuss R (1943) *Birth, Poverty and Wealth: a Study of Infant Mortality*. Hamish Hamilton London.

Woods R I, Watterson P A and Woodward J H (1988). The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part I. *Population Studies* 42, 343-366.

Woods R I, Watterson P A and Woodward J H (1989). The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part II. *Population Studies* 43, 113-132.

Woods R I (1993). On the Historical Relationship Between Infant Mortality and Adult Mortality. *Population Studies* 47(2), 195-219.

Woods R I and Shelton N (1997). *Atlas of Victorian Mortality*. Liverpool University Press, Liverpool.

Woods R I (2000). *The demography of Victorian England and Wales*. Cambridge University Press, Cambridge.

## Biography

*Catherine Porter holds a PhD in Geography and works as a Research Associate for the 'Spatial Humanities: Text, GIS, Places' project at Lancaster University. Her research interests involve the use of Geographic Information Systems (GIS) in the digital humanities, particularly the use of quantitative methodologies and spatial analysis for the investigation of historical geography and early maps.*

*Paul Atkinson holds a PhD in Modern History and works as a Senior Research Associate on the Spatial Humanities project at Lancaster University. He researches the nineteenth-century social history of Britain, and is currently working on infant mortality, using the Great British Historical GIS database and the analysis of digitised text.*

*Ian Gregory is professor of Digital Humanities in the Department of History at Lancaster University. He is PI on the ERC-funded 'Spatial Humanities: Text, GIS, Places' project. His main research interests concern applying GIS to the humanities.*

# ALTERNATIVE APPROACHES TO FORECASTING MIGRATION: FRAMEWORK AND UK ILLUSTRATIONS

Philip Rees<sup>11</sup>, Nikolas Lomax<sup>12</sup>, Peter Boden<sup>23</sup>

<sup>1</sup>Centre for Spatial Analysis and Policy, School of Geography, University of Leeds, Leeds LS2 9JT

<sup>2</sup>Edge Analytics Ltd, Leeds Innovation Centre, 103 Clarendon Road, Leeds LS2 9DF

February 2015

## Summary

This paper is a review of the migration component of population projection models. A general population accounting framework is defined that underpins projection models. The framework identifies migration variables internal to a country, international migration to and from a country and its constituent regions and international migration between other countries. These different types of migration can be represented in projection models as flows (migration numbers), flows projected by a time series model, migration transmission rates multiplied by the origin population at risk, migration admission rates multiplied by the destination population at risk, or through an explanatory model. The arguments for and against each migration projection model are discussed through an analysis of 16 projection examples linked to published studies. The importance of understanding the forces affecting migration at origin and destination is stressed. Simulation experiments are carried out for a UK population disaggregated by the ethnicity, testing model results against 2001-2011 estimates. The paper shows how a proper understanding of the spatial context is needed for successful population projections.

**KEYWORDS: Population Projection Models, Accounting Frameworks, Internal Migration, International Migration**

## 1. INTRODUCTION

Population projections are carried out by international agencies (e.g. UN 2014, World Bank 2014), demographic research centres (e.g. Lutz et al 2014), national statistical offices (e.g. ONS 2013 for the UK and Home countries; ONS 2014a for subnational areas), local authorities (e.g. GLA 2014 for London Boroughs) or academic teams (e.g. Rees et al 2011 for local authorities in England by ethnicity). These projections figure significantly in policy debates on climate change, food crises, demographic ageing, ethnic transitions, regional development, planning and housing.

The research question posed in this paper is “which is the best migration model to use in a population projection?” Of course, this is an impossible question to answer without knowing the objectives of the projection exercise, the nature of the populations being studied and the character of the migration data available. So instead we evaluate the range of alternative methods available. Along the way we point to some ingredients that should be included in the migration models for projection: full understanding of migration data available, the need to embed migration variables within a population accounts framework,

---

<sup>1</sup> p.h.rees@leeds.ac.uk

<sup>2</sup> n.m.lomax@leeds.ac.uk

<sup>3</sup> pete@edgeanalytics.co.uk

and testing which migration model has performed best against a historical series, with attention to the limits within which we can be confident future migration flows lie.

Except at the very smallest spatial scales, the usual projection method used is the cohort-component model. This model tracks by age and sex the changes to the population due to births, deaths, in-migration flows and out-migration flows. Age is a vital ingredient because fertility, mortality and migration components vary in intensity across the life span, in different ways. Sex is important because there are differences between men and women in their exposure to mortality and migration risks. Most fertility analyses use female populations at risk. It is always useful to distinguish between internal migration between areas within a country and international migration between countries. This is because these two migration flows have different drivers and constraints.

Demography has paid relatively little attention has been paid to the proper measurement and projection of migration. Academics have been good at building theories of migration such as the mobility transition (Zelinsky 1971), the cascading hierarchy of counter-urbanization migration flows (Champion 1998) or spatial interaction models of migration flows (Stillwell and Congdon 1991, Champion et al 2002). Very little of this has been incorporated into projection methodology.

The aim of this paper is to fill this gap in knowledge by reviewing the ways in which migration has been represented and modelled in population projections, assessing their advantages and disadvantages. We first outline the framework of population accounts within which migration must be embedded (section 2). We then consider the choice between representing migration as a flow of people or as a population at risk multiplied by a suitable migration rate (section 3). Section 4 reviews systems of interest in population projection, distinguishing migration flows internal to the system from those that are external. Section 5 shows how the different migration models can be used in single region, bi-regional, multi-regional and multi-country projection models. Section 6 illustrates the impact of migration model choice using a cohort-component model for the four countries of the UK. A final section discusses choice of a “best” migration model and what factors need to be considered.

## **2. A POPULATION ACCOUNTING FRAMEWORK**

Population accounts are frameworks for assembling the components of change within which migration data are located in a consistent way. Migration can be measured in censuses and surveys as *transitions* over (1) a fixed time interval (“Where did you live one year ago?”), (2) over an uncertain interval (“Where did you live before your present residence?”) or (3) over a person’s lifetime (“Where were you born?”). Migration can also be measured as (4) *moves* in registration systems (“What was your previous registered residence?”). Measures based on census or survey questions are subject to survivor bias and estimates must be made of non-surviving migrants. Registration systems capture the migration event as it occurs. It

is essential that type of migration measure and design of projection model are matched (Rees 1985). Only measures (1) and (4) are easy to use in projection models.

Migration statistics must be placed within the context of population change along with births and deaths. The essential feature of demographic accounts is that they must count all people or events that leave the population of interest and count all people or events that enter the population of interest. Here we focus only on demographic accounts based on moves, because this is the standard adopted by National Statistics (ONS 2012).

In Table 1 is set out a population accounting framework including migration data of the movement type. Although no projection will be as broad in scope as this framework, we later explain how aggregations or sub-sets underpin projection models in practice. The table rows list the regions in a country of interest and a set of countries in the rest of the world which send migrants to the regions. The set of regions and set of countries together cover the whole world. A final row holds births which add new-born infants to the regional and national populations. The regions and countries appear also in the columns of the table as destinations of the migrations and births. A final column contains the exits to death from the regions and countries.

Entries in the table are migration border-crossing moves and life state moves for births and deaths. An individual can experience only one birth or one death in a time interval but can make many migrations. This feature distinguishes movement accounts from transition, in which only one transition can occur in a time interval. We focus subsequent discussion on migration only. The migration part of Table 1 is divided into four quadrants holding different migration variables, about which decisions must be made in designing population projection models. The *top left quadrant* holds migration variables,  $M$ , that count migrations between  $n$  regions within the country of interest. There should be at least two regions – where there is only one region, the population accounts refer to a single country. In the UK projections are routinely carried out for around 400 lowest tier local authorities but separately in each home country<sup>4</sup>. The superscripts  $r1r2$  in variable  $M^{r1r2}$  indicate the migrations that take place between region  $r1$  and region  $r2$ . Counts of number of migrations (moves) between regions are found in the cells off the principal diagonal. The variables in the principal diagonal are residual counts derived from the accounting relationships for the table rows (see the Table 1 notes). The migration variables in the *top right quadrant* are labelled  $E$  and represent emigration, migration out of a country to other countries. The superscripts, e.g.  $r1c2$ , indicate region of origin and country of destination. The migration variables in the *bottom left quadrant* refer to immigration – migration into a region within the country of interest from another country, for example, from country  $m$  to region  $n$ ,  $I^{cmrn}$ . The *bottom right quadrant* completes the set of migration variables, holding counts of migration between origin and destination countries besides the country of

---

<sup>4</sup> The Home Countries are England, Wales, Scotland and Northern Ireland.

**Table 1: Population accounting framework incorporating migration (moves) data**

To	Country 1 of interest				Other countries				
From	Region r1	Region r2	...	Region r n	Country c2	...	Country cm	Deaths	Totals
Region r1	$R^{r1}$	$M^{r1r2}$	...	$M^{r1rn}$	$E^{r1c2}$	...	$E^{r1cm}$	$D^{r1}$	$P_t^{r1+}$
Region r2	$M^{r2r1}$	$R^{r2}$	...	$M^{r2rn}$	$E^{r2c2}$	...	$E^{r2cm}$	$D^{r2}$	$P_t^{r2+}$
:	:	:		:	:		:	:	:
Region rn	$M^{rn r1}$	$M^{rn r2}$	...	$R^{rn}$	$E^{rn c2}$	...	$E^{rn cm}$	$D^{rn}$	$P_t^{rn+}$
Country c2	$I^{c2r1}$	$I^{c2r2}$	...	$I^{c2rn}$	$R^{c2}$	...	$M^{c2cm}$	$D^{c2}$	$P_t^{c2+}$
:	:	:		:	:		:	:	:
Country cm	$I^{cmr1}$	$I^{cmr2}$	...	$I^{cmrn}$	$M^{cm c2}$	...	$R^{cm}$	$D^{cm}$	$P_t^{cm+}$
Births	$B^{r1}$	$B^{r2}$	...	$B^{rn}$	$B^{c2}$	...	$B^{cm}$	0	$B^+$
Totals	$P_{t+1}^{+r1}$	$P_{t+1}^{+r2}$	...	$P_{t+1}^{+rn}$	$P_{t+1}^{+c2}$	...	$P_{t+1}^{+cm}$	$D^+$	$T^{++}$

Notation

P	Population	r1, r2, ..., rn	subscripts for particular regions
R	Residual balance	c2, ..., cm	subscript for rest of country
M	Migration (moves)	n	number of regions
I	Immigration (moves)	m	number of countries
E	Emigration (moves)	+	summation over subscript replaced
B	Births	t	time at start of interval
D	Deaths	t+1	time at end of interval
T	Total flows	ri, rj, ck, cl	general subscripts for regions and countries

Notes

Accounting relationships for rows in Table 1 (example for region ri) are:
$P_t^{r1+} = R^{r1} + M^{r1r2} + \dots + M^{r1rn} + E^{r1c2} + \dots + E^{r1cm} + D^{r1}$ (1)
$R^{r1} = P_t^{r1+} - (M^{r1r2} + \dots + M^{r1rn}) - (E^{r1c2} + \dots + E^{r1cm}) - D^{r1}$ (2)
Accounting relationships for columns in Table 1 are:
$P_{t+1}^{+r1} = R^{r1} + M^{r2r1} + \dots + M^{rn r1} + I^{c2r1} + \dots + I^{cmr1} + B^{r1}$ (3)
With knowledge of opening and closing populations and natural increase, net migration can be derived as a residual:
$N^{ri} = (P_{t+1}^{+ri} - P_t^{ri+}) - (B^{ri} - D^{ri})$
This derivation involves summation of both internal and external flow balances:
$N^{ri} = \left( \sum_{rj \neq ri} M^{rjri} - \sum_{rj \neq ri} M^{rirj} \right) + \left( \sum_{ck} I^{ckri} - \sum_{ck} E^{rick} \right)$
The term "region" is used to indicate a territorial unit in general rather than a particular territorial unit such as "Standard Region", "Government Office Region" or "English Region".



interest (e.g.  $M^{c2cm}$ , migrations from country 2 to country  $m$ ). As in the top left quadrant, residual terms appear in the diagonal.

It is a requirement for constructing population accounts that the whole world is covered. However, it is usual in projection models to lump places outside the country of interest together as the “Rest of the World”. Although this aggregation is forced on the analyst by the poverty of international migration statistics, this is unsatisfactory as both policy makers and citizens have a considerable interest in the origins of immigrants and the destinations of emigrants.

The four quadrants of the accounts differ in terms of the source of the migration data. In the case of the UK, within country migrations are estimated from counts in the National Health Service (NHS) Patient Register(s), supplemented by information from the Higher Education Statistics Agency (HESA) on student entry to and exit from higher education institutions and locations. Some information is also used on migrations between the Armed Forces and prisoner populations and civilian locations. The immigration variables are estimates based on the national Long Term International Migration (LTIM) totals, which use data from the International Passenger Survey (IPS) (migrants intending to stay for 12 months or more), the Home Office (Asylum Seekers), HESA (foreign student immigration), the Labour Force Surveys of the UK and the Republic of Ireland. These national totals are distributed to local areas using data on new NHS registrations with a previous address abroad, new National Insurance registrations with a previous address outside the UK (Department of Work and Pensions), HESA (foreign student registrations) and the Ministry of Defence (migrations from overseas military bases to the UK (ONS 2011). Emigration variables are estimated using a time series model based on a set of determinants and proxies for emigration, such as immigration in the previous year and total internal out-migration, constrained to national total estimates based on the IPS (ONS 2010).

No reliable disaggregation of immigration or emigration by country of origin or destination can be made, beyond broad groupings (European Union, Australia-Canada-New Zealand, South Asia). Recent research combined IPS information with NHS, DWP and HESA administrative data using Bayesian statistical methods (Disney 2014). Immigration visa statistics could also be used. The Home Office has been working to develop a system for recording entry and exit via electronic passport records (Project Semaphore, the e-Borders project) though there have been difficulties which have delayed its implementation (Wikipedia 2014). The system, when developed, will record trips into and out of the UK rather than migrations (which make up about 1% of trips), so further information will be needed to produce comprehensive international migration statistics on origins and destinations for the UK.

Statistics on migration between countries are based on data gathered from national statistical agencies by international statistical agencies such as the United Nations, the World Bank and Eurostat. There are

considerable challenges in harmonizing these data, which have been tackled by academic teams for inter-country migration with Europe (see NIDI 2011 for the MIMOSA database of migration flows between European countries, see CPC 2013 for the similar IMEM database) and for all countries in the world by the Wittgenstein Centre (Abel 2013, Abel and Sander 2014, Sander, Abel and Bauer 2014a). The European country estimates rely heavily on population register data collected by a majority of European countries. It is necessary to resolve differences in the estimate of flows by origin and destination countries. The global migration flow estimates rely on use of time series of tables of lifetime migration from censuses and surveys. These are tables of country of residence against country of birth/citizenship. An innovative model deduces migration flows for five year intervals from successive censuses, supplemented with a gravity model to fill in missing migration flows.

### 3. MODELS OF MIGRATION FLOWS FOR POPULATION PROJECTIONS

The Table 1 framework is useful for organising historic statistics on populations and migrations but for future values of the migration flows we need models of migration. Table 2 sets out alternative approaches.

The *first approach* (Table 2, section 2.1) assumes that recent migration flows continue into the future. This is not a good model because it ignores variation in the populations generating the migrations. In addition it fails to consider changes in migration due to economic cycles or secular trends in migration rates. Rates in some developed countries may be in decline, controlling for population size and ageing.

However, if migration flows are subject to government control, then there is a need to project changes in migration. This is the basis of the *second approach* (Table 2, section 2.2) which assumes a trajectory of migration flows. This approach has been used, in particular, to model the migration flows in the emigration and immigration quadrants. Formal time series models may be employed with greater weight given to recent years and lesser weight to distant years. Usually, some plausible limit is set beyond which the first approach is employed. Judgement by statisticians or by a pool of experts is often used to set the trajectory of change in migration and its upper or lower limit. It is claimed, though rarely through assembly of evidence, that immigration policy limits immigration. To a frustrated intending immigrant or a failed asylum seeker it certainly feels like that. The 2010-15 Coalition Government policy of setting a target of less than 100 thousand for net immigration in the UK can be regarded as a failure, given that net international migration has averaged 208 thousand per annum for 2010-2013 (ONS 2014b). However, set against a projection model design which lets the growing population of the Rest of the World determine future migration, the policies of the Coalition Government might be judged successful.

**Table 2: Models of migration flows used in population projections**

2.1 Assume flows continue at historic levels (h =time horizon of projections)	
$M_{t=0}^{rirj} = M_{t=1}^{rirj} = \dots = M_{t=h}^{rirj}$ $E_{t=0}^{ricl} = E_{t=1}^{ricl} = \dots = E_{t=h}^{ricl}$ $I_{t=0}^{ckrj} = I_{t=1}^{ckrj} = \dots = I_{t=h}^{ckrj}$ $M_{t=0}^{ckcl} = M_{t=1}^{ckcl} = \dots = M_{t=h}^{ckcl}$	
2.2 Assume flows vary according to judgment or a time series model	
$M_{t=0}^{rirj}, M_{t=1}^{rirj}, \dots, M_{t=h}^{rirj}$ $E_{t=0}^{ricl}, E_{t=1}^{ricl}, \dots, E_{t=h}^{ricl}$ $I_{t=0}^{ckrj}, I_{t=1}^{ckrj}, \dots, I_{t=h}^{ckrj}$ $M_{t=0}^{ckcl}, M_{t=1}^{ckcl}, \dots, M_{t=h}^{ckcl}$	
2.3 Assume flows are rates multiplied by a population at risk	
Transmission rates, time constant or varying input to a projection model	
$tm^{rirj} = M^{rirj} / PAR^{ri}$ $te^{ricl} = E^{ricl} / PAR^{ri}$ $ti^{ckrj} = I^{ckrj} / PAR^{ck}$ $tc^{ckcl} = M^{ckcl} / PAR^{ck}$	$M^{rirj} = tm^{rirj} \times PAR^{ri}$ $E^{ricl} = te^{ricl} \times PAR^{ri}$ $I^{ckrj} = ti^{ckrj} \times PAR^{ck}$ $M^{ckcl} = tc^{ckcl} \times PAR^{ck}$
Admission rates, time constant or varying input to a projection model	
$am^{rirj} = M^{rirj} / PAR^{rj}$ $ae^{ricl} = E^{ricl} / PAR^{rj}$ $ai^{ckrj} = I^{ckrj} / PAR^{cl}$ $ac^{ckcl} = M^{ckcl} / PAR^{cl}$	$M^{rirj} = am^{rirj} \times PAR^{rj}$ $E^{ricl} = ae^{ricl} \times PAR^{rj}$ $I^{ckrj} = ai^{ckrj} \times PAR^{cl}$ $M^{ckcl} = ac^{ckcl} \times PAR^{cl}$
Populations at risk	
$PAR^{ri} = \frac{1}{2}(P_t^{ri} + P_{t+1}^{ri})$ $PAR^{ri} = \frac{1}{2}(P_t^{ri} + P_{t+1}^{ri})$ $PAR^{ck} = \frac{1}{2}(P_t^{ck} + P_{t+1}^{ck})$ $PAR^{ck} = \frac{1}{2}(P_t^{ck} + P_{t+1}^{ck})$	$PAR^{rj} = \frac{1}{2}(P_t^{rj} + P_{t+1}^{rj})$ $PAR^{rj} = \frac{1}{2}(P_t^{rj} + P_{t+1}^{rj})$ $PAR^{cl} = \frac{1}{2}(P_t^{cl} + P_{t+1}^{cl})$ $PAR^{cl} = \frac{1}{2}(P_t^{cl} + P_{t+1}^{cl})$
2.4 Assume an explanatory model for flows (e.g. a spatial interaction model)	
$M^{rirj} = f(w_O^{ri}, w_D^{rj}, c^{rirj}, K)$	
<p>where</p> <p><math>w_O^{ri}</math> is a vector of origin determinants of out-migration/emigration</p> <p><math>w_D^{rj}</math> is a vector of destination determinants of in-migration/immigration</p> <p><math>c^{rirj}</math> is matrix of variables estimating the “costs” of migration</p> <p><math>K'</math> is a set of constraints on the model predictions, which vary depending on the context the model is used in</p>	
2.5 Definitions	
<p>tm = transmission migration rate between regions</p> <p>te = transmission emigration rate from regions</p> <p>ti = transmission immigration rate to regions</p> <p>tc = transmission migration rate between countries</p> <p>am = admission migration rate between regions</p> <p>ae = admission emigration rate from regions</p> <p>ai = admission immigration rate to regions</p> <p>ac = admission migration rate between countries</p>	

The *third approach* (Table 2, section 2.3) is to adopt a model that projects migration as the product of migration rate, for a recent period or projected into the future, multiplied by a population at risk, generated by the projection model. Two kinds of rates can be used: the first uses as its denominator the population at risk of the origin region (within the country) or of the origin country in the rest of the world. This we call the *transmission* migration rate. The second kind of rate uses the population at risk of the destination region (within the country) or of the destination country in the rest of the world. This we call the *admission* migration rate. Use of transmission rates of migration implies that the capacity of the destination to absorb in-migration or immigration is not constrained, for example, by housing availability or by visa quotas. The use of admission rates implies there is some control of in-migration or immigration to the capacity of the receiving population. In particular situations, we can observe capacity constraints at work. In the UK Higher Education (HE) sector, quotas of places for home students at each university were set centrally for each university. So, a projection (Rees 1986) using transmission rates over-projected the number of UK students entering HE in the late 1980s and early 1990s on the basis of trends in the 1980s because the system was constrained at University destinations. No quotas were applied to students from outside the UK and between 2004 and 2013 study was the main reason for migration for 28% of immigrants (ONS 2014b). A recurring theme in much projection work is the need to understand migration streams by reason for migration versus the paucity of information about the numbers in each reason group by country. So a careful choice must be made between transmission and admission rates in designing models for projecting migration.

To compute migration rates requires a population at risk (PAR) denominator to match the migration numerator. The ideal PAR is the total person-years of risk in the region or country (per unit of time) but this needs to be sourced from longitudinal population registers which record residence spells in regions in the time interval. An approximation to this PAR measure is the average of start and end populations in the interval (Table 2, section 2.3). In the UK the historic series of rate measures are computed using mid-year populations, which over short intervals are close to the average population. How is the future average population in an interval computed, if the end population is unknown? The most general way is to set up a simple iterative algorithm. In the first iteration the start population is used as the first estimate of the PAR. An end population is then projected. An average PAR can now be computed and the end population is projected for a second time. The computations continue until the PAR converges on a stable value. Normally, this only takes a few iterations. The alternative is to work out an analytical version of the PAR so that the projection computations only need to be carried out once. The disadvantage is that a different analytical specification is needed depending on the migration model used.

In Table 2, section 2.3 are set out the migration rates used in projection in transmission (t) form (labelled  $t_m$ ,  $t_e$ ,  $t_i$  and  $t_c$ ) and in admission (a) form (labelled  $a_m$ ,  $a_e$ ,  $a_i$  and  $a_c$ ) where m refers to between region migration, e to emigration from regions, i to immigration to regions and c to between country migration.

To obtain projected migration flows the migration rates are multiplied by the matching PAR of origin region for transmission and of destination region for admission.

Table 2, section 2.4 proposes a final choice in the menu of migration models: models of migration in which flows are projected using the underpinning determinants, specified in a general way in the table. This choice has only rarely been used in a projection context for two reasons: first, the analyst needs to project the determinants into the future and this presents additional challenges; second, in many cases, a purely demographic model based on recent trends usually performs better than an explanatory model. However, such an explanatory model is needed if the aim is to explore the consequences for migration of regional and country employment, housing and social developments or plans. The elements that enter into such an explanatory model consist of a vector of origin attributes, a vector of destination attributes and a vector of factors which impede or expedite migration between origins and destinations. In some contexts it may be necessary constrain the migration model predictions to external forecasts or controls. A model of migration between former health authority areas developed for the Office of Deputy Prime Minister by a Newcastle-Leeds research team (Champion et al 2002) tests a very large set of determinants, using a two stage model. Stage one predicted the total out-migration from the origins; stage two distributed these projected origin totals across destination regions. The coefficients of both parts of the model are specific to origins, so this is an example of a model where the parameters vary by geographical areas.

#### 4. SYSTEMS OF INTEREST

The population accounts of Table 1 contain both migration flows within countries (domestic migration) and between countries (international migration). However, populations and flows of interest and those not of interest may not coincide with this domestic and international distinction. Figure 1 draws a distinction between internal flows between areas for which the populations are projected and external flows where the populations of source or destination areas for external flows are not projected. The examples in Figure 1 show how the distinction between internal and external flows can overlap international boundaries.

Figure 1.1 provides text labels for the algebraic variables in Table 1. Figure 1.2 shows a system of interest that is just a single region (or a single region repeated a number of times). Internal flows (in green) are confined to the region of interest (birth and death flows). External flows (in orange) are in-migration and immigration streams into the region and out-migration and emigration streams out of the region. Normally these are aggregated if the only migration information available consists of residual net migration. Completely outside the system of interest are flows between other areas (shown in red).

**Figure 1: Examples of systems of interest distinguishing internal and external migration flows**

1.1 Types of migration flows between regions and countries

GENERAL		DESTINATIONS						
MIGRATION FLOWS			Regions in a country	Sum for country	Countries in a cluster	Sum for cluster	Other clusters of countries	Sum for world
ORIGINS		1	2 ... n	+	2 ... m	+	2 ... c	+
	1							
Regions in a country	2 ... n		Flows between regions in a country		Flows to countries in a cluster		Flows to countries in other clusters	
Sum for country	+							
Countries in a cluster	2 ... m		Flows from countries in a cluster		Flows between countries in a cluster		Flows to other clusters to countries in a cluster	
Sum for cluster	+							
Other clusters of countries	2 ... c		Flows from countries in other clusters		Flows from other clusters to countries in a cluster		Flows between countries in other clusters	
Sum for world	+							

Key to Figures 1.2 to 1.5

	Internal flows in system of interest	
	External flows in system of interest	
	Outside system of interest	

1.2 Single region system

SINGLE REGION		DESTINATIONS						
MIGRATION FLOWS			Regions in a country	Sum for country	Countries in a cluster	Sum for cluster	Other clusters of countries	Sum for world
ORIGINS		1	2 ... n	+	2 ... m	+	2 ... c	+
	1							
Regions in a country	2 ... n							
Sum for country	+							
Countries in a cluster	2 ... m							
Sum for cluster	+							
Other clusters of countries	2 ... c							
Sum for world	+							

1.3 Regions used in a POPGROUP application

REGIONS IN POPGROUP		DESTINATIONS						
MIGRATION FLOWS			Regions in a country	Sum for country	Countries in a cluster	Sum for cluster	Other clusters of countries	Sum for world
ORIGINS		1	2 ... n	+	2 ... m	+	2 ... c	+
	1							
Regions in a country	2 ... n							
Sum for country	+							
Countries in a cluster	2 ... m							
Sum for cluster	+							
Other clusters of countries	2 ... c							
Sum for world	+							

1.4 Multi-region system within a whole country

MANY REGIONS		DESTINATIONS						
MIGRATION FLOWS			Regions in a country	Sum for country	Countries in a cluster	Sum for cluster	Other clusters of countries	Sum for world
ORIGINS		1	2 ... n	+	2 ... m	+	2 ... c	+
	1							
Regions in a country	2 ... n							
Sum for country	+							
Countries in a cluster	2 ... m							
Sum for cluster	+							
Other clusters of countries	2 ... c							
Sum for world	+							

1.5 Multi-region system in a cluster of countries

MANY REGIONS, COUNTRIES		DESTINATIONS						
MIGRATION FLOWS			Regions in a country	Sum for country	Countries in a cluster	Sum for cluster	Other clusters of countries	Sum for world
ORIGINS		1	2 ... n	+	2 ... m	+	2 ... c	+
	1							
Regions in a country	2 ... n							
Sum for country	+							
Countries in a cluster	2 ... m							
Sum for cluster	+							
Other clusters of countries	2 ... c							
Sum for world	+							

1.6 Multi-country system in all clusters of countries

MANY COUNTRIES		DESTINATIONS						
MIGRATION FLOWS			Regions in a country	Sum for country	Countries in a cluster	Sum for cluster	Other clusters of countries	Sum for world
ORIGINS		1	2 ... n	+	2 ... m	+	2 ... c	+
	1							
Regions in a country	2 ... n							
Sum for country	+							
Countries in a cluster	2 ... m							
Sum for cluster	+							
Other clusters of countries	2 ... c							
Sum for world	+							

Figure 1.3 shows a system of interest consisting of a set of regions within a country between which internal flows are of interest. However, the set of regions do not cover the whole national territory. This system of interest is often constructed by users of the POPGROUP software for population projection (LGA 2014), where the future population of a local authority of interest will be dominated by flows to and from neighbouring local authorities. The local authorities in the rest of the country can be treated as a single combination. The difficulty of such a system of interest is that it does not automatically generate the populations at risk for areas that strongly or weakly interact with the local authority of interest. Information is often borrowed from a bigger sub-national projection (e.g. ONS 2014a) to supply the populations at risk to be used in migration rate models. However, there will be issues of consistency between the local projections and the official sub-national projections to be resolved.

Figure 1.4 shows a system of interest which divides a country into a number of regions and models the flows between them simultaneously. External flows to or from these regions from other countries complete the system. The alternatives for modelling these external flows are discussed in the next section of the paper. Flows between other countries (highlighted in red) are ignored. The official sub-national projections for local authorities in England adopt such a multi-regional system (ONS 2014a).

Figure 1.5 extends the multi-region system at regional scale to countries which belong to a cluster with special arrangements for international migration, such as the freedom of labour migration which is a key foundation of the European Union. Such a system was used in the DEMIFER project (ESPON 2010, Rees et al 2012) that developed reference and policy scenario projections for European Union regions at NUTS2 level.

Figure 1.6 concentrates on country level populations, dropping the regional level, but for the whole world (Sander et al 2014b, Lutz et al 2014). All migration flows between countries are internal to the system of interest (highlighted in green) and regional flows are outside the system (highlighted in red).

## 5. PUTTING THE INGREDIENTS TOGETHER

In Table 1 we outlined the general accounting framework for population projection and in Table 2 we specified the various ways in which migration could be modelled/projected. In practice, projection models will only use part of the Table 1 framework which covers the whole world and may use different models to project migration in the different quadrants. Table 3 sets out a range of framework-model combinations that have been used recently. Figure 2 gives graphic representation to this range, as an aid to description. This menu, as in a restaurant, is not exhaustive but sets out a wide range of models, from which a projection analyst can choose. We set out the advantages and disadvantages of each migration model to help in this choice.

**Table 3: A classification of selected migration models used in projection**

<b>Classes</b>	<b>Territorial units</b>	<b>Representation of migration: Internal</b>	<b>Representation of migration: International</b>
<b>A</b>	<b>Uni-regional models</b>		
A1	One region	Net flows (combined)	Net flows (combined)
A2	One region	Net rates (combined)	Net rates (combined)
<b>B</b>	<b>Bi-regional models</b>		
B1	Two regions	Out-migration (transmission) rates	Immigration and emigration flows
B2	Two regions	Out-migration (transmission) rates	Immigration flows and emigration (transmission) rates
B3	Two regions	Out-migration (transmission) rates	Immigration (admission) rates and emigration (transmission) rates
B4	Two regions	Out-migration (transmission) rates	Emigration (transmission) rates for rest of world; emigration (transmission) rates for regions
<b>C</b>	<b>Multi-regional models</b>		
C1	Many regions	Out-migration (transmission) rates	Immigration flows and emigration flows
C2	Many regions	Out-migration (transmission) rates	Immigration flows and emigration (transmission) rates
C3	Many regions	Out-migration (transmission) rates	Immigration (admission) rates and emigration (transmission) rates
C4	Many regions	Out-migration (transmission) rates	Emigration (transmission) rates for rest of world; emigration (transmission) rates for regions
<b>D</b>	<b>Multi-country models</b>		
D1	Many countries	Not applicable	Net migration flows
D2	Many countries	Not applicable	Net migration rates
D3	Many countries	Not applicable	Inter-country migration (transmission) rates
D4	Many regions and countries	Out-migration (transmission) rates	Inter-country migration (transmission) rates; immigration and emigration flows



Figure 2: Graphical representation of the migration models

Net migration				Gross migration				Other terms		Abbreviations			
Net migration flows				Gross migration flows				Fertility and mortality rates		RoC Rest of Country			
Net migration admission rates				Gross migration admission rates				0 Not considered		RoW Migration			
Net migration transmission rates				Gross migration transmission rates						Rates Transmission migration rates			
										Arates Admission migration rates			
										Immig Immigration			
										Emig Emigration			
A1 Single Region, Net Mig Flows				A2 Single Region, Net Mig Rates									
From/To	Region			From/To	Region								
Region				Region									
RoW				RoW									
B1 Two Regions, Internal rates, Immig Flows, Emig Flows				B2 Two Regions, Internal Rates, Immig Rates, Emig Rates									
From/To	Region	RoC	RoW	From/To	Region	RoC	RoW						
Region				Region									
RoC				RoC									
RoW			0	RoW			0						
C1 Many Regions, Internal rates, Immig Flows, Emig Flows				C2 Many Regions, Internal Rates, Immig Rates, Emig Rates									
From/To	Region 1	Region 2	...	From/To	Region 1	Region 2	...						
Region 1			...	Region 1			...						
Region 2			...	Region 2			...						
...	...	...	...	...	...	...	...						
Region n			...	Region n			...						
RoW			...	RoW			0						
D1 Many Countries, Net Mig Flows				D2 Many Countries, Net Mig Rates									
From/To	Country 1	Country 2	...	From/To	Country 1	Country 2	...						
Country 1			...	Country 1			...						
RoW			...	RoW			...						

Table 3 is organised into four sets of models: *uni-regional models* (set A), in which only one region is of interest; *bi-regional models* (set B), consisting of a region and the rest of the country; *multi-regional models* (set C) in which all regional populations in a country are projected simultaneously; and *multi-country models* (set D) in which many countries are projected together.

### 5.1 Uni-regional models

Single region models (A1, A2) use net migration flows or rates in the absence of any better information on migration. These models are appropriate when net migration is derived as a residual (see the notes to Table 1). Net migration is the combination of sets of internal and external inflows and outflows, which could be trending in different directions. It is better to estimate historically and to project into the future the four gross migration flows rather than just one net migration figure. Model A1 assumes a trajectory of net migration flows, while model A2 projects the net migration flow as a net admission migration rate multiplied by the single region population. The trajectory of net migration flows or trajectory of net admission rates may be constant or time varying.

### 5.2 Bi-regional models

If the uni-regional model is rejected as inadequate or inconsistent, the next alternative is to model the population of the region of interest along with that of the rest of the country, because country level data on births, deaths and international migration are readily available. Internal migration can be modelled by multiplying transmission migration rates by origin region PARs, capturing the linkages between the regions. This method is used in the four types of bi-regional model identified in Table 3.

The B1 model uses trajectories of immigration and emigration flows. If this choice is made, it implies that both immigration and emigration are constrained at the destination, normally at country level. For example, immigration into the UK is controlled by national and international laws. Entry is managed through a visa system which differentiates potential immigrants by nationality (UK, EU and other) and by main reason for immigration (for work, for education, to join family members, for marriage, to seek asylum). Permitted lengths of residence are stipulated. Work permits favour occupations where there is a shortage of labour or which are important for international business. These controls serve to limit immigration and reflect past policy decisions. Policy is also framed in terms of targets but, in the absence of new laws, administrative measures are used such as delays in processing or raising the charges for visas.

The origin country has little influence on emigration, except in times of conflict when affected population groups may be forced to leave. Current concerns about British citizens emigrating to join Jihadi terrorist groups have led to some restrictions. More control on emigration is exerted by the destination countries but there are a large number of possible destinations which weaken the effect of controls in any one. This perspective leads to the second international migration model in which emigration flows are modelled as

the product of emigration transmission rates multiplied by the origin region or origin country population. In the ETHPOP population projections (Rees et al. 2011, ETHPOP 2014), this migration model produced increasing emigration flows for ethnic minority populations as these grew rapidly. With immigration flows fixed (to correspond with official net international migration assumptions in the 2008-Based National Population Projections), this meant shrinking net international migration and lower population growth in the B2 model than in the B1 model.

Which model should be preferred? The ETHPOP team gives different advice, recommending model B1, model B2 or an average on different occasions. An issue is the extent to which immigration raises emigration through return migration to a country of origin. ONS (2010) built in a dependence of emigration from England's local authorities on immigration in previous years into its model of sub-national emigration, specific to the main countries of origin. Considerable return migration to the West Indies has occurred at ages around retirement, but it is not yet clear whether this return migration will also occur for South Asian groups. For one immigrant group, students entering higher education, return migration is a condition of their visas, though many are able to gain leave to remain through government programmes such as the Fresh Talent initiative in Scotland (The Scottish Government 2008), which has been extended to the whole UK. The argument for the B1 model is that it follows National Statistics practice and projections, while model B2 has yet to be verified as better.

The third two region model, B3, uses immigration admission rates with emigration transmission rates. The argument for immigration admission rates is that, as a country's population increases, so its demand for immigrant labour and its capacity to teach international students expands. However, this argument is untested.

The final two-region model, B4, which treats immigration as a product of emigration rates from the Rest of the World multiplied by Rest of the World populations, is also untested. However, the model is inherently unlikely given the controls applied by destination countries. On the other hand, it is a model that indicates how immigration flows could potentially increase without controls and so provides a benchmark against which the impact of current policy can be measured.

### **5.3 Multi-regional models**

All of the multi-regional models (Table 3, type C; Figure 2, third row of tables) project migration flows for all regions within a country simultaneously, using internal out-migration rates and origin region populations. This methodology has been widely adopted since development between 1968 and 1986 by Rogers and colleagues (see Rogers 1995 for a summary of the methods). The four models differ in the way they handle international migration. The alternatives follow the B1 to B4 sequence of two-region models; the discussion and arguments in section 5.2 apply as well to multi-regional models, so they are

not repeated here. Multi-regional models are used by the Office for National Statistics for projecting local authority populations in England (ONS 2014a). The UK National Population Projections (NPP) up to the 2012-based projections consisted of four single region projections for each home country (England, Wales, Scotland, Northern Ireland) linked by a matrix of flows (ONS 2013). Bijak (2012), in a report on how the migration assumptions of the NPP might be improved, suggested that a multi-regional cohort-survival model with explicit inter-country migration rates be adopted. ONS (2014c) reports that this recommendation will be implemented in the 2014-based NPP, though empirical investigation led to the rejection of Bijak's recommendation of model C2 for handling international migration because it did not fit recent experience and did not generate plausible future international migration flows.

#### 5.4 Multi-country models

Population projections for sets of countries are undertaken by international agencies (e.g. UN 2014), national statistical agencies (e.g. US Census Bureau 2014), population research centres (e.g. Kupiszewski 2013; Lutz et al 2014) and multi-national research teams (ESPON 2010, Rees et al 2012). In Table 3 and Figure 2, four illustrative models are presented. D1 is used in UN 2014. D3 is used in the medium term in the world country projections (Lutz et al 2014) after 2050 the UN 2014 model of convergence of net migration flows to zero in all countries is used. Kupiszewski 2013 uses model D3 for European flows and model D1 for extra-Europe flows. Model D4 is used in ESPON (2010), which projects regional and national populations simultaneously for Europe using a model developed by Kupiszewski and Kupiszewska (2011), though extra-European migration is handled using model C3. This last example corresponds most closely with the general accounting framework of Table 1.

Model D1 is exemplified by the latest UN Population Division projections of the populations of UN Member States (UN 2014). Virtually all attention in the projection methodology is given to the fertility and mortality components with medium, high and low variants. Uncertainty is handled by developing probabilistic projections centred on the medium projection using Bayesian methods applied to the UN estimates time series. Virtually, no attention is given to international migration. Migration is not considered in framing the variants or probabilistic projections: one short-term trajectory of net migration balances is assumed for each country and balances are converged to zero for all countries in the long run, without much justification. The idea is probably that over time all countries will reach similar levels of development and the incentive to move from poorer countries to richer countries will disappear. The international migration literature offers little support for such a convergence hypothesis. In practice, international migration makes little difference to countries in the Global South (less developed states) where natural increase is the dominant component contributing to continuing though diminishing growth. In the Global North (more developed states) where natural increase is low or negative, net immigration makes a significant contribution. So there can be a considerable difference between the UN and national statistical agency projections for more developed countries.

Recently the UN and World Bank have assembled and published time series from 1960 to 2010 of country of birth and country of residence statistics. This time series of migrant stock data has been used by Wittgenstein Centre researchers (Abel and Sander 2014, Sander et al 2014a) to estimate migration flows over decadal intervals, interpolated for five year intervals. Using model D3, Sander et al (2014b) undertake an analysis of the forces affecting country to country flows, drawing on the literature, a meta-expert workshop and a web survey of general expert views. Future gross migrations were generated from a combination of (1) a time series statistical model, (2) meta-expert views about the forces affecting international migration and (3) ordinary expert views with relative weights of four ninths, four ninths and one ninth respectively. The resulting trends were allowed to run until 2050, when a model of convergence to eventual zero net migration was adopted (switching to a D1 model).

Model D4 (ESPON 2010, Kupiszewski and Kupiszewska 2012) combines several migration models to generate sets of population projections for the countries of the European Union in 2010, EFTA and Switzerland and for their NUTS 2 regions. Within each country, a multi-regional model is employed, using out-migration transmission rates. Within Europe country to country flows are modelled using a similar model. The projections are connected by sub-models allocating country flows to regions using destination regional populations. Immigration flows from outside Europe and emigration flows to countries outside Europe are handled by developing trajectories of migration numbers. Seven scenarios were developed – three reference and four policy scenarios. The policy scenarios combined assumptions about the rate of change in migration rates or flows linked to economic/environmental performance and assumptions about convergence or divergence between countries and regions within countries linked to the strengthening or weakening of welfare/social solidarity policies (Rees et al 2012b).

## 5.5 Explanatory models of migration

We briefly referred earlier in the report to the large body of work on this type of migration model, remarking that little of this work had been used for forecasting. However, there is a tradition in the forecasting of local authority populations in the UK of linking the projections to housing plans (Rees 1994). A model used widely by Local Authorities (LAs) in the 1990s and 2000s was the CHELMER Model, developed by the Population and Housing Research Group under the leadership of Professor Dave King, latterly at Anglia Ruskin University. Cambridge Econometrics took responsibility for the model in 2008 (Cambridge Econometrics 2014). The methodology is probably as follows<sup>5</sup>. A housing plan for an LA provides numbers and sizes of planned housing. This is converted using survey or census information into the likely number and size of households and their composition in terms of age and sex. These constitute people who will be added to the local population. They will be found by applying migration admission rates to the planned populations, including from within the same LA. Two problems

---

<sup>5</sup> Unfortunately, no recent documentation seems to be available.

are faced by the user of such a model: (1) it needs a forecast of housing units to be built, usually based on a judgement about how many housing permissions will result in housing completions and (2) those permissions are useful only in the short term (5 years) beyond which local strategic plans may guide the analyst.

## **6. ILLUSTRATIONS OF THE IMPACT OF MODEL CHOICES**

[This section is to be added. It will illustrate the impact of migration model choice using a cohort-component model for the four countries of the UK.]

## **7. DISCUSSION AND ADVICE**

We now attempt an answer the question posed in the introduction: “which migration models should be used in population projections?”

It is clear that there is no such thing as a “best” model. Rather it is a matter of “horses for courses”. That is, the model must be chosen in the light of driving forces of the migration system and of the data available. In particular, different models may be chosen to forecast different migration streams. Usually, a different model will be adopted for domestic migration from that for international migration. So for internal migration the most common choice is a model employing out-migration transmission rates (as in Rogers 1995). Applying such a model to the whole world assumes an absence of destination controls that is rather unlikely. A whole world model therefore indicates the potential for migration flows rather than future reality.

Raymer et al (2012) compare the forecasts of four simple regional population projection models: an overall growth rate model, a component model with net migration, a component model with in-migration and out-migration rates, and a multiregional model with destination-specific out-migration rates. The authors show how both the forecast subpopulation totals differ between the multiregional model and the simpler models, as well as for different assumptions about international migration. These experiments demonstrate that the ways in which migration is handled in projections really matter but offer little in the way of guidance.

We argue, following Rogers (1990), that net migration models hide too much. No net migrants exist. It is difficult to estimate age distributions for net migration. Net migration formulations may hide compositional effects of gross migration. The multi-regional model is therefore preferred.

Why should we therefore consider the bi-regional model? The bi-regional model requires far fewer variables which can be estimated more reliably. Rogers (1976) and Wilson and Bell (2004) both report that the results of a bi-regional model are close to those a full multiregional model when the model is

used for all pairs of regions and in-migrations are adjusted so that the total over all regions matches the total for out-migrations.

The experiments of Van Imhoff et al (1997) investigated the impact of adding age disaggregation to a multi-regional projection model. The conclusions were that knowledge of the all age flow matrix plus matrices of origins by age and destinations by age yielded, using iterative proportional fitting, provides good fits to full arrays. This is the “three faces of the migration cube” solution. It is not always ideal because origin-destination structures differ according to life course stages (Rees et al 1996). Van Imhoff et al (1997) focussed on different representations (aggregations) of the full multiregional model. Though useful, the work did not get to the heart of the matter: namely, which migration model is most accurate in producing forecasts. Here we need experiments that use a first part of a time series to calibrate the model and a following part to test the model, assessing the degree and sources of error. To compare the effect of different migration models on future migration requires use of the same overall projection model and assumptions about components in a systematic experimental design (cf Bongaarts and Bullateo 1999, Rees et al 2013).

We conclude with a summary of advice, gleaned from the work reviewed, on which migration model to use in a population projection.

- It is better to use gross migration flows rather than net.
- It is better to use gross migration rates than flows, as long as there are no constraints on the number of migrants who can be received.
- If there are constraints, then a model of future in-migration or immigration totals should be developed. Migration admission rates can be used to source migrants by origin.
- A trade-off needs to be considered between more detail (when input variables are harder to estimate) and less detail (easier to estimate).
- The bi-regional model has less detail than the multi-regional model but gives results that are close.
- The three faces version of the multi-regional model is a reasonable approximation to the full model.
- Different models may be needed for the different sets of flows (internal, international, or intermediate e.g. within Europe).
- The best choice for internal migration (migration transmission rate in a bi-regional or multi-regional model) may not be the best choice for international migration. A world model of inter-country migration may exaggerate migration to richer destinations which may be subject to controls.

We hope that the ideas presented in this report will help in the design of future population projection models incorporating migration.

## **Acknowledgements**

The conceptual analysis reported in this paper was commissioned by Edge Analytics Ltd, which has given permission for its inclusion in this paper. The work producing the illustrative ethnic projections was funded by the ESRC (Grant Ref ES/L013878/1, 2015-2016) as part of project *Evaluation, Revision and Extension of Ethnic Population Projections – NewETHPOP*. We also acknowledge the advice offered by Denise Williams of the Office for National Statistics, who provided access to ONS experimental work in preparation for the UK 2014-based National Population Projections.

## **Biographies**

Philip Rees is Emeritus Professor of Population Geography at the University of Leeds, with interests in ethnic population projections, health outcomes and ageing of the population.

Nik Lomax is a lecturer in Population and Migration. His research focuses on the demographic composition of local areas, which influences policy and resource allocation decisions. This work incorporates measurement and estimation of migration, births and deaths as well as assessment of how these patterns change over time.

Peter Boden is Managing Director of Edge Analytics Ltd, a consultancy that provides demographic analytics to clients in business and government, specializing in demographic forecasting, scenario planning, optimization, planning policy and project management. They maintain and develop the POPGROUP projection software for the Local Government Association.



## References

- Abel G (2013) Estimating global migration flow tables using place of birth data, *Demographic Research*, 28(18): 505-546
- Abel G and Sander N (2014) Quantifying global international migration flows, *Science*, 28 March, 2014, 343(6178): 1520-1522
- Bijak J (2012) Migration Assumptions in the UK National Population Projections: Methodology Review. Report to the Office for National Statistics. Southampton: University of Southampton, <http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/population-projections/npp-migration-assumptions-methodology-review/index.html>
- Bongaarts J and Bulatao R (1999) Completing the demographic transition. *Population and Development Review* 25, 515–29
- Cambridge Econometrics (2014) The Chelmer Population and Housing Model, [http://camecon.com/SubNational/SubNationalUK/ModellingCapability/Chelmer\\_Model.aspx](http://camecon.com/SubNational/SubNationalUK/ModellingCapability/Chelmer_Model.aspx)
- CEFMR (2014) *Central European Forum for Migration and Population Research*. Warsaw, Poland <http://www.cefmr.pan.pl/>
- Champion A (1998) Population trends of small and medium-sized towns in nonmetro regions / Les tendances démographiques des villes petites et moyennes en régions non métropolitaines. *Revue de Géographie de Lyon*, 73(1), 5-16
- Champion, T., Fotheringham, S., Rees, P., Bramley, G. and others (2002) *Development of a Migration Model*. The University of Newcastle upon Tyne, The University of Leeds and The Greater London Authority/London Research Centre. Office of the Deputy Prime Minister, London. [http://www.ncl.ac.uk/curds/publications/pdf/DevelopmentofamigrationmodelPDF2404Kb\\_id1153369.pdf](http://www.ncl.ac.uk/curds/publications/pdf/DevelopmentofamigrationmodelPDF2404Kb_id1153369.pdf)
- CPC (2013) Integrated Modelling of European Migration (IMEM) Database. Centre for Population Change, University of Southampton, <http://www.imem.cpc.ac.uk/>
- Disney NG (2014) *Model Based Estimation of UK Immigration*. PhD Thesis, Social Statistics, University of Southampton
- ESPON (2010) DEMIFER - *Demographic and Migratory Flows Affecting European Regions and Cities*. [http://www.espon.eu/main/Menu\\_Projects/Menu\\_AppliedResearch/demifer.html](http://www.espon.eu/main/Menu_Projects/Menu_AppliedResearch/demifer.html)
- ETHPOP (2014) ETHPOP database, publications and presentations. [www.ethpop.org](http://www.ethpop.org)
- GLA (2014) *GLA 2013 Round Population and Household Projections*. Greater London Authority, <http://data.london.gov.uk/dataset/gla-2013-round-population-and-household-projections>
- Kupiszewski M (ed.) (2013) *International Migration and the Future of Populations and Labour in Europe*. The Springer Series on Demographic Methods and Population Analysis Vol. 32, Springer
- Kupiszewski M and Kupiszewska D (2011) MULTIPOLES: A Revised Multiregional Model for Improved Capture of International Migration. In J Stillwell and M Clarke (ed.) *Population Dynamics and Projection Methods*, Springer, pp.41-60
- LGA (2014) POPGROUP. Local Government Association, <http://www.local.gov.uk/popgroup>
- Lutz W, Butz W and KC S (ed.) (2014) *World Population and Human Capital in the Twenty-First Century*. Oxford, Oxford University Press
- NIDI (2011) Project: MIMOSA: Modelling migration and migrant populations. Netherlands Interdisciplinary Demographic Institute, <http://www.nidi.knaw.nl/en/research/mm/230211>
- ONS (2010) *Estimating International Long-term Emigration by Local Authority: Update*. Office for National Statistics, <http://www.ons.gov.uk/ons/guide-method/methodquality/imps/msiprogramme/communication/improvements-mid-2008/methodology-papers/emigrationdetailedmethodology-update.pdf>
- ONS (2011). *Improved Methodology for Estimating Immigration to Local Authorities in England and Wales*. Web page describing a suite of methodological papers. Office for National Statistics, <http://www.ons.gov.uk/ons/guide-method/method-quality/imps/improvements-to-local-authorityimmigration-estimates/index.html>
- ONS (2012) *A Conceptual Framework for UK Population and Migration Statistics*. Office for National Statistics, <http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/conceptual-framework/index.html>
- ONS (2013) *National Population Projections, 2012-based: Statistical Bulletin*. Office for National Statistics, [http://www.ons.gov.uk/ons/dcp171778\\_334975.pdf](http://www.ons.gov.uk/ons/dcp171778_334975.pdf)

- ONS (2014a) *2012-based Subnational Population Projections for England: Statistical Bulletin*. Office for National Statistics, [http://www.ons.gov.uk/ons/dcp171778\\_363912.pdf](http://www.ons.gov.uk/ons/dcp171778_363912.pdf)
- ONS (2014b) *Migration Statistics Quarterly Report, November 2014*. Office for National Statistics, [http://www.ons.gov.uk/ons/dcp171778\\_386531.pdf](http://www.ons.gov.uk/ons/dcp171778_386531.pdf)
- ONS (2014c) *An Examination of Options for Setting International Migration Assumptions for the National Population Projections*. NPP Migration assumptions rates-based modelling project 14(02), Office for National Statistics
- Raymer J, Abel G, Rogers A (2012) Does specification matter? Experiments with simple multiregional probabilistic population projections. *Environment and Planning A* 44(11) 2664–2686
- Rees P (1985) Does it really matter which migration data you use in a population model? In White P and Van der Knaap G (eds.) *Contemporary studies of migration*. Geobooks, Norwich, pp.55-77
- Rees P (1986) A geographical forecast of the demand for student places. *Transactions, Institute of British Geographers*, NS11, 5-26
- Rees P (1994) Estimating and projecting the populations of urban communities. *Environment and Planning A*, 26, 1671-1697
- Rees P, Durham H, Kupiszewski M (1996) Internal migration and regional population dynamics in Europe: United Kingdom case study. *Working Paper 96/20*, School of Geography, University of Leeds, Leeds, UK, <http://eprints.whiterose.ac.uk/3734/>
- Rees P, Wohland P, Norman P and Boden P (2011) A local analysis of ethnic group population trends and projections for the UK. *Journal of Population Research*, 28(2-3), 149-184
- Rees P, van der Gaag N, de Beer J, Heins F (2012) European regional populations: current trends, future pathways and policy options. *European Journal of Population*, 28(4), 385-416
- Rees P, Wohland P and Norman P (2013) The demographic drivers of future ethnic group populations for UK local areas 2001–2051. *The Geographical Journal*, 179(1), 44–60
- Rogers A (1976) Shrinking large-scale population-projection models by aggregation and decomposition. *Environment and Planning A*, 8, 515-541
- Rogers A (1990) Requiem for the net migrant. *Geographical Analysis*, 22(4), 283-300
- Rogers A (1995) *Multiregional Demography: Principles, Methods and Extensions*. Chichester, John Wiley
- Sander N, Abel G and Bauer R (2014a) The global flow of people, the Wittgenstein Centre for Demography and Global Human Capital, Vienna, <http://www.global-migration.info/>
- Sander N, Abel G and Riosmena F (2014b) The future of international migration. Chapter 7, pp.333-396 in Lutz W, Butz W and KC S (ed.) (2014) *World Population and Human Capital in the Twenty-First Century*. Oxford, Oxford University Press
- Stillwell J and Congdon P (ed.) (1991) *Migration Models: Macro and Micro Approaches*. London, Belhaven Press
- The Scottish Government (2008) *Fresh Talent: Working in Scotland Scheme: An Evidence Review*. Authored by Cavanagh L, Eirich F and McLaren J-G, Scottish Government Social Research, <http://www.scotland.gov.uk/resource/doc/235857/0064664.pdf>
- UN (2014) *World Population Prospects – The 2012 Revision: Methodology of the United Nations Population Estimates and Projections*. Department of Economic and Social Affairs, Population Division, ESA/P/WP.235, New York, United Nations, <http://esa.un.org/wpp/documentation/publications.htm>
- US Census Bureau (2014) *International Data Base – World Population: 1950-2050*. <http://www.census.gov/population/international/data/idb/worldpopgraph.php>
- van Imhoff E, van der Gaag N, van Wissen L and Rees P (1997) The selection of internal migration models for European regions. *International Journal of Population Geography*, 3, 137-159
- Wikipedia (2014) *e-Borders*. <http://en.wikipedia.org/wiki/E-Borders>
- Wilson T and Bell M (2004) Comparative empirical evaluations of internal migration models in subnational population projections. *Journal of Population Research*, 21(2), 127-159
- World Bank (2014) *Population Estimates and Projections*. <http://data.worldbank.org/data-catalog/population-projection-tables>
- Zelinsky W (1971) The hypothesis of the mobility transition. *Geographical Review*, 61(2), 219-249

# Learning Lessons from Population Projections: How Well Did We Forecast the Ethnic Transition?

Philip Rees<sup>1\*</sup>, Pia Wohland<sup>2†</sup> and Paul Norman<sup>1‡</sup>

<sup>1</sup>Centre for Spatial Analysis and Policy, School of Geography, University of Leeds, Leeds

<sup>2</sup>Institute for Health and Society (IHS) & Newcastle University Institute for Ageing (NUIA)

January 2015

## Summary

Population projections are rarely evaluated. Yet evaluations are an essential means for understanding how projections depart from reality. In this paper we describe a set of local population projections by ethnicity for England, based on the 2001 Census population, which can be compared after 10 years with the results of the 2011 Census. We examine the differences at national and local levels and for broad and detailed aggregate groups. Some differences are small such as the all group population of England but others are large such as the populations of Other Asians or Other Blacks, indicating considerable uncertainty about the components of change estimated for 2001 to 2006 and the projection assumptions for 2006 to 2011. By looking at the ethnic-local differences by age and by geographic pattern we can make some deductions about the sources of difference. These lessons enable us to plan a set of simulations with a new version of our projection model over 2001 to 2011 that can isolate more precisely the reasons for the 2011 projection versus 2011 census differences. This analysis will help improve the next round of ethnic population projections upon which we are currently engaged.

**KEYWORDS:** Population Projections, Ethnic Projections, 2011 Census, Evaluation

## 1. The Context for Ethnic Population Projections

Most European countries have entered what Coleman (2006) has termed the Third Demographic Transition, in which a combination of low fertility and increasing longevity results in an ageing population and labour supply problems, which stimulate inward migration. Because of this migration of foreigners the ethnic (national) origin of the population changes as the immigrants settle, form families and grow because of their reproductive-friendly age structure. A transition ensues from an ethnically homogeneous population to an ethnically heterogeneous one. The process of compositional change does not stop with the immigration of the foreign born. When the new immigrants settle and raise families, a new generation of native-born children of foreign born parents will be created. Later this new generation will marry or partner and have children who will be the grandchildren of the foreign born immigrants, though, of course, many family histories will be more complex than this. In most European countries there is no official tracking of the country of birth ancestry of these new generations. Further generations are assigned to the native born part of the population. However, communities with such foreign birth origins continue to retain cultural and spatial characteristics associated with their origins, while at the same time assimilating into or integrating with the native born population. In the UK it was recognised in the 1970s that there was a need to monitor groups of distinct ethnic origin because of obvious

---

\* p.h.rees@leeds.ac.uk

† pia.wohland@newcastle.ac.uk

‡ p.d.norman@leeds.ac.uk

discrimination in employment and housing markets. An ethnic classification was introduced in the 1980s into the UK Labour Force Survey, using a self-reporting question. After an attempt to introduce an ethnic question into the 1981 Census failed because of some ethnic minority anxiety, a question was used in the 1991 Census, with positive ethnic minority support, and (with revisions) has been used in the 2001 and 2011 Censuses. In the UK we can therefore examine the progress of the ethnic transition from 1991 onwards and use this knowledge to forecast the country's ethnic composition at national and subnational levels.

It is important to be able to monitor ethnic group numbers in order to measure the degree of (dis-) advantage that groups face and discrimination on grounds of ethnicity after controlling for other factors that determine achievement or well-being. Of course, other dimensions of difference such as disability or sexuality or social class also produce disadvantage and discrimination, but the projection of the population by these dimensions is less developed.

For what purposes is ethnic population information needed? The health needs of groups different by ethnicity that is related to genetics (sickle cell anaemia), food consumption (in relation to cardio-vascular disease) though socio-economic deprivation plays the most important role in determining health status. Businesses also benefit from knowledge of the ethnic composition of their potential consumers where tastes differ in food or clothing choices. Politicians need to know about the future ethnic compositions of constituency electorates because voting preferences vary by ethnic group. In 2014 the think tank Policy Exchange commissioned a re-working of previously published ethnic projections for Parliamentary constituencies (Rees and Clark 2014). Another important use for ethnic population outputs is to provide context variables for individual level studies; a framework which recognizes the need to measure the impact of spatial community characteristics on behaviour or conditions in addition to individual level variables. Information on future ethnic composition of local populations is needed for planning and consultation purposes, as is testified by their production for a number of years by the Greater London Authority (GLA 2014). Ethnic group projections also have a role to play in informing public debate about the way national and subnational populations are changing.

If ethnic population projections are to fulfil these many roles, they must of the highest quality. The aim of this chapter is to pause between a completed round of such projections based on the 2001 Census (Wohland et al 2010, Wohland et al 2014, Rees et al. 2011, 2012a, 2012b) and a new round of projections based on the 2011 Census (Rees et al. 2015) to evaluate how well the 2001 based projections reproduced the 2011 Census results and to interpret the differences.

Section 2 provides some brief background on 2001-2011 changes in the ethnic composition of the population. In section 3, we summarise the way we implemented the projections and discuss key results.

In section 4, we compare our projections with the results of the 2011 Census and suggest reasons for the differences observed. Section 5 discusses how fast the ethnic transition in the population of England and Wales is happening.

## 2. THE CHANGING ETHNICITY OF THE POPULATION

Our projections cover the whole of the UK, but we focus our evaluation on England and Wales, which use the same classifications of ethnicity. The *White* share of the England and Wales population has declined from 93% in the 1991 Census to 91% in the 2001 Census and 85% in 2011 Census (Jivraj 2012, from Census statistics). The *White British* share of the 2011 Census population was 80% of the 85. So the (not White) ethnic minority share of the England and Wales population increased from 7% in 1991 to 9% in 2001 and 15% in 2011. These statistics indicate some speeding up of the transition during the 2001-2011 period compared with 1991-2001.

By ethnic group we mean people who identify themselves as belonging to a sub-population with national origins outside the United Kingdom; both those born abroad and their descendants. Ethnicity is based on self-identification when answering a survey or census question. Ethnic status has legal recognition in the Equality Act 2010 (EHRC 2013) which replaces earlier acts with a unified framework for monitoring and acting on evidence of disadvantage or discrimination on grounds of gender, ethnicity, disability or sexual orientation. It is important to compare the numbers of ethnic group members in work compared with the numbers available for work, for example.

The decennial census has been the main source of information on local and national ethnic populations but society can benefit from more frequent knowledge of changing ethnic composition. Updates can be provided through two methods. The first method is to roll forward ethnic group populations from the latest census to successive mid-years, using estimates of the components of population change, which has been done for mid-year 2002 to 2009 for local authorities by the Office for National Statistics (ONS 2011b). The second method is to carry out a representative social survey each year, such as ONS's Annual Population Survey (APS) (ONS 2013b). ONS is currently carrying out an assessment of the reliability of their ethnic population estimates (ONS 2011c) and has suspended the production of local estimates, pending an evaluation against the results of the 2011 Census (ONS 2011d).

We need a view about how the ethnic composition of the population of England and Wales is likely to change in the future. As ONS is doing, we have embarked on an evaluation of our projections through comparison with the results of the 2011 Census. The main research question we try to answer is "How well did we do in projecting ethnic group populations for local authorities in England?"

### 3. METHODS, DATA AND RESULTS OF THE ETHNIC PROJECTIONS

The projection model used is a bi-regional cohort component model (Rees et al. 2012). This means we project the population disaggregated by age and sex, accounting for migration within the United Kingdom as outflows from each area and as inflows from the rest of the country. The model is applied separately to each ethnic group using suitable estimates of rates and flows. However, the groups are connected when babies are born to mothers and fathers from different groups and are assigned by their parents in mixed ethnic categories. We use single years of age to 100+ and the sixteen ethnic groups in the 2001 Census. The coverage of the projections is the United Kingdom. For spatial units, we use 352 local authorities<sup>§</sup> in England together with Wales, Scotland and Northern Ireland (355 zones in total).

Selecting from a wide set of projections, we focus here on two scenarios (Table 1) which use assumptions aligned to those in the National Population Projections (2008-based) but which differ in one model feature. The first scenario (TREND) projects the future flows of emigrants along with an equivalent immigration series, which together match the ONS net international migration assumption of +180 thousand net international migrants. Note that since 2008 the net immigration level has been mainly higher than the 180 thousand net immigration assumption (ONS 2013b). The second scenario (UPTAPER) projects emigration as a product of emigration rates and the changing populations at risk in the 355 zones. The argument for the second approach is that we should make the emigration flows functions of the population at risk resident in the UK zones, which will change over the projection. The argument against such an approach is that emigrants face many barriers to migration, some determined by destination country policy just as immigration to the UK is influenced by UK government policy.

**Table 1** ETHPOP scenario projections: assumptions

Scenario Projection	Assumptions
TREND	Fertility, mortality and international migration assumptions follow those of ONS's 2008-based National Population Projections 2008 factored to reflect local authority differences. Internal migration assumptions are based on the 2001 Census updated using NHS Patient Register data to 2008
UPTAPER	Uses the same assumptions as the TREND projection by changes the model for projecting emigration from assumptions about emigration flows to assumptions about emigration rates which are multiplied by local authority populations

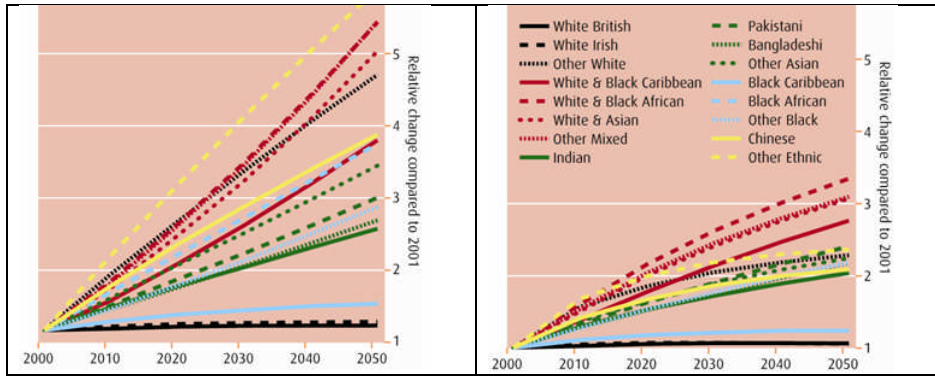
**Sources:** For details see Wohland et al. (2010) and Rees et al. (2012a)

Figure 1 shows the results of these two projections for the UK starting from a base population at mid-year 2001. The graphs show the projected changes for the 16 ethnic groups relative to the population at mid-year 2001. The White British and White Irish populations fail to grow in the UPTAPER projection and increase only a little in the TREND projection, falling back toward mid-century. The Black Caribbean population grows quite slowly, experiencing high emigration, low fertility and a loss of children to the

<sup>§</sup> Two LAs with very small populations are merged with a larger neighbour: City of London with City of Westminster and Isles of Scilly with Penwith.

mixed White and Black Caribbean group. The other minority ethnic populations have much higher future trajectories, experiencing growth of 2 to 3.2 times in the UPTAPER projections between 2001 and 2051 and 2.5 to 6 times in the TREND projections. These projections show that the diversity of the UK population will increase substantially over the first half of this century.

**Figure 1** Projected populations of ethnic groups under the emigration flows scenario (TREND in the LH Graph) and the emigration rates scenario (UPTAPER in the RH graph)



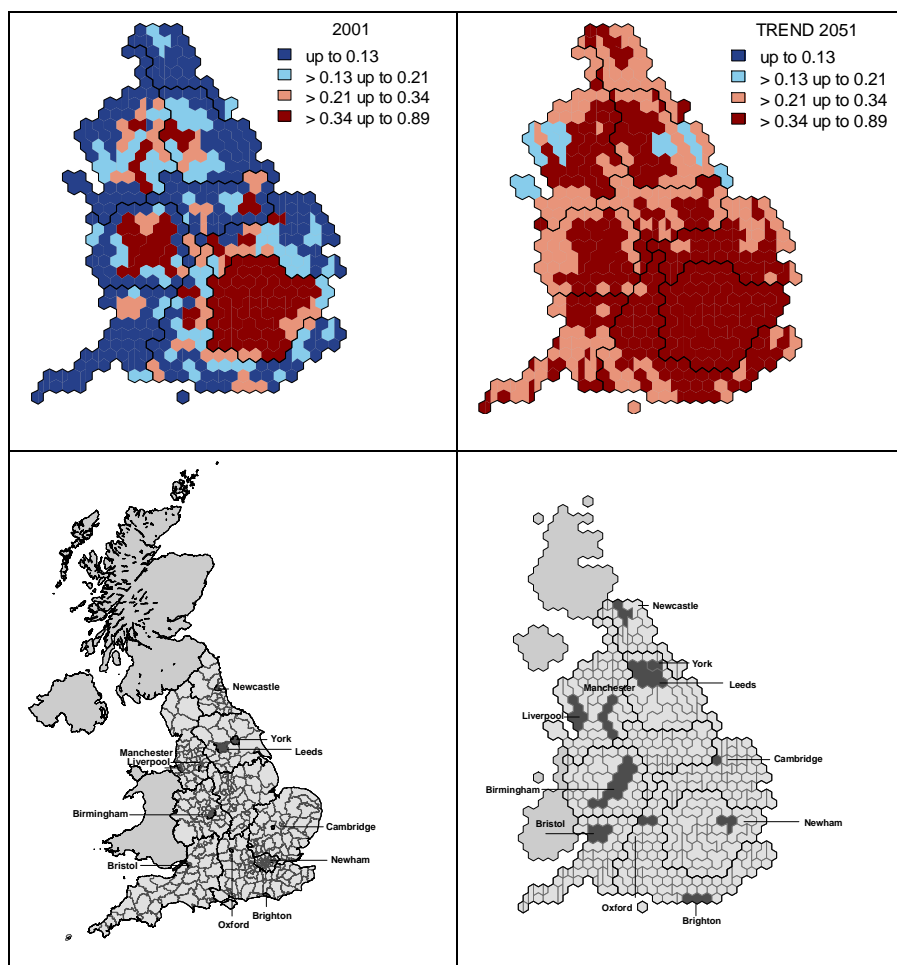
Source: Rees et al. (2011).

Figure 2 shows diversity patterns for local authorities in England, measured using the Simpson Index of Diversity (DI), mapped using a population cartogram, in which the extent on the map of a zone is proportional to a population at risk (2001 population). The minimum DI is zero (no diversity) when only one group is present in an area. The maximum diversity depends on the number of groups considered. With 16 groups each having an equal share of the population, the DI would be 0.9375. The ethnic diversity of the local authority populations increases substantially over the 50 years. By 2051, under the TREND projection, almost all areas are in the top two diversity quartiles, which were found in London only and some big cities in 2001. In the UPTAPER projection the increase in diversity is more subdued.

A note of caution is necessary before we accept this picture of the future ethnic composition of the UK population. The projections are liable to error from three sources: (1) the estimates of the ethnic fertility, mortality, internal and international migration indicators used as projection inputs may be wrong, (2) the assumptions about the future behaviour of the inputs may be wrong, and (3) the model may be wrongly specified. So, it is sensible to validate the projections against the recent 2011 Census. We report on this evaluation in the next section.

**Figure 2:** Ethnic diversity for LAs in England: TREND projection, 2001 and 2051

Note: Index of Diversity =  $1 - \sum_e r_e^2$ , where  $r_e$  = proportion of the population in ethnic group  $e$ .



Source: Wohland et al. (2010).

#### 4. COMPARISON OF CENSUS 2011 AND ETHPOP PROJECTED POPULATIONS

To compare the 2011 Census results and the ETHPOP projections, we harmonize the ethnic group definitions and the boundaries of the local authorities. Table 2 shows the correspondence of the 16 ethnic groups in the 2001 Census and the 18 ethnic groups in the 2011 Census in England and Wales. We merge the two new groups in 2011 into the matching larger group in 2001. To harmonize the zones used in the two censuses we aggregated the 2001 Census local authorities into unitary county authorities created in April 2009 and used in the 2011 Census. In 2001 England and Wales was made up of 352 zones; in 2011 there were 327 zones. The final adjustment made was to combine the 2011 TREND and UPTAPER projected populations because these figures bracketed the 2011 Census population for England and Wales as a whole and because the arguments for and against each approach are unresolved. We did not correct for the 3 month time difference between the Census (27 March) and the projections (30 June/1 July).



**Table 2:** The correspondence of ethnic groups in the 2001 and 2011 Censuses

2001 Census Ethnic Group (16 Groups)	2011 Census Ethnic Group (18 Groups)
White: British	White: English/Welsh/Scottish/Northern Irish/British
White: Irish	White: Irish
White: Other white	White: Gypsy or Irish Traveller
	White: Other White
Mixed: White and Black Caribbean	Mixed/multiple ethnic group: White and Black Caribbean
Mixed: White and Black African	Mixed/multiple ethnic group: White and Black African
Mixed: White and Asian	Mixed/multiple ethnic group: White and Asian
Mixed: Other Mixed	Mixed/multiple ethnic group: Other Mixed
Asian or Asian British: Indian	Asian/Asian British: Indian
Asian or Asian British: Pakistani	Asian/Asian British: Pakistani
Asian or Asian British: Bangladeshi	Asian/Asian British: Bangladeshi
Chinese or Other Ethnic Group: Chinese	Asian/Asian British: Chinese
Asian or Asian British: Other Asian	Asian/Asian British: Other Asian
Black or Black British: Black African	Black/African/Caribbean/Black British: African
Black or Black British: Black Caribbean	Black/African/Caribbean/Black British: Caribbean
Black or Black British: Other Black	Black/African/Caribbean/Black British: Other Black
Chinese or Other Ethnic Group: Other Ethnic	Other ethnic group: Arab
	Other ethnic group: Any other ethnic group

Note: Darker lines show how the 16/18 detailed groups are combined into 5 broad groupings.

In Table 3 we compare results for England and Wales from the 2011 Census with the average of our two ETHPOP projections for five broad ethnic groupings. Our projection of the All Groups population is very close (a difference of less than 3 in 10,000). Our projection of the White groups is close (3% error) but the projected population is greater than the census population. This means that we under-projected the growth of the ethnic minority groups. Our projection of the Asian groups was 18% under the 2011 Census figure. For the Black groups the under-projection was also 18%. The under-projection of the Mixed groups was 17%. The projection of the “Other” groups was the worst at 23%. Overall we have under-projected the growth of ethnic minority populations and hence the pace at which England and Wales population has diversified. Instead of the White/Ethnic Minority population split being 88.5%/11.5% in 2011 (our projections), the Census measured the split as 86.0%/14.0% .

Table 4 shows comparisons for the 16 harmonized groups. Within the White groups, the White British group is over-projected by 3% and the White Irish group by 25%. We may have under-estimated the emigration of White British people and over-estimated their survival into old age. For the White Irish group we may have under-estimated the extent that their children born in the 2000s were assigned White British ethnicity and some of the group may have changed identities between censuses. The White Other group we under-projected by 7%, probably because we under-estimated the level of immigration from other European countries and over-estimated emigration.

**Table 3:** The England and Wales populations of five ethnic groupings: Census 2011 and ETHPOP 2011

Ethnic grouping (2011 definitions)	Census Population CD 2011  (thousands)	Average of TREND and UPTAPER Projections, MY2011 (thousands)	Difference = Average Projection minus Census (thousands)	100 × (Difference /Census) (%)
All Groups	56,076	56,057	-18	-0.03
White	48,209	49,609	1,400	2.90
Black	1,865	1,521	-344	-18.44
Asian	4,214	3,469	-744	-17.66
Mixed	1,224	1,017	-207	-16.90
Other	564	440	-124	-21.92

Sources: Census 2011 – ONS (2013c), Crown Copyright.

Projections – ETHPOP (2013), funded by ESRC.

Notes: CD = Census date (27 March 2011), MY = Mid-YEAR (30 June/1 July).

**Table 4** The England and Wales populations of 16 harmonized ethnic groups: Census 2011 and ETHPOP 2011

16 Harmonized Ethnic Groups	Census CD2011 Populations  (thousands)	Average of TREND and UPTAPER Projections MY2011 (thousands)	Difference = Average minus Census (thousands)	100 × (Difference/ Census) (%)	Difference Multiplier = (Census Change /ETHPOP Change) 2001-2011
All Groups	56,076	56,057	-18	0.0	1.00
White British	45,135	46,572	1,437	3.2	0.97
White Irish	531	663	132	24.8	0.80
White Other	2,544	2,374	-169	-6.7	1.07
White and Black Caribbean	427	331	-96	-22.5	1.29
White and Black African	166	132	-34	-20.2	1.25
White and Asian	342	299	-43	-12.6	1.14
Other Mixed	290	256	-34	-11.9	1.13
Indian	1,413	1,394	-19	-1.3	1.01
Pakistani	1,125	990	-134	-11.9	1.14
Bangladeshi	447	369	-78	-17.4	1.21
Chinese	393	361	-33	-8.3	1.09
Other Asian	836	355	-481	-57.6	2.36
Black Caribbean	595	641	46	7.8	0.93
Black African	990	754	-236	-23.8	1.31
Other Black	280	126	-154	-55.0	2.22
Other Ethnic Group	564	440	-124	-21.9	1.28

Source and Notes: See Table3.

Within the Mixed groups, under-projections of the White and Black Caribbean and White and Black African groups were twice as large as the under-projection of White and Asian and Other Mixed groups. These groups grow because children are born to parents of two different ethnicities. The only

information available on this process came from a 2001 Census commissioned table. Mixed partnerships may have increased from the 2001 level because more opportunities for mixing became available.

Within the Asian groups, the Indian group were only slightly under-projected (by 1%), but the Pakistani and Bangladeshi groups were under-projected by 12 and 17% respectively. It is likely we under-estimated the strength of continuing immigration associated with marriage for these groups and over-estimated the falls in their fertility. The Chinese group was under-projected by 8%; there was a growing influx of students from China after 2001, which we may have under-estimated. The Other Asian group, with varied origins in the smaller countries in Asia, was the most under-projected of all the groups at 58%. We likely under-estimated the inflows from crisis countries such as Iraq, Afghanistan, Iran and the student intake from emerging countries in South-East and East Asia (Malaysia, Thailand, Vietnam, South Korea).

The Black Caribbean group was over-projected by 8%. We may have under-estimated the return emigration stream to the West Indies. This is an important process for the older members who arrived in England and Wales in the 1950s and who entered the retirement ages in the 2000s. We under-projected the Black African group by 24%, probably under-estimating immigration. Immigration from sub-Saharan Africa is composed of increasing student numbers, (e.g. from Nigeria and Ghana), flows of refugees and asylum seekers (e.g. from Somalia, Eritrea and Zimbabwe). The Other Black group is under-projected by 55%. People in other Black groups may have changed their identity between censuses and new groups may have started immigration in the 2000s from Latin America and the South West Pacific, for example.

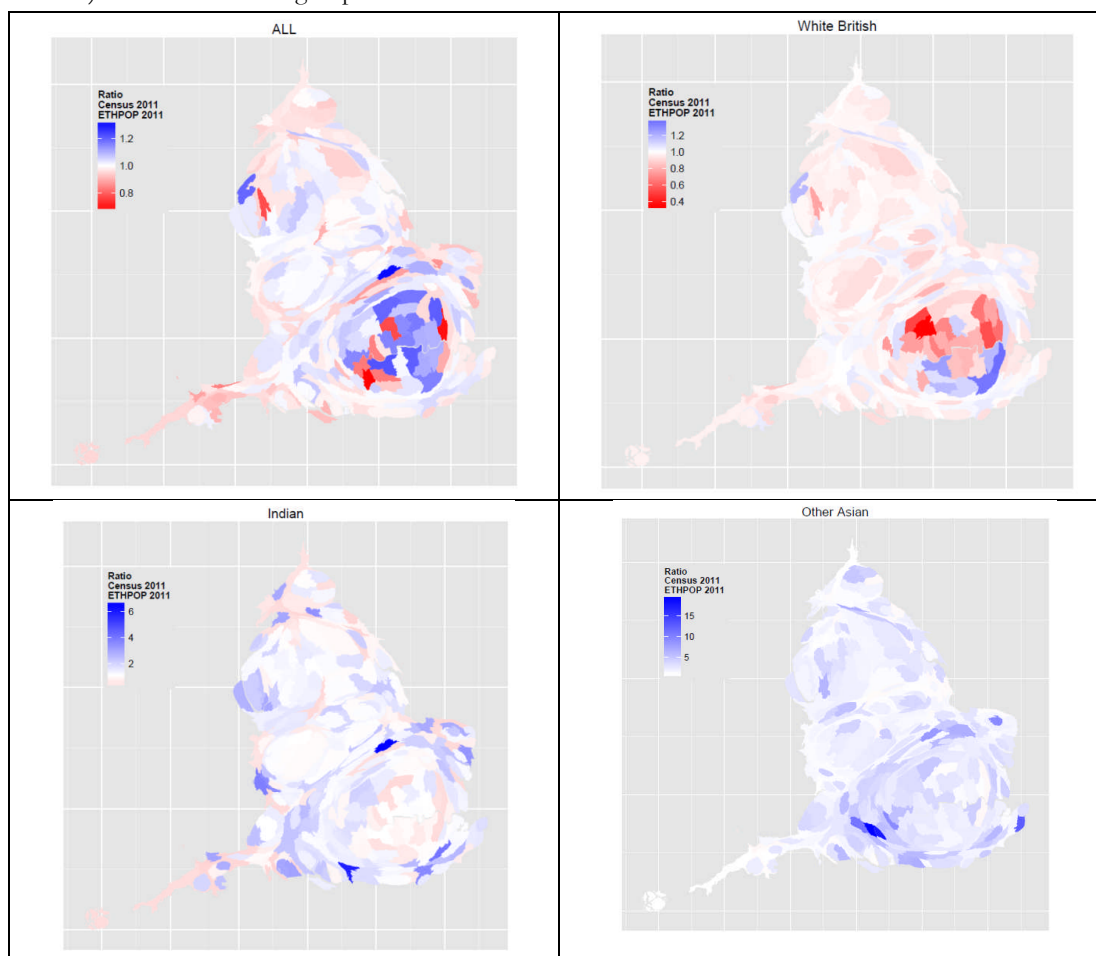
Finally, the residual group, the Other Ethnic Group was under-projected by 22%, reflecting the increasing diversity of origins of immigration to England and Wales.

Separate comparisons can be made for age groups (not reported here in detail). The patterns of over-projection (White groups) and under-projection (Black and Minority Ethnic (BAME) groups) are repeated for children ages 0-15), working age adults (ages 16-64) and the old (ages 65+). The relative differences from Census populations are a little higher for the childhood ages than for all ages, suggesting some under-estimation of ethnic minority fertility. For the working ages, the relative differences are a little lower than for all ages but with the same pattern across groupings. For the older ages we have over-projected the White groups relatively more than the total population and over-projected the Black groups as well. For the Asian grouping the older ages are closely projected but for the Mixed and Other Ethnic groups there is substantial under-projection. These differences suggest we have been too optimistic in our mortality decline assumptions and again failed to capture increasing immigration of Other Ethnic groups.

So far we have compared Census and ETHPOP results for England and Wales. Data are available for repeating this analysis across all 326 local authorities in England. Here we just look at the pattern for All

Groups, the White British, Indian and Other Asian groups in series of maps (Figure 3), based on population cartograms, using the method of Gastner and Newman (2004). The indicator used in the cartograms is the ratio of Census 2011 Population to ETHPOP 2011 Populations. Blue shades indicate that the Census population is higher than the ETHPOP projection, i.e. we have under-projected. Red shades indicate the Census population is lower than the ETHPOP projection, i.e. we have over-projected.

**Figure 3** Maps of local authorities showing over-projection (red shades) and under-projection (blue shades) for selected ethnic groups



Source: authors' computations.

In the top left map for All Groups, we see that a large of the population lives in local authorities which are reasonably well projected (White shades). Most local authorities which have been under-projected are found in Greater London. However, Greater London also contains Boroughs in which the populations have been over-projected, including Westminster, and Kingston upon Thames. It is likely these differences stem from errors in the projection of the internal migration component between 2001 and 2011 or in the estimation of the immigration and emigration components between 2001 and 2005, before revisions improving reliability were introduced (ONS 2011d).

The top right map presents the projection errors for the White British group. The most extreme values are seen in Great London, with over-projection in most London Boroughs, except for outer boroughs in the south east and south centre of Greater London. Clearly, our projections have under-estimated the increase in ethnic diversity in London, in which the share of the White British population has shrunk. The Indian group (bottom left map) was, by and large, well projected. There are some local authorities around the conurbation cores in which the group was under-projected (blue shades). In some London Boroughs there was some over-projection, along with some remoter rural local authorities in Cornwall, Devon, Norfolk, Cumbria and the North East. It is likely that the internal migration pattern has shifted over the decade in ways we were unable to capture. Our projections perform worst for the Other Asian group (bottom right map). The local authorities where the under-projection is greatest are in southern England outside of Greater London. However, there are many Local Authorities in the Midlands and northern England as well where the Other Asian population was under-projected.

The larger, longer established, groups are projected relatively well compared with the smaller, newer groups. Some of this is because there is less potential for small number exaggeration with groups such as the White British and Indian and some is because the demographic component evidence may be better estimated for these groups.

## 5. DISCUSSION

Here we summarise the findings of our analysis and experiment with various adjustments of our ethnic population projections. We also reflect on the question posed in the title of the paper.

Overall the average of our two projections was aligned closely to inter-census change. The White groups were over-projected and the ethnic minority groups were under-projected. The ordering from least to most under-projected was: Asian, Black, Mixed and Other, though the differences in degree of under-projection were not large. Examination of the detailed ethnic groups suggests that as time passes from the first wave of immigration the growth of a group slows down. The White Irish and Black Caribbean groups have been in England and Wales longest and have the slowest growth. This is a function of much lower current immigration and ageing of the settled populations of the group. The biggest wave of South Asian immigration was in the 1960s and 1970s and these groups have aged and converged in fertility towards the national norm. They experienced moderate growth. The groups which have experienced immigration most recently such as the Black African, Other Asian or Other Ethnic groups, had the greatest growth between 2001 and 2011, which our projections under-estimated. The Mixed groups also grew more than we had anticipated, which suggests we had did not capture fully the increase in mixed partnerships. Overall, the England and Wales population is diversifying much faster than we projected.

The most important conclusion is that the ethnic group projections need to be revised in the light of the 2011 Census. This will involve re-estimating the components of change by ethnicity and locality using a full decade of demographic information and data from the two “book-end” censuses. We need also to make assumptions for the future informed by the errors of the past decade, developing fully the transition theory suggested in the previous paragraph. This work is underway in a new ESRC funded project (Rees et al. 2015a).

Is there any way to fix our ETHPOP projections until revisions can be effected? Table 5 presents some alternatives applied to the 2051 average of the TREND and UPTAPER projections (second column). The third column applied the difference ratio for 2001-2011 given in the last column of Table 4 to the 2051 projected populations over 5 decades. The results are implausible: the White British population is halved to 27 million and the Other Asian population grows to 25 million. More realistic are the adjustments in the fourth column which add or subtract 5 times the 2001-2011 error. The results are more plausible but it is unlikely that the same errors will occur over the next four decades as happened in the previous decade. The fifth column of Table 5 simply uplifts the 2051 projected populations by the 2011 error, which is probably too conservative an adjustment. We have carried out further experiments with uprating the ETHPOP projections using knowledge of the 2001-2011 differences for the Policy Exchange (Rees and Clark 2014), the Government Office for Science (Rees et al 2015b) and as part of a PhD thesis (Clark 2015).

So, the answer the question posed in the title of this piece is that “our multi-ethnic future” is arriving a good deal faster than we thought. There is an urgent need to revise the ethnic group projections, which will be accomplished in 2015-16.

**Table 5** Alternative corrections to the ETHPOP 2051 average projection, England and Wales

16 Harmonized Ethnic Groups	ETHPOP populations Average 2051  (thousands)	ETHPOP populations multiplied by the multiplier 5 times and adjusted to All groups total (thousands)	ETHPOP populations uplifted by 5 × difference 2001- 2011 (thousands)	ETHPOP populations uplifted by difference 2001- 2011 (thousands)
All Groups	74,477	74,600	74,569	74,496
White British	57,604	27,008	50,417	56,167
White Irish	1,540	279	881	1,408
White Other	3,974	3,074	4,819	4,143
White and Black Caribbean	720	1,411	1,200	816
White and Black African	308	522	476	341
White and Asian	668	717	883	711
Other Mixed	563	581	735	598
Indian	2,380	1,395	2,474	2,399
Pakistani	1,916	1,984	2,587	2,050
Bangladeshi	669	953	1,057	746
Chinese	619	523	782	651
Other Asian	632	25,113	3,036	1,112
Black Caribbean	754	284	522	707
Black African	1,211	2,592	2,391	1,447
Other Black	231	6,857	1,002	385
Other Ethnic Group	691	1,306	1,309	815
Non-White British	16,873	47,591	24,152	18,329

Sources: see Table 3.

Notes: The multiplier for 2001-2011 is shown in the sixth column of Table 4. The difference is shown in the fourth column of Table 4.

## Acknowledgements

The ethnic projections presented in this paper were results from the ESRC Funded project, *Ethnic group population trends and projections for UK local areas: dissemination of innovative data inputs, model outputs, documentation and skills*, 1 October 2010 to 30 September 2011, ESRC Research Award RES-165-25-0162. The boundaries for Figure 2's population cartograms were supplied by Bethan Thomas of the University of Sheffield.

## Biographies

Philip Rees is Emeritus Professor of Population Geography at the University of Leeds, with interests in ethnic population projections, health outcomes and ageing of the population.

Dr Pia Wohland is a Senior Research Associate at the Institute for Health and Society (IHS) and Newcastle University Institute for Ageing (NUIA). She has researched ethnic mortality and health differences for local areas in the UK and developed software for ethnic population projections.

Dr Paul Norman is a Lecturer in Human Geography (Applied Spatial Analysis & Policy), with research interests in harmonisation of small area level socio-demographic, morbidity and mortality data to enable time-series analysis of demographic and health change, and in using individual level microdata to understand aggregate differences in population stratification and characteristics over time.

## References

- Clark, S. (2015) *Modelling the Impacts of Demographic Ageing on the Demand for Health Care Services*. PhD dissertation, School of Geography, University of Leeds
- Coleman D (2006) Immigration and ethnic change in low-fertility countries: a third demographic transition *Population and Development Review* 32 401–446
- EHRC (2013) *Equality Act*. Equality and Human Rights Commission. Online at: <http://www.equalityhumanrights.com/legal-and-policy/equality-act/>
- ETHPOP (2013) ETHPOP database. Online at: <http://www.ethpop.org/>
- Gastner, M.T. and Newman, M.E.J. (2004) Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20): 7499-7504
- GLA (2014) *GLA Demographic Projections*. Greater London Authority. Online at: <http://data.london.gov.uk/dataset/gla-demographic-projections>
- Jivraj, S. (2012) How has ethnic diversity grown 1991-2001-2011? Briefing in the *Dynamics of Diversity: Evidence from the 2011 Census* series, Prepared by the ESRC Census on Dynamics of Ethnicity (CoDE). Online at: [http://www.ethnicity.ac.uk/census/869\\_CCSR\\_Bulletin\\_How\\_has\\_ethnic\\_diversity\\_grown\\_v4NW.pdf](http://www.ethnicity.ac.uk/census/869_CCSR_Bulletin_How_has_ethnic_diversity_grown_v4NW.pdf)
- ONS (2011a) *Population Estimates by Ethnic Group 2002 – 2009*. Date: 18 May 2011, Coverage: England and Wales, Theme: Population. Online at: <http://www.ons.gov.uk/ons/rel/peeg/population-estimates-by-ethnic-group--experimental-/current-estimates/index.html>
- ONS (2011b) *Assessment of Reliability of the Population Estimates by Ethnic Group*. Online at: <http://www.ons.gov.uk/ons/rel/peeg/population-estimates-by-ethnic-group--experimental-/current-estimates/index.html>
- ONS (2011c) *Population Estimates by Ethnic Group: Important Note on Reliability of Estimates for Subnational Areas*. Online at: <http://www.ons.gov.uk/ons/rel/peeg/population-estimates-by-ethnic-group--experimental-/current-estimates/index.html>
- ONS (2011d) Improved Immigration Estimates to Local Authorities in England and Wales: Overview of Methodology. Online at: <http://www.ons.gov.uk/ons/guide-method/method-quality/imps/improvements-to-local-authority-immigration-estimates/index.html>
- ONS (2013a) *Annual Population Survey (APS)*. NOMIS Official Labour Market Statistics. Online at: <http://www.nomisweb.co.uk/articles/676.aspx>
- ONS (2013b) *Statistical Bulletin: Migration Statistics Quarterly Report, August 2013*. Online at: <http://www.ons.gov.uk/ons/rel/migration1/migration-statistics-quarterly-report/august-2013/msqr-august-2013.html>
- ONS (2013c) 2011 *Census Data for England and Wales on Nomis*. Online at <http://www.nomisweb.co.uk/census/2011>
- Rees, P., Wohland, P., Norman, P. and Boden, P. (2011) A local analysis of ethnic group population trends and projections for the UK *Journal of Population Research* 28 149–84
- Rees, P., Wohland, P., Norman, P. and Boden, P. (2012a) Ethnic population projections for the UK, 2001–2051. *Journal of Population Research*, 29(1): 45-89. DOI: 10.1007/s12546-0111-9076-z
- Rees, P., Wohland, P. and Norman, P. (2012b) The demographic drivers of future ethnic group populations for UK local areas 2001–2051. *Geographical Journal*, 179(1): 44-60. DOI: 10.1111/j.1475-4959.2012.00471.x
- Rees, P. and Clark, S. (2014) The projection of ethnic group populations aged 18 and over for Westminster Parliamentary Constituencies in Great Britain for election years 2015, 2020, 2025, 2030 and 2035. A Report to the Policy Exchange, 10 Storey's Gate, London SW1P 3AY. School of Geography, University of Leeds
- Rees, P., Wohland, P., Norman, P., Lomax, N. (2015a) *Evaluation, Revision and Extension of Ethnic Population Projections – NewETHPOP*. Funded by ESRC, Grant Ref ES/L013878/1, 1 Jan 2015 to 31 May 2016
- Rees, P., Norman, P. and Durham, H. (2015b) Urban Social Disparities. Working Paper for the Foresight Future of Cities Project, Government Office for Science
- Wohland P., Rees P., Norman P., Boden P. and Jasinska M. (2010) Ethnic population projections for the UK and local areas, 2001–2051. *Working Paper 10/02*, School of Geography, University of Leeds, June 2010. Online at: <http://www.geog.leeds.ac.uk/fileadmin/documents/research/csap/10-02.pdf>
- Wohland, P., Norman, P. and Rees P. (2014) ETHPOP Database (projected populations, presentations and publications). Online at: <http://www.ethpop.org/>



# Spatially modelling dependent infrastructure networks

Robson C.<sup>1</sup>, Barr SL.<sup>2</sup>, James P.<sup>3</sup> and Ford A.<sup>4</sup>

School of Civil Engineering and Geosciences, Newcastle University

November 07, 2014

## Summary

The resilience of infrastructure networks to different types of perturbation is of significant interest due to the importance of these systems for the efficient functioning of modern society. Significant failures such as power blackouts and their subsequent knock-on effects on other infrastructures are particularly important to understand. We present a spatially explicit method for representing infrastructure spatial dependencies and modelling of failure impacts using a PostgreSQL-PostGIS database coupled with spatial network failure models. The utility of the methodology is shown by simulating how the London underground may respond to different failure types on the South East electricity transmission grid.

**Keywords:** Dependency, resilience, critical infrastructures, spatial networks, spatial database

## 1. Introduction

Critical infrastructures (CI) facilitate active economies and underpin the quality of life in modern nations. Failures or disturbances to these CI, such as the Italian and North American electricity blackouts of 2003 (Andersson *et al.*, 2005), can lead to far reaching societal and economic consequences. Such failures affect not only a single CI, but instead failures propagate to other infrastructure networks exacerbating the effects. For example, failure of electricity transmission systems affect multiple CI, including communications (the internet) and transport systems, and in the case of the 2003 blackout in USA and Canada it was estimated to have had an economic cost of 4-10 billion dollars (U.S.-Canada Power System Outage Task Force, 2004).

Understanding the relationships between interdependent infrastructure systems is vital to ensure the repercussions of failure in one network does not adversely affect the multiple systems which may be dependent upon it. Dependencies between CI can take multiple forms (Rinaldi *et al.*, 2001), from physical connections to cyber dependencies and geographic dependencies, all of which can result in failures propagating between infrastructure systems. An ever increasing number of CI exhibit dependencies, and thus the analysis of these systems should account for this (Dueñas-Osorio *et al.*, 2007) to improve our ability to understand the consequences of failures in CI.

In this paper we introduce a spatial analysis of the relationship between two CI, the National Grid and the London tube network using a spatial-topological based analysis approach to highlight the potential vulnerability of the tube system to electricity substation failures. This is facilitated using a PostgreSQL PostGIS enabled database within a python driven framework.

## 2. Method

Interdependent spatial networks were constructed using a dedicated postgresSQL PostGIS enabled interdependent network database schema model (Barr *et al.*, 2013). National Grid shapefiles (National

---

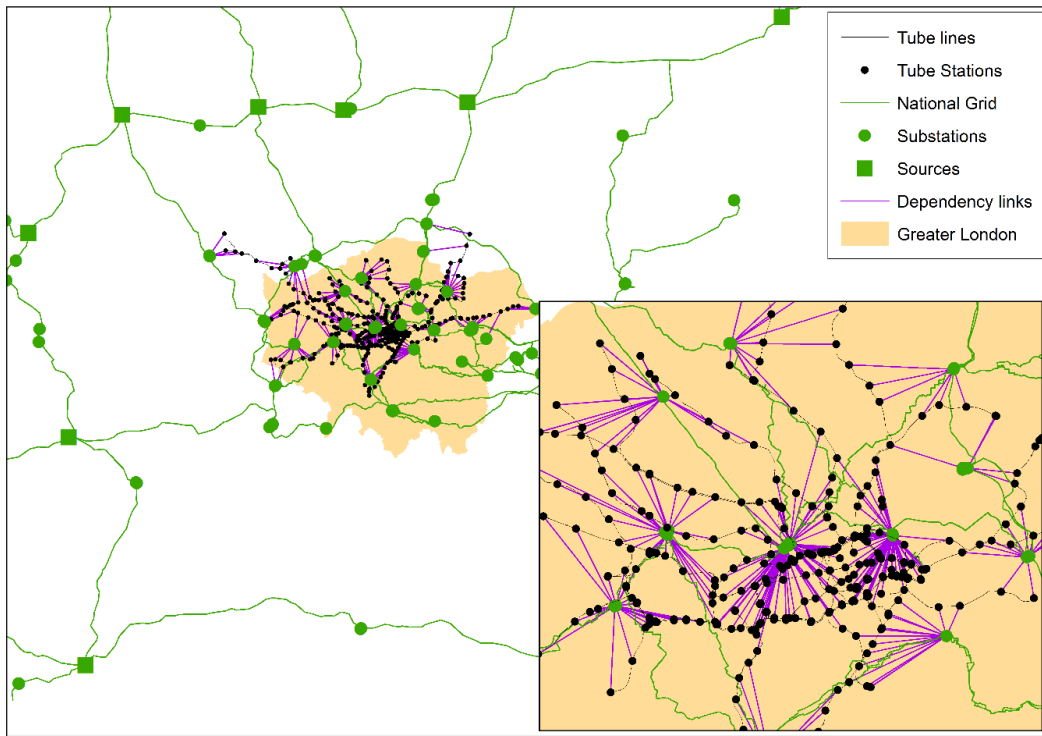
<sup>1</sup> [c.a.robson1@ncl.ac.uk](mailto:c.a.robson1@ncl.ac.uk)

<sup>2</sup> [stuart.barr@ncl.ac.uk](mailto:stuart.barr@ncl.ac.uk)

<sup>3</sup> [philip.james@ncl.ac.uk](mailto:philip.james@ncl.ac.uk)

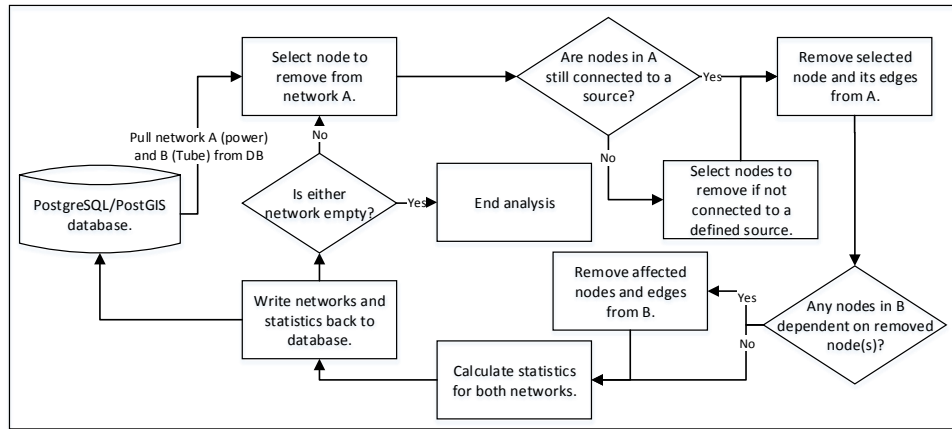
<sup>4</sup> [alister.ford@ncl.ac.uk](mailto:alister.ford@ncl.ac.uk)

Grid, 2014) were processed using a suite of software tools for spatio-topological network construction (Barr *et al.*, 2013) to generate a power network for the South East of England comprising of 709 nodes and 877 edges. Data from Transport for London (TFL) was employed to build a representation of the tube network (Transport for London, 2014) that consists of 436 nodes and 466 edges. To build the relationship between the two networks (Figure 1) we make a number of assumptions; (i) each tube station is powered by its geographically closest electricity substation, (ii) there are no other sources of electricity for each station, and (iii) the substations at the edge of power network are the sources of power for the South East area. Within the interdependent network database schema the dependency relationships between the electricity and tube networks are stored in a separate interdependency table using node id's to record the substation that each tube station is dependent on.



**Figure 1:** The electricity grid for South East England and the London tube network. Inset: Central London showing spatial dependency edges (mappings) between electricity substations and tube stations.

In order to perform interdependent failure analysis networks are retrieved from the database forming NetworkX instances (NetworkX, 2014) and the particular failure model of choice initiated: either (i) random, (ii) degree (a measure of the most connected nodes) or (iii) betweenness (an estimate of the load on each component) (Boccaletti *et al.*, 2006) (Figure 2). Each failure model involves removing a selected/targeted substation every iteration, checking to ensure remaining substations are still connected to a 'source' node (i.e., that the electricity network is still functional), and then via the dependency list identifying those tube stations that will fail as a result of a dependency to the selected/targeted substation. The failed station nodes are removed along with their coincident edges (tube lines). Performance metrics are then computed, and along with the current instance (state) of both networks and their dependencies, written back to the database. The process is repeated until no nodes/edges are left in one network.

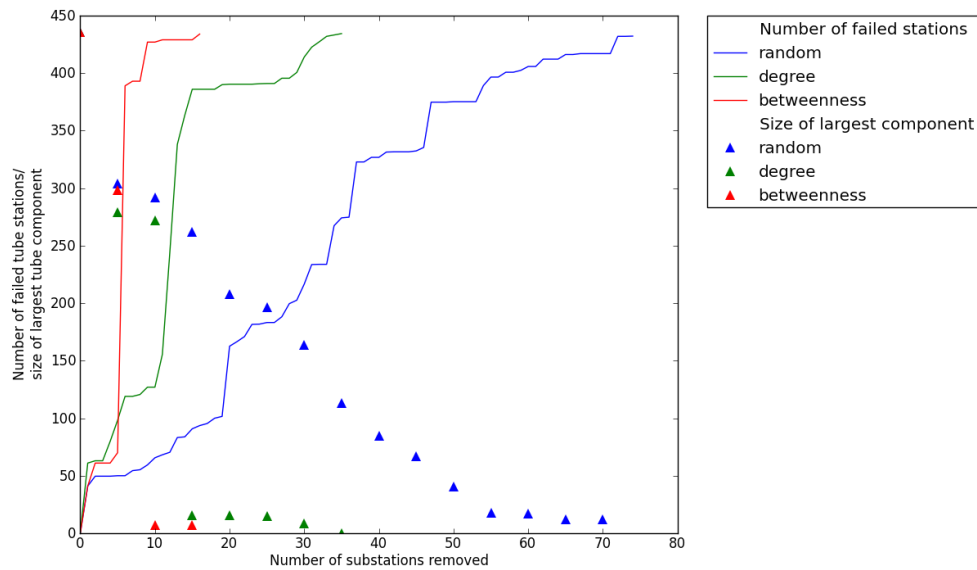


**Figure 2:** Flow diagram of interdependent failure model.

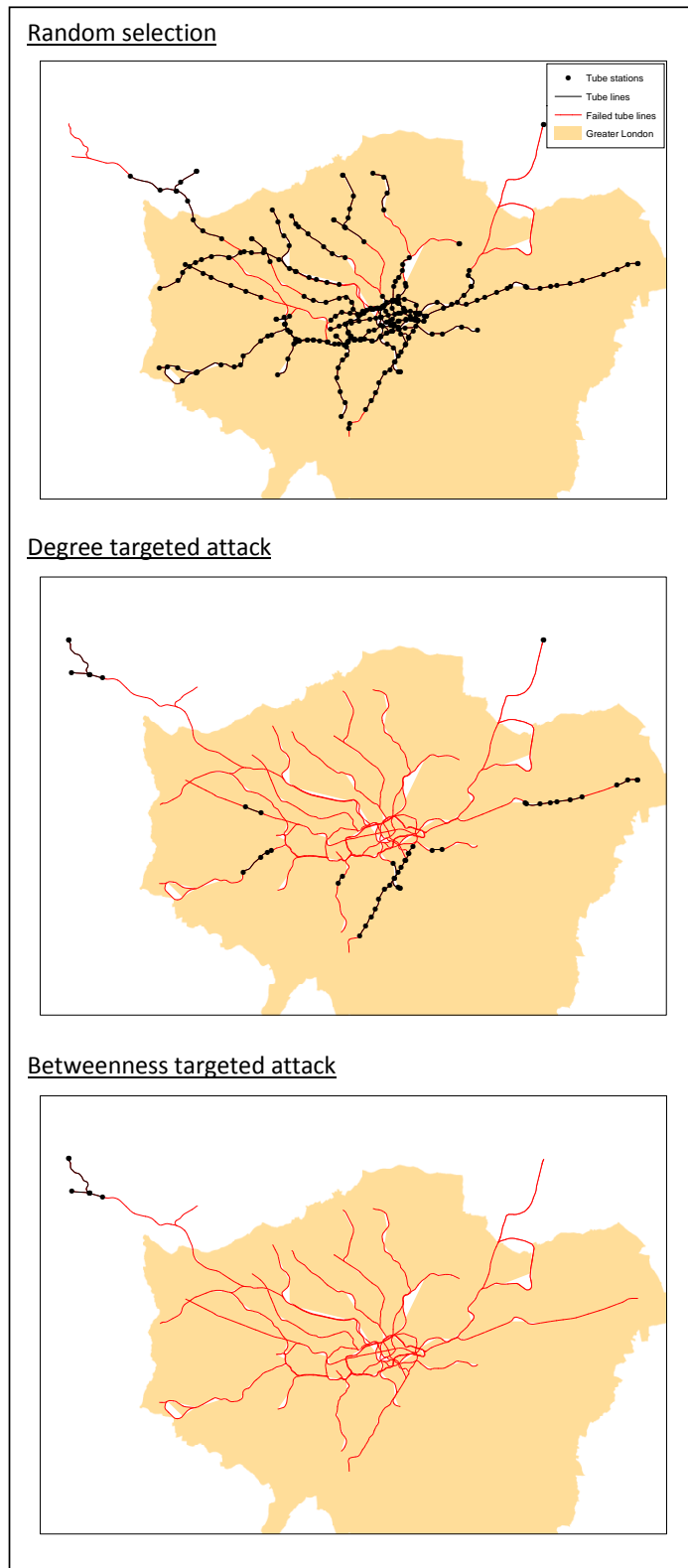
### 3. Results

The results show that the degree and betweenness failure models have a greater impact on the tube network (Figure 3) with all tube stations becoming disconnected earlier (17 and 10 substations removed respectively) compared to the random model (70 substations removed before complete failure). Not only does the betweenness model lead to quicker total failure but that the effects of the failures in this model are far more dramatic at early stages of failure, as shown by the fact that the largest connected functioning component of the tube network decreases significantly quicker for the betweenness model than for the random (Figure 3).

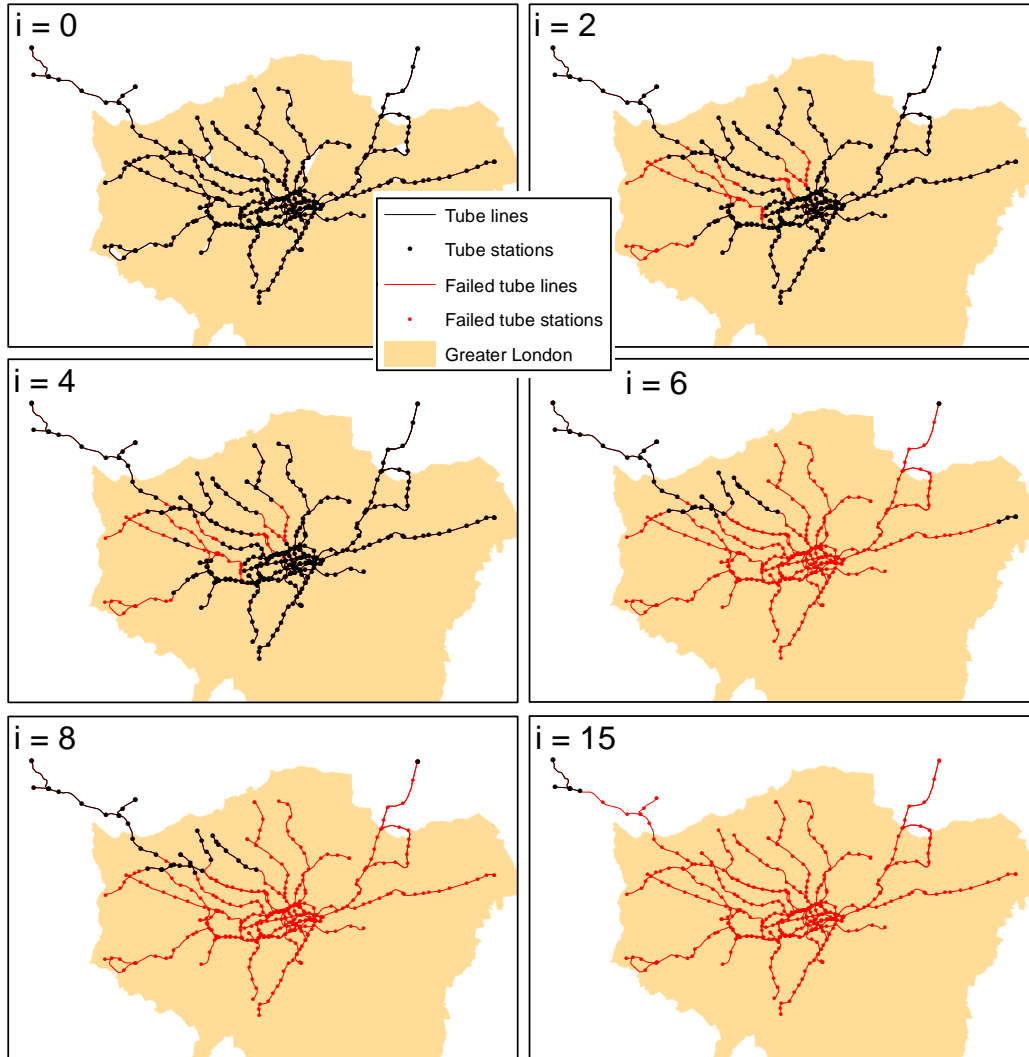
The spatial pattern of failures for the three failure models shows that both the degree and betweenness methods have a greater impact spatially in the centre of London, whereas the random failure model does not target a specific spatial area thus allowing travel across the network to remain possible for longer (Figure 4). The sensitivity of the tube network to the betweenness failure type is shown in Figure 5 with 61 tube stations failing after only four iterations and 319 by the sixth.



**Figure 3:** Change in the number of failed tube stations and the size of the largest group of connected tube stations as the number of failed substations increase.



**Figure 4:** The state of the tube network after fifteen iterations for a single run for each failure model.



**Figure 5:** Evolution of the tube network failure when the power network is exposed to the betweenness failure model.

#### 4. Conclusion

This paper has presented a software framework that couples a new open-source database schema representation of interdependent spatial networks with iterative network failure models. The resulting software not only allows for several different types of network failure to be run, but the use of a database management system allows a persistent record of the failure dynamics at each iteration to be recorded and analysed further. In the case of this study, this allows a user to not only understand the global characteristics of the failures, but also critically to understand the evolving spatial pattern of failure and how this differs between the three models investigated. The results for the electricity and tube networks show that potentially dependent systems are very vulnerable to failures on the most highly connected nodes but robust to spatially random failures in the supplying network. Future work will extend this spatio-topological analysis to develop models with a stronger physical-basis.

## 5. Biography

Mr Craig Robson received a B.Sc. (Hons) degree in Geographic Information Science from Newcastle University in 2011. He is currently studying for a Ph.D. in spatial infrastructure network modelling at Newcastle University.

Dr Stuart Barr is a Senior Lecturer in Geographic Information Science at Newcastle University.

Mr Philip James is a Senior Lecturer in Geographic Information Science at Newcastle University.

Mr Alistair Ford works as a Researcher in Geomatics at Newcastle University.

## Bibliography

Andersson, G., Donalek, P., Farmer, R., Hatziargyriou, N., Kamwa, I., Kundur, P., Martins, N., Paserba, J., Pourbeik, P., Sanchez-Gasca, J., Schulz, R., Stankovic, A., Taylor, C. and Vittal, V. (2005) 'Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance', *Power Systems, IEEE Transactions on*, 20, pp. 1922-1928.

Barr, S.L., Alderson, D., Robson, C., Otto, A., Hall, J., Thacker, S. and Pant, R. (2013) 'A National Scale Infrastructure Database and Modelling Environment for the UK', *International Symposium for Next Generation Infrastructure*. Wollongong, New South Wales, Australia.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.U. (2006) 'Complex networks: Structure and dynamics', *Physics Reports*, 424, pp. 175-308.

Dueñas-Osorio, L.A., Craig, J.I., Goodno, B.J. and Bostrom, A. (2007) 'Interdependent Response of Networked Systems', *Journal of Infrastructure Systems*, 13, pp. 185-194.

National Grid (2014) *National Grid - Services - Shape-Files*. Available at: <http://www2.nationalgrid.com/uk/services/land-and-development/planning-authority/shape-files/> (Accessed: 24/10).

NetworkX (2014) *NetworkX: Overview*. Available at: <https://networkx.github.io/>. (Accessed: 24/10).

Rinaldi, S.M., Peerenboom, J. and Kelly, T. (2001) 'Identifying, understanding and analysing Critical Infrastructure Interdependencies', *IEEE Control Systems Magazine*, 21, pp. 11-25.

Transport for London (2014) *Transport for London*. Available at: <https://www.tfl.gov.uk/> (Accessed: 28/10).

U.S.-Canada Power System Outage Task Force (2004) *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*.

# The Complexity of Exclusion

J.M. van Rooyen<sup>\*</sup>, Joana Barros<sup>†</sup>

Department of Geography, Environment and Development Studies  
Birkbeck, University of London

June 05, 2015

## Summary

Two decades have passed since South Africa celebrated its first democratic election and witnessed the emergence of a new government. However, the legacy of spatial segregation and economical exclusion persisted. Formulating an understanding of this phenomenon provides the basis for this research project. By applying agent-based modelling, the study aims to analyse the complex composition of two societies and to demonstrate the potential for integrated future development.

**KEYWORDS:** segregation, complexity, agent-based modelling, integration

## 1. Introduction

The subjects of South African planning policy and complex social systems are both considered in the subsequent paper. The aim is to study the complexity of socio-economic disparity at a micro-scale, by considering individual preference in terms of household relocation and potential emergent patterns as result. The following discussion focuses on the development of an agent-based model to study the dynamics of the study area and provides a concurrent focus on a specific part of planning legislation implemented to promote urban integration.

## 2. An Urban Policy for Renewal

Before the transition to a new democratic country began in 1994, an extremely complex apartheid system resulted in the underdevelopment of rural areas and urban degeneration and exclusion. In 2001 the URP (Urban Renewal Programme) was launched, which reflected the main objective of poverty alleviation through investment in the economic and social infrastructure of most deprived areas. In the Western Cape, the government identified the suburbs of Khayelitsha and Mitchell's Plain as focus areas for the renewal initiative (Urban Renewal Programme, 2006). Both of these areas are situated within the municipal boundaries of the City of Cape Town. The general aims of the URP for these areas are to empower systematic and sustained interventions in order to address poverty and also to address under-development and socio-economic exclusion.

## 3. Methodology

The objective of the research study is to explore the dynamics of segregation of the above-mentioned communities. A prototype agent-based model, based on Thomas Schelling's "preferential" segregation modelling (Schelling, 1971), is applied to the case study area. In addition the phenomenon of segregation by exclusion from opportunities is studied simultaneously. In this particular study the phenomenon of preferential segregation is considered only in terms of cultural choice and not racial prejudice. The development of the Delft South neighbourhood in Cape Town serves as good example. Cultural integration was encouraged by the provision of housing to Black African and Coloured families in a random (mixed) way, but segregated self-organisation still occurred (Oldfield, 2004).

---

<sup>\*</sup> c.vanrooyen@mail.bbk.ac.uk

<sup>†</sup> j.barros@bbk.ac.uk

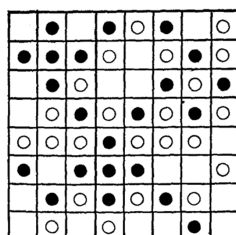
Hence, the prototype model tests Schelling's concept in the neighbourhood environment of Khayelitsha and Mitchell's Plain, where the different types of agents (racial groups) are initially apart from each other (spatially) and not randomly mixed as found in the Schelling model.

## 2.1. Primary Data

The study area boundaries were obtained in GIS format and contain the size and also the neighbourhood type of the study area. Census boundaries and local land use classification boundaries were obtained to study the dynamics of the study area.

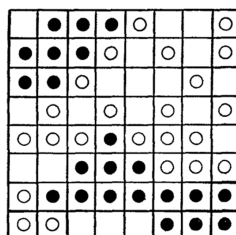
## 2.2. Schelling model

Thomas Schelling (1969) developed the model of segregation to point out how even slight individual preferences and perceptions of difference can collectively result in segregation. He constructed a two dimensional model (Schelling, 1971) and divided the study area into a matrix of similar sized cells. Various cells were left vacant (Figure 1), while the rest were randomly occupied by individuals from two different groups.



**Figure 1** Initial Condition of Schelling's Model (Schelling, 1974)

The movement rule was established and dictated that an individual, discontent with his neighbourhood (less than 50% similar neighbours), will move to the nearest vacant location, surrounding himself with a neighbourhood meeting his needs. The pattern resulting from this process reflected a highly segregated state (Figure 2).



**Figure 2** Outcome of Schelling's Model (Schelling, 1974)

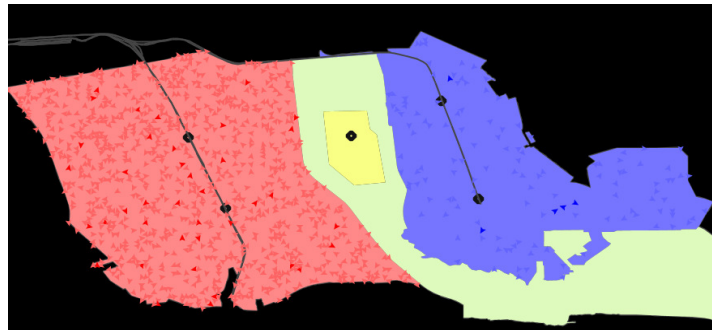
## 4. An Agent-Based Model for the Suburbs of Khayelitsha and Mitchell's Plain

This model, developed in Netlogo (Wilensky, 1999), simulates the dynamics of two racial groups (Black African and Coloured) in two suburbs of the City of Cape Town. The prototype model assumes that each agent (here representing a household) wants to ensure that it lives in proximity to its own culture. Furthermore, the assumption is also made that all agents desire to better their socio-economic circumstances. Thus, certain agents are happy where they live and others want to move closer to better opportunities, such as more affordable housing or better job opportunities. The



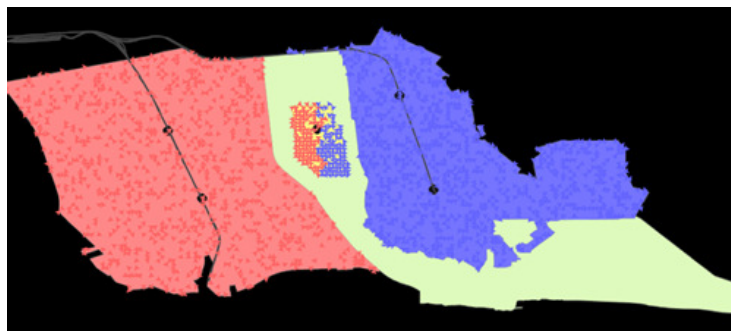
unhappy agents of both cultures are portrayed as darker features, compared to the lighter coloured happy agents. The simulation shows how individual preferences at a local scale can influence community development and evolution at a larger scale.

Agents are initially located in their respective neighbourhoods (Figure 3). These neighbourhoods reflect GIS polygons obtained to establish their extent in reality and are the same colour as the agents. The green area between the two neighbourhoods represents vacant land and the yellow area contained within the vacant land represents a new housing and business development.



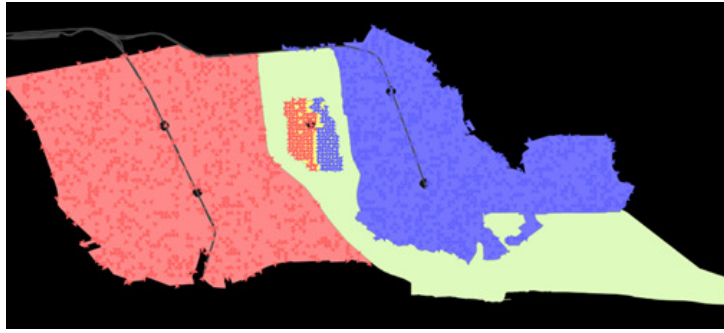
**Figure 3** Agents at Initial Setup Stage (Black-African agents represented in blue and Coloured agents in red)

In the first simulation run only 11% of the overall households for each cultural group was specified as having a desire to move to the new development, with a 30% preference for their own culture. The outcome in Figure 4 shows a small number of households moving to the new development and more importantly signs of integrated patterns. The conclusion is that the desire for a better home or employment outweighs the desire to live exclusively amongst neighbours of a similar cultural group. The reason for specifying only 11% is to ensure that the new development can accommodate all households with the desire to move there.



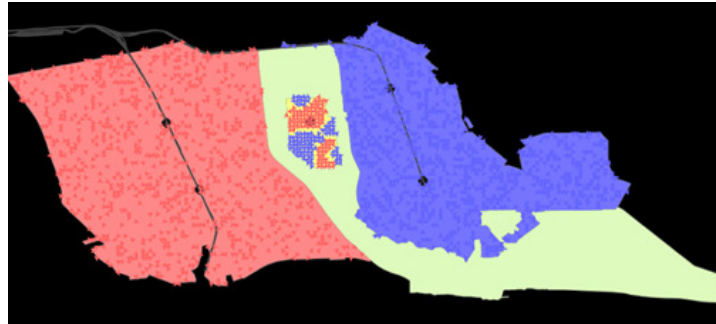
**Figure 4** 11% of Population Desiring New Development (30% preference of own kind)

In the second simulation scenario that was tested, 11% of the overall households of each cultural group with a desire to move to the new development was specified again (for the same reason as before). However, a 60% preference for own kind is applied this time. The outcome in Figure 5 indicates the same number of households relocating to the new area, but with a distinct open space between the two sets of agents.



**Figure 5** 11% of Population Desiring New Development (60% preference of own kind)

Finally, a third scenario was tested (Figure 6) whereby 11% of the households of both societies relocated yet again. However, during this run the percentage preference for their own kind was gradually decreased from 90% to 30%. The outcome in Figure 6 provided quite a different pattern than the initial run at 30% preference for own kind, which occurred in Figure 4.



**Figure 6** 11% of Population Desiring New Development (preference reduced from 90% to 30% during simulation run)

## 5. Discussion and Conclusion

Although the model presented in this paper is only in the initial stages of development (prototype), the outcome of the scenarios above demonstrate that the individual preferences and choices of the agents at a local scale result in emergent patterns at a macro scale. By merely applying the variables of “preference to own culture” and “desire to relocate” is it evident that interaction amongst agents is complex and inter-dependent. This is particularly apparent in the last simulation run where a different pattern formed, due to a different initial condition in household preferences.

However, the limitation is that the neighbourhood scale applied to the prototype model does not take into account the more realistic diversity in race, culture and variation in land use that is evident in the city as a whole.

The next stage of the research project expands the model to a city-wide scale. Hence, agents comprise of three different racial groups (households) and in addition the variables and parameters of income, residential dissonance and dwelling type are applied.

## **6. Biography**

Jacobus M. van Rooyen is a part-time PhD researcher at Birkbeck, University of London and in his fifth year of study. Interests include agent-based modelling of social systems, complex adaptive systems and the phenomenon of urban emergence.

Joana Barros is a lecturer in GI Science at Birkbeck, University of London and Jacobus' supervisor. Her areas of expertise are urban planning and modelling, more specifically agent-based and cellular automata models applied to urban systems and urbanisation in developing countries.

## **References**

National Urban Renewal Programme: Implementation Framework (2006) DPLG, Pretoria

Oldfield, S., 2004. Urban networks, community organising and race: an analysis of racial integration in a desegregated South African neighbourhood. *Geoforum*, Themed Section on: Differentiation in South Africa and Indian Cities 35, 189–201. doi:10.1016/j.geoforum.2003.08.006

Schelling, T.C., 1969. Models of Segregation. *Papers and Proceedings of the Eighty-First Annual Meeting of the American Economic Association* 59, 488–493.

Schelling, T.C., 1971. Dynamic models of segregation†. *Journal of mathematical sociology* 1, 143–186.

Schelling, T.C., 1974. On the Ecology of Micromotives. *The Corporate Society*. Marris, R. (ed). London: Macmillan 19–64.

Wilensky, U. (1999). NetLogo. <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

# A self-exciting point process model for predictive policing: implementation and evaluation

G. Rosser<sup>\*1</sup> and T. Cheng<sup>†1</sup>

<sup>1</sup> SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering, University College  
London, London WC1E 6BT

November, 2014

## Summary

The self-exciting point process (SEPP) model has recently been shown to perform well in predicting spatiotemporal crime patterns. However, this model has not been widely applied to crime data and many open questions remain about how best to implement it in a real setting. In this work, we consider a range of practical implementation details relating to the application of SEPP models to real crime data. We propose a robust protocol that optimises the performance of the method, and suggest guidelines for parameter selection.

**KEYWORDS:** predictive policing, self-exciting point process, kernel density estimate, machine learning

## 1. Introduction

The criminological theory of near repeat victimisation states that the occurrence of certain crimes increases the risk of further crimes within the local neighbourhood for some ensuing time period (Johnson and Bowers, 2004; Youstin et al., 2011). As a result of this process, crime events tend to cluster in space and time. Predictive policing is concerned with identifying emerging crime ‘hotspots’ using forecasting methods. This has been the target of much recent research effort (S. Chainey et al., 2002; Bowers et al., 2004) as such a method would be of great utility to police forces worldwide. Existing methods commonly apply statistical analysis or heuristic algorithms to crime data with the aim of identifying hotspots and localising them in time and space (see, for example, Bowers et al., 2004). Such approaches are valuable and have shown reasonable predictive power, however they are better suited to retrospective analysis than forecasting since the underlying methods are not based on well-stated models. Furthermore, such analyses give little insight into the underlying method of generation of crime patterns.

The subject of this work is the application of a self-exciting point process (SEPP) model to crime data. The point process framework is well suited for time and geolocation tagged data, such as records of crime. Methods based on point processes have previously been developed to detect space-time clustering (Diggle et al., 1995), which is useful for retrospective analysis. The SEPP model has been used in the field of seismology to predict earthquake sequences for several decades. As we describe in detail below, this model describes a dynamic point process in space and time in which events may trigger further events within their spatial and temporal neighbourhood (self-excitation). In a promising advance in the field of criminology, Mohler et al. noted the similarity between this model and the criminological theory of near repeat victimisation, and applied the SEPP model to the predictive modelling of crime data (Mohler et al., 2011). Their method outperforms a kernel-based hotspot

---

<sup>\*</sup> g.rosser@ucl.ac.uk

<sup>†</sup> t.cheng@ucl.ac.uk

detection approach in terms of predictions made on real crime data.

Despite the apparent advantages of the SEPP framework for predictive crime modelling, there are several open questions and issues preventing the widespread adoption of the method. Most notably, the process of training the model (i.e. inferring parameters) on data involves the use of kernel density estimates (KDE), whose underlying kernel functions and bandwidths may have a significant effect on the predictive performance of the model (Mohler et al., 2011), or prevent the training algorithm from terminating successfully. In addition, the method is computationally intensive due to the necessity of repeatedly evaluating the KDE at a large number of data points (typically millions to tens of millions per iteration). Finally, there is no available open source implementation of the SEPP model for crime data, which hampers further research and development of the methods discussed.

The subject of this abstract is the development of a robust computational tool to apply the SEPP to crime data. We consider the following real-world implementation issues: (a) the effect of imposing upper thresholds on the temporal and spatial maximum triggering extents; (b) the effect of changing the temporal or areal domain upon which the model is trained; (c) the effect of modifying the kernel functions in the KDE.

We assess the predictive performance of our method using appropriate validation methods, such as the measure of search efficiency rate. We apply our method to open crime data provided by the city of Chicago, USA, to demonstrate its effectiveness.

## 2. Materials and Methods

### 2.1. Self-exciting point process

At the core of the SEPP model of crime is the conditional intensity,  $\lambda(t, x, y)$ , which gives the density of the expected rate of occurrence of crimes in a small neighbourhood around the region  $(x, y)$  at time  $t$ , conditional upon the history of all occurrences up to that time. The conditional intensity may be described as the sum of background and triggered events:

$$\lambda(t, x, y) = \mu(t, x, y) + \sum_{\{k: t_k < t\}} g(t - t_k, x - x_k, y - y_k), \quad (1)$$

where  $\mu$  denotes the background occurrence rate and  $g$  denotes the triggering function. Thus all crimes that have occurred prior to a given time may theoretically contribute some additional expectation of the current crime activity, though in practice this may vanish over some period of time and/or distance.

In order to apply this theory to real data we must estimate the functional forms of  $\mu$  and  $g$ . In practice, this entails declustering the data (Zhuang et al., 2002) to identify those events arising from background activity and those triggered by previous events. Common approaches in the seismology literature involve maximum likelihood estimates based on assumed forms of  $\mu$  and  $g$  (Daley and Vere-Jones, 2003). An alternative approach, employed by Mohler et al. (2011), employs KDEs to avoid this necessity (Zhuang, 2006).

Let  $p_{ji}$  denote the probability that event  $i$  was triggered by event  $j$ . By convention,  $p_{ii}$  denotes the probability that event  $i$  is a background event. Furthermore,  $p_{ji} = 0$  if  $t_i < t_j$ , so that all of the probabilities may be encoded in an upper triangular matrix  $P$ . Under the assumptions of equation (1), these probabilities are given by

$$p_{ii} = \frac{\mu(t_i, x_i, y_i)}{\lambda(t_i, x_i, y_i)} \quad (2)$$

$$p_{ji} = \frac{g(t_i - t_j, x_i - x_j, y_i - y_j)}{\lambda(t_i, x_i, y_i)}. \quad (3)$$

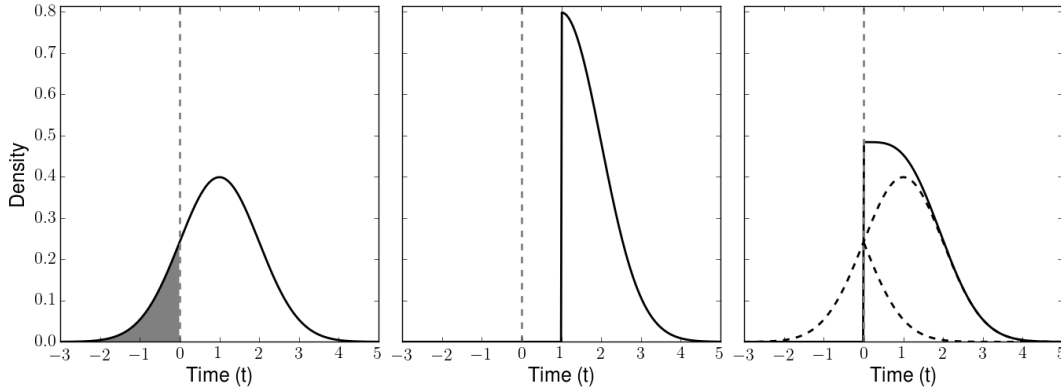
In (Mohler et al., 2011), an optimisation routine is proposed in which the background and parent/child events are sampled randomly from the data using the probabilities in  $P$ . From these samples, a KDE is computed and  $P$  is updated following equations (2) and (3). This algorithm has been validated using simulated data.

### 2.1.1. Triggering thresholds

In order to calculate  $p_{ji}$  (equations (2) and (3)), the KDE  $g$  must be evaluated  $N(N - 1)/2$  times, where  $N$  is the number of data points in the dataset. This step becomes prohibitively computationally expensive with increasing datasets. To limit the number of evaluations, it is necessary to impose maximum temporal and spatial extents,  $\Delta t_{max}$  and  $\Delta d_{max}$ , respectively, above which triggering is discounted. The matrix  $P$  thus becomes sparse, with non-zero entries only where pairs of data points lie within the threshold.

### 2.1.2. Kernel function

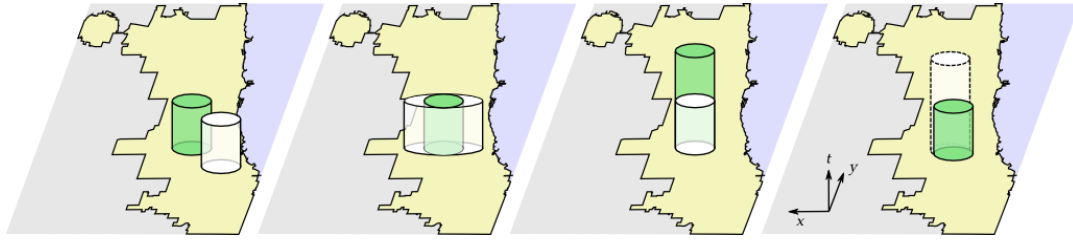
Both  $\mu$  and  $g$  are inferred using a KDE. In the case of  $g$ , triggering is only realistic when the time difference,  $\Delta t = t_i - t_j$ , is positive. However, a three-dimensional multivariate Gaussian kernel function is used in (Mohler et al., 2011), which permits density at negative  $\Delta t$ . We consider the effect of changing the kernel function used for the temporal component of the triggering KDE, shown in Figure 1. The spatial kernel functions are unchanged.



**Figure 1** Three kernel functions considered for the temporal component of the triggering KDE. The function plotted is the marginal pdf in the temporal dimension, with a mean of 1. (Left) standard Gaussian, shaded region indicates density at negative time differences; (centre) one-sided Gaussian; (right) Gaussian reflected at  $\Delta t = 0$ , dashed lines illustrate the reflected portion.

### 2.1.3. Effect of temporal and spatial domain

Varying the spatial or temporal domain bounding the training data affects the performance of the SEPP model (Mohler et al., 2011) in an unknown manner. We are currently assessing the effect of spatial and temporal domain translation and enlargement, as indicated in **Figure 2**. Results will be presented at the GISRUK 2015 conference.



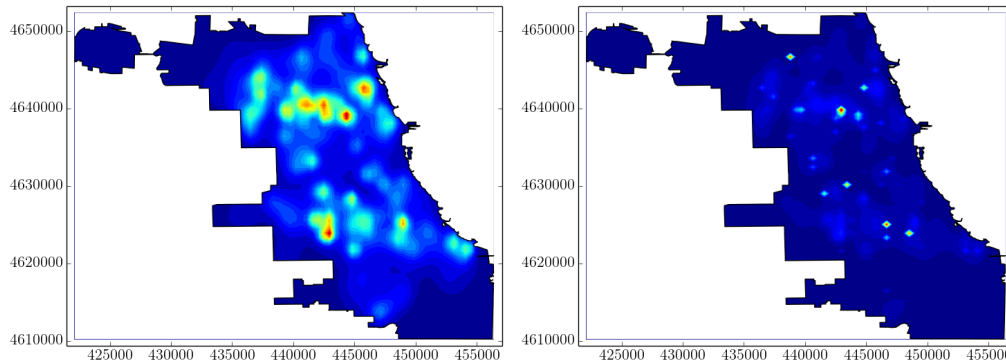
**Figure 2** Illustration of the spatial and temporal domains within the city of Chicago used to train the SEPP model. (From left): spatial translation; spatial enlargement; temporal translation; temporal enlargement.

### 3. Results

#### 3.1. Applying SEPP to Chicago crime data

For this study we are using crime data available on the City of Chicago's online data portal at <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>. Georeferenced, timestamped data are available from 2001 to the present.

Figure 3 shows the crime density heatmap computed using the SEPP for burglaries in February 2001.



*Figure 3* Density maps computed using the SEPP for burglaries in the Chicago region in February 2001. (Left) background density; (right) combined background / trigger density.

#### 3.2. Varying triggering thresholds

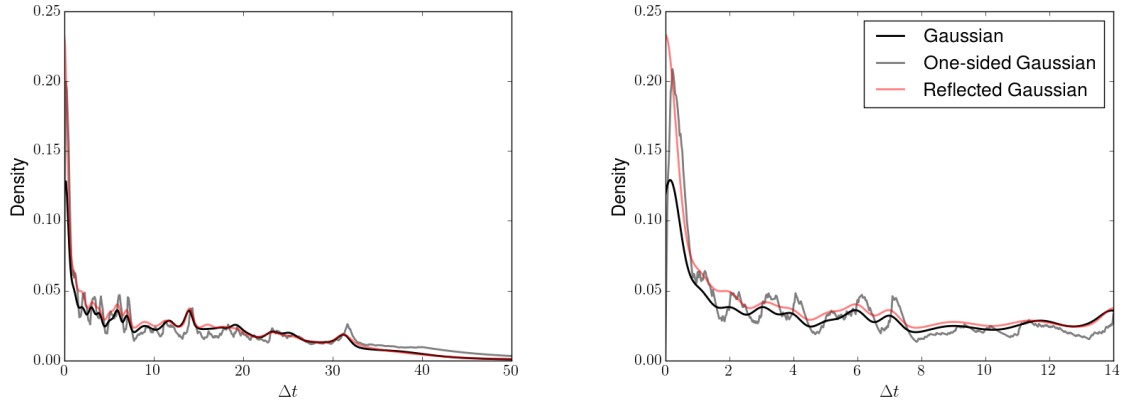
As Table 1 indicates, the number of permissible triggering pairs in the SEPP model changes significantly with the spatial and temporal threshold limits. Even with unrestrictive values of these limits, the number of links only reaches 10% of the maximum. Work is ongoing to assess the effect of these thresholds on the predictive performance of the SEPP model.

**Table 1** Variation of the number of triggering pairs in the SEPP model with trigger thresholds. The number of data points included was 11902, giving a maximum of 70822851 pairs.

		$\Delta t_{max}$ (days)						
$\Delta d_{max}$ (metres)		10	20	30	40	50	60	90
	20	528	773	948	1068	1202	1304	1567
	50	842	1336	1713	2047	2352	2621	3257
	100	1545	2652	3589	4439	5231	5936	7587
	200	4478	8162	11362	14409	17159	19750	25830
	300	8732	16248	22878	29206	34886	40207	53078
	500	20942	39913	56860	72646	87090	100485	133398
	1000	70184	134824	194304	248863	299809	346543	464228
	5000	1144490	2206618	3192283	4113140	4978984	5781365	7800521

### 3.3. Effect of the choice of kernel function

Figure 4 shows the effect of the choice of temporal kernel function on the inferred form of the triggering function  $g$ . The kernel functions used are those shown in Figure 1. In all cases, the triggering intensity is greatest immediately following a burglary, decreasing over the course of 3 days. However, the kernels lead to different estimates for  $g$ , with the one-sided variant being less smooth and the reflected variant having greater density within the first day following a burglary.



**Figure 4** The effect of kernel function selection on the temporal component of the triggering intensity  $g$ , computed using Chicago burglary data from 1/1/2010 to 1/7/2010. The two plots show the same data on different scales.

## 4. Acknowledgements

This work is part of the project Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks ([www.ucl.ac.uk/cpc](http://www.ucl.ac.uk/cpc)), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1).

## 5. Biography

Tao Cheng is a Professor in GeoInformatics and Director of the SpaceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimeLab>) at University College London. Her research interests span network complexity, geocomputation, integrated spatio-temporal analytics and big data mining



(modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

Gabriel Rosser is a research associate in the SpaceTimeLab working on the Crime, Policing and Citizenship (CPC) project. His research interests include GIScience, probabilistic modelling, predictive algorithms and machine learning.

## References

Bowers KJ, Johnson SD and Pease K, 2004. Prospective hot-spotting: The future of crime mapping? *British Journal of Criminology*, 44:641-658.

Chainey S, Reid S and Stuart N, 2002. When is a hotspot a hotspot? In *Socio-Economic Applications of Geographic Information Science*, CRC Press.

Daley D and Vere-Jones D, 2003. An Introduction to the Theory of Point Processes (2nd ed.), Springer, New York.

Diggle PJ, Chetwynd AG, Haggkvist R and Morris SE, 1995. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research*, 4:124.

Johnson SD and Bowers KJ, 2004. The burglary as clue to the future: The beginnings of prospective hot-spotting. *European Journal of Criminology*, 1:237.

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP and Tita GE, 2011. Self-exciting point process modelling of crime. *Journal of the American Statistical Association*, 106(493):100-108.

Youstin TJ, Nobles MR, Ward JT and Cook CL, 2011. Assessing the generalizability of the near repeat phenomenon. *Criminal Justice and Behaviour*, 38:1042.

Zhuang J, Ogata Y and Vere-Jones D, 2002. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369-380.

Zhuang J, 2006. Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society B*, 68(4):635-653.

# Constrained clustering of the precipitation regime in Greece

Eftychia Rousi<sup>\*1</sup>, Christina Anagnostopoulou<sup>†1</sup>,  
Angelos Mimis<sup>‡2</sup> and Marianthi Stamou<sup>§2</sup>

<sup>1</sup>Department of Meteorology and Climatology, Aristotle University of Thessaloniki, Greece

<sup>2</sup>Department of Economic and Regional Development, Panteion University of Athens, Greece

November 4, 2014

## Summary

The aim of this paper is an objective clustering of the precipitation regime in Greece. The data consists of winter daily precipitation values obtained from a Regional Climate Model, the RACMO2/KNMI, for the period 1971-2000. The constrained clustering method is implemented by using three different linkages, single, complete and average, and for three different cluster numbers, 10, 20 and 30. Average and complete linkage both performed well, with the latter proving to be more detailed and its spatial resolution presents many similarities to the original data. The 20 and 30 clusters are clearly more representative than the 10 cluster results.

**KEYWORDS:** Constrained Clustering, Precipitation, Greece, Regional Climate Model.

## 1. Introduction

The problem of defining the various climate zones has a wide range of uses (Iyigun et al., 2013). These include the redefinition of climate zones and rainfall regimes as a result of ongoing climate changes while at the same time examining the reasons that lead to those changes. Also these have a direct effect to hydrology and flora. So the regional water management as well the farming strategies are affected.

In this context, the famous classification system of Köppen – Geiger has emerged, which was originally published by Köppen in 1918. This system provides a set of rules applied to variables derived from long term values for temperature and precipitation. In these, with several rules at hand, various locations are classified into climate types (Cannon, 2012). This rule based approach has been adopted and extended by various researchers, as for example by Thornthwaite who by following manual classifications projected the various locations into climate regions which exhibit climate homogeneity.

With the widespread use of personal computers a different approach has emerged. In this, climate classification is performed by clustering algorithms based on the assumption that areas with similar values of variables characterizing climate, such as temperature or precipitation, can be classified in the same climate type. In this way, the climate types are directly defined by the data. In this methodology, usually a two step approach is adopted. Firstly, a principal component analysis (PCA), followed by clustering analysis (CA) (Fovell and Fovell, 1993; Cannon, 2012).

Those approaches treat the spatial problem of climate zones in an aspatial way, meaning that an area

---

\* erousi@geo.auth.gr

† chanag@geo.auth.gr

‡ mimis@panteion.gr

§ marianthi.stamou@panteion.gr

is included into a class regardless of its location. As a result, the existence of small patches of different classes within regions of a specific climate type that can be seen in Fovell and Fovell (1993) and an attempt to treat this was proposed by Pawitanand Huang (2003).

In our approach, we are extending the preliminary study of Pawitan and Huang (2003) by applying a robust constrained hierarchical algorithm and by examining the validity of the results in the wider area of Greece. Further we test the methodology in a regular grid of points and most importantly, as Fovell and Fovell (1993) encourages, data are not limited to the main land but climate characteristics in the Mediterranean Sea are also considered.

## 2. Materials and Methods

The data used in this study are provided by the Regional Climate Model KNMI-RACMO2 of the Royal Netherlands Meteorological Institute (van Meijgaard et al., 2008). The model has a spatial resolution of 25km x 25km and it is driven by the General Circulation Model ECHAM5/MPI. A broader area around Greece has been chosen, composed by 1064 grid points. The data consist of winter daily precipitation values over Greece for the 30 year period 1971-2000 (Figure 1).

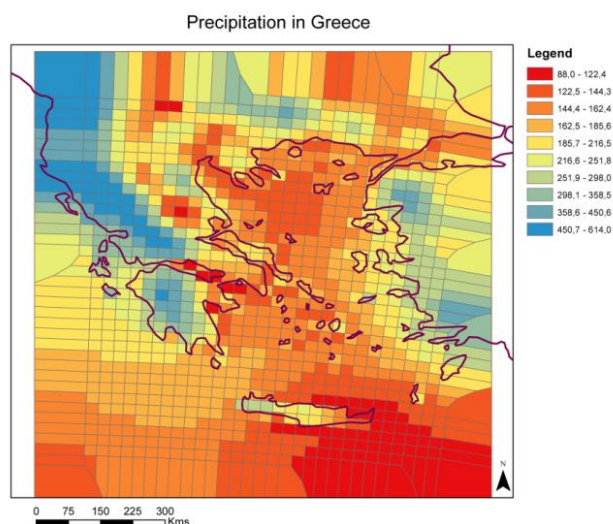


Figure 1. Choropleth map of winter daily precipitation values (in mm).

In our case a ‘flavor’ of agglomerative hierarchical clustering is used (Murtagh 1985). In this method, at the beginning, each area is considered as a region and at each step (iteration) the closest regions, in terms of a given metric, are merged till only one region is left. Different techniques can be devised by following different strategies to define the new distance between the newly merged and the other regions. So, by merging region  $R_i$  and  $R_j$ , the newly defined region  $R_i \cup R_j$  will have distance in respect with the other regions  $R_k$  given by the Lance and Williams formula (1967):

$$d(R_i \cup R_j, R_k) = a_i \cdot d(R_i, R_k) + a_j \cdot d(R_j, R_k) + b \cdot d(R_i, R_j) + c \cdot |d(R_i, R_k) - d(R_j, R_k)| \quad (1)$$

where  $d(R_i, R_j)$  is the distance between regions  $R_i$  and  $R_j$  in terms of the characteristics of each region which are precipitation values in our case. Also  $a_i$ ,  $a_j$ ,  $b$  and  $c$  are parameters whose values depend on the method (Gordon 1996). For example for  $a_i=a_j=1/2$ ,  $b=0$  and  $c=-1/2$  you have single linkage.

Having described the hierarchical clustering techniques, we have to incorporate explicitly the spatial contiguity constraint. For that purpose a “Sorted Dictionary” structure has been adopted where we only keep the pair of regions that are contiguous and are sorted by their relative distance. Following this approach, at each step of the algorithm, we know the pair of regions having the minimum distance while satisfying the contiguity constraint. After removing from the structure the pair with minimum distance, a new region is created by merging them and the structure is updated accordingly (by using Equation 1).

In our implementation, the single, average and complete linkages are used by adopting the Lance and Williams formula. It has been developed in Python 2.7 has been imported as a script in ArcGIS.

### 3. Results and Discussion

The maps of the hierarchical constrained clustering method on the data are presented in Figures 2, 3 and 4, where winter daily precipitation values in Greece are illustrated for 10, 20 and 30 clusters. The average precipitation (in mm) of all grid points grouped in each cluster is shown in the legend of each figure. The daily precipitation simulated values are point data and in order to use them in the analysis, a Dirichlet tessellation was applied to define the neighbourhood structure.

Figure 2 presents the maps for 10, 20 and 30 clusters based on single linkage constrained clustering. It is obvious that this kind of linkage does not provide representative results for the precipitation regime in Greece.

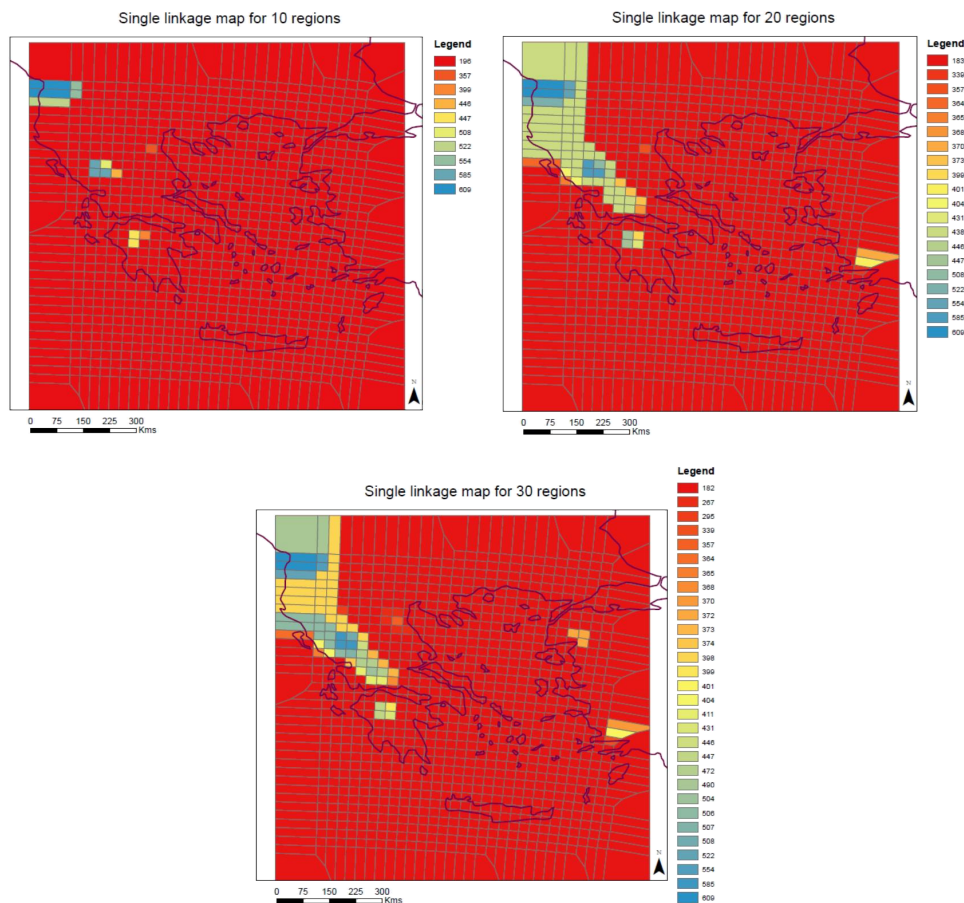


Figure 2. Single linkage constrained clustering.

Results of the average linkage for 10, 20 and 30 clusters are presented in the three maps of Figure 3. This linkage provides a better classification of the precipitation in Greece, especially for 20 and 30 clusters. In these corresponding two maps, the methodology captures not only the high precipitation amounts in the windward areas of Pindus mountains, but it also distinguishes the Olympus peak, the eastern Macedonia-western Thrace region and the increase of precipitation in the eastern Aegean and Asia Minor (Metaxas and Kallos, 1980).

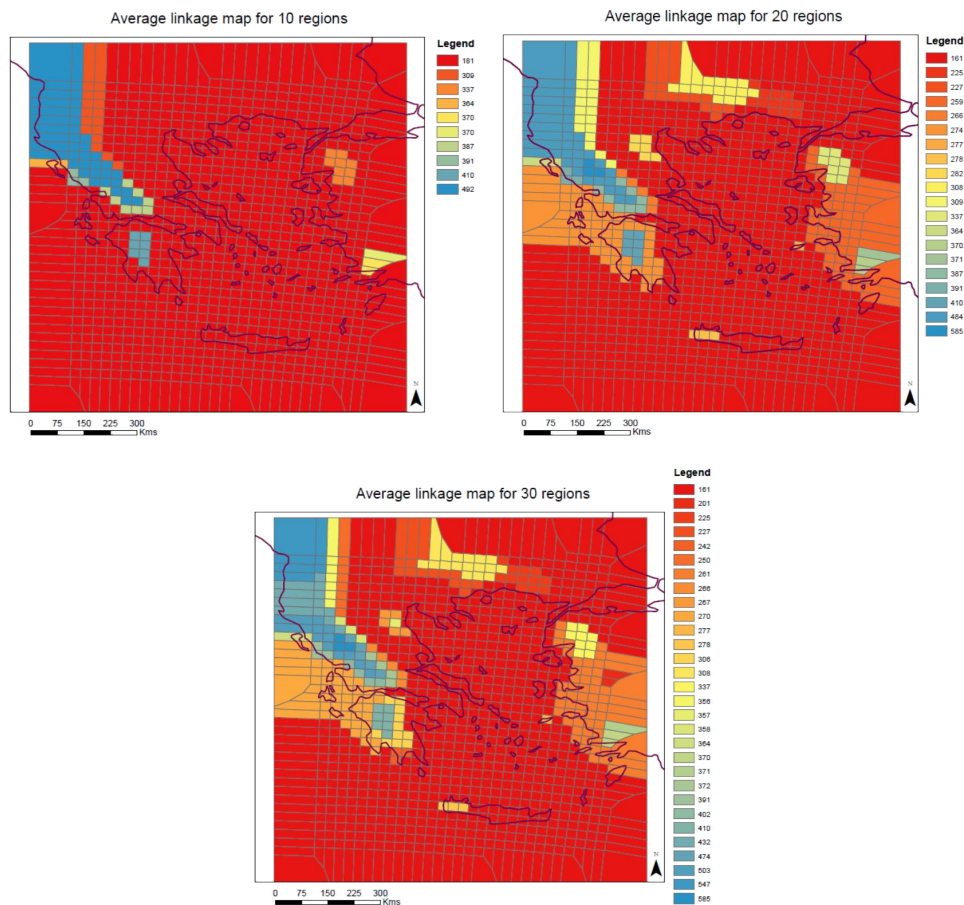


Figure 3. Average linkage constrained clustering.

Finally, the complete linkage was tested and the results are shown in the maps of Figure 4. This method seems to provide greater resemblance to the spatial distribution of the precipitation in Greece (Hatzianastassiou et al., 2008). In this case, even the 10 cluster solution gives a good image, although, 20 and 30 clusters are clearly more representative.



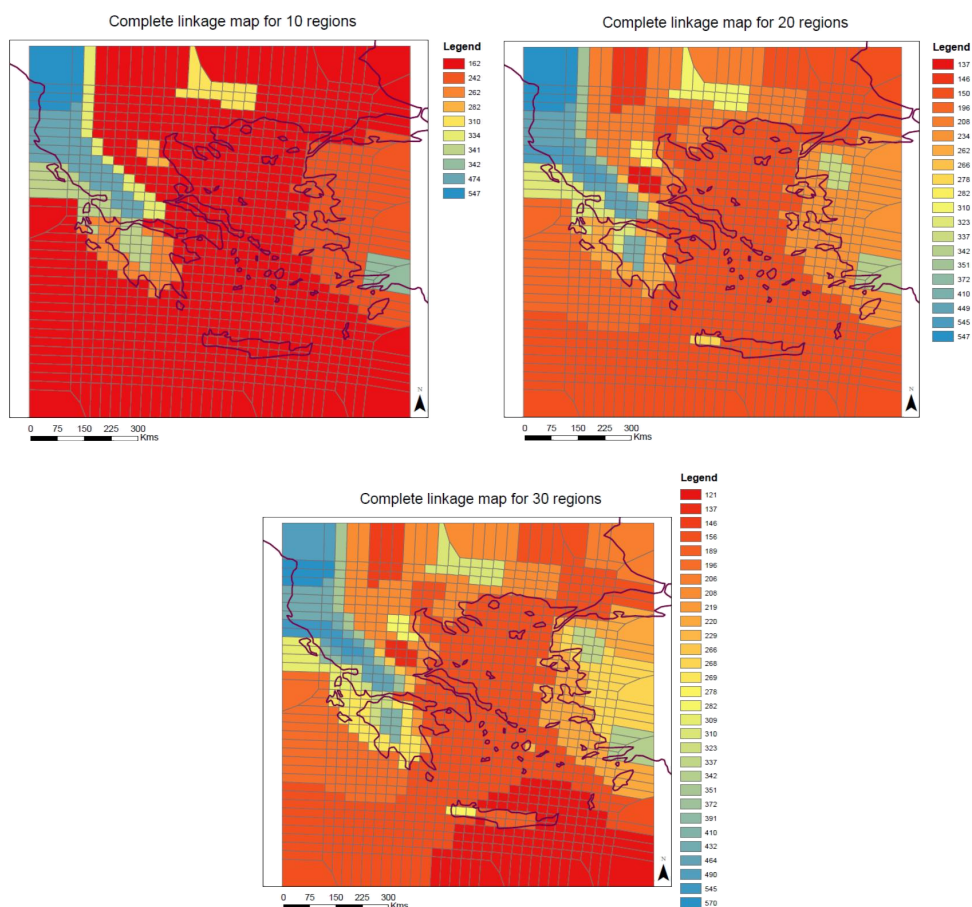


Figure 4. Complete linkage constrained clustering.

#### 4. Conclusions

A hierarchical constrained clustering was performed in order to classify winter precipitation in Greece. Of the three different linkage methods tested, the single linkage was the one that performed poorly. Average and complete linkage both performed well, with the latter proving to be more detailed and its spatial resolution presents many similarities to the original data. Regarding the number of clusters, the 10 cluster solution is clearly not suitable for the precipitation data since it does not capture any of the data variability. On the contrary, 20 clusters could be considered good while 30 clusters do not add any value to the classification results. Future work could test the hierarchical constrained clustering on a combination of meteorological parameters that affect Greek climate.

#### 5. Biography

Efi Rousi is a geographer with a Phd in Climatology. Her research interests include climate change, climate modelling, dynamic and synoptic climatology.

Christina Anagnostopoulou is an Assistant Professor of Climatology at the Department of Meteorology and Climatology, Aristotle University of Thessaloniki, Thessaloniki, Greece. Her research interests include synoptic and dynamic climatology, extreme weather events, methods of climate analysis, climate change.

Angelos Mimis is an assistant professor of spatial analysis in Panteion University of Athens, Greece. His interests include GIS, spatial analysis, computational geometry and optimization.

Marianthi Stamou is a PhD candidate in the Department of Economic and Regional Development, Panteion University of Athens, Greece. Her research interests include spatial econometric models and GIS.

## References

- Cannon A J (2012). Köppen versus the computer: comparing Köppen-Geiger and multivariate regression tree climate classifications in terms of climate homogeneity. *Hydrology and Earth Systems Science*, 16, 217–229.
- Fovell R G and Fovell M-Y C (1993). Climate zones of the conterminous United States Defined using cluster analysis. *Journal of Climate*, 6, 2103-2135.
- Gordon AD (1996). A survey of constrained clustering. *Computational Statistics and Data Analysis*, 21, 17-29.
- Hatzianastassiou N Katsoulis B Pnevmatikos J and Antakis V (2008). Spatial and temporal variation of precipitation in Greece and surrounding regions based on global precipitation climatology project data. *Journal of Climate*, 21, 1349-1370.
- Iyigun C Turkes M Batmaz I Yozgatligil C Purutcuoglu V Koc E K and Ozturk M Z (2013). Clustering current climate regions of Turkey by using a multivariate statistical model. *Theoretical and Applied Climatology*, 114(1-2), 95-106.
- Lance G N and Williams W T (1967). A general theory of classificatory sorting strategies, I. Hierarchical Systems. *The Computer Journal*, 9, 373-380.
- Metaxas D A Kallos G (1980). High rainfall amounts over western Greek mainland during December and January. In Proceedings of the 1st Hellenic–British Climatological Meeting, Hellenic Meteorological Society, Athens, 5–11 September, 125–137.
- Murtagh F (1985). A Survey of Algorithms for Contiguity-constrained Clustering and Related Problems. *The Computer Journal*, 28(1), 82-88.
- Pawitan Y and Huang J (2003). Constrained clustering of irregularly sampled spatial data. *Journal of Statistical Computation and Simulation*, 73(12), 853-865.
- van Meijgaard E van Uft L van den Berg W Bosveld F van den Hurk B Lenderink G and Siebesma A (2008). *The KNMI regional atmospheric climate model RACMO version 2.1*. Tech. Rep. TR-302, Royal Netherlands Meteorological Institute.

# Accessibility-based simulation of urban expansion in Brazil

M. Saraiva<sup>\*1,2</sup>, J. Barros<sup>†1</sup> and M. Polidori<sup>‡2</sup>

<sup>1</sup>Geography, Environment and Development Studies – Birkbeck, University of London

<sup>2</sup>Faculdade de Arquitetura e Urbanismo – UFPel, Brazil

November 7, 2014

## Summary

This work proposes an urban growth simulation model based on a weighted accessibility measure, which is calculated based on the characteristics of the landscape surrounding the city. Two types of features were considered: natural and human-made. The model was tested for the city of Bagé, in the Southern Brazil. The model was largely able to replicate the urban expansion pattern of the case study, suggesting that the proposed weighted accessibility measure is a suitable way to capture the impact of surrounding areas in the process of urban expansion.

**KEYWORDS:** urban growth, accessibility, natural environment, transport system, urban modelling.

## 1. Introduction

Accessibility measures have been largely used in urban studies, mainly for urban transportation (Ritsema van Eck & Geurs, 2001), and natural environment studies (Bunn et al., 2000). The present study proposes a simple model for simulating urban expansion based on accessibility. It assumes that the relationship between an urban area and its surroundings (natural and human-made features) is an important aspect of urban growth. The relationship between natural environment and urbanization has been demonstrated by Alberti et al (2003) who have called for more studies integrating urban and natural environments. This study assumes that urban expansion is partially influenced by the characteristics of the surrounding area, which may attract urban development (river, beach or proximity to existing transport network) or offer resistance to urbanization (flooding places or lack of infrastructure). In the present work, the proximity to such features is quantified using an accessibility measure.

## 2. Method

In this study, accessibility is used to consider the characteristics of the natural and urban environment as a single spatial differentiation measure. Urban expansion is then simulated assuming that areas with higher accessibilities are more likely to be urbanized. In the model the rate of urban expansion is an exogenous variable, which can be defined by the user. The same applies for the amount of randomness in the selection of areas to convert process. The model is recursive and accessibility is recalculated every iteration taking into consideration the newly converted urban areas.

Accessibility can be defined as the property of a particular element to be closest to all other elements of the network, considering minimal (or preferred) paths between them (Ingram, 1971). For this work, the measure will be calculated in a grid space. Mathematically, the cellular accessibility described here is defined in Equation 1.

---

<sup>\*</sup> m.saraiva@mail.bbk.ac.uk

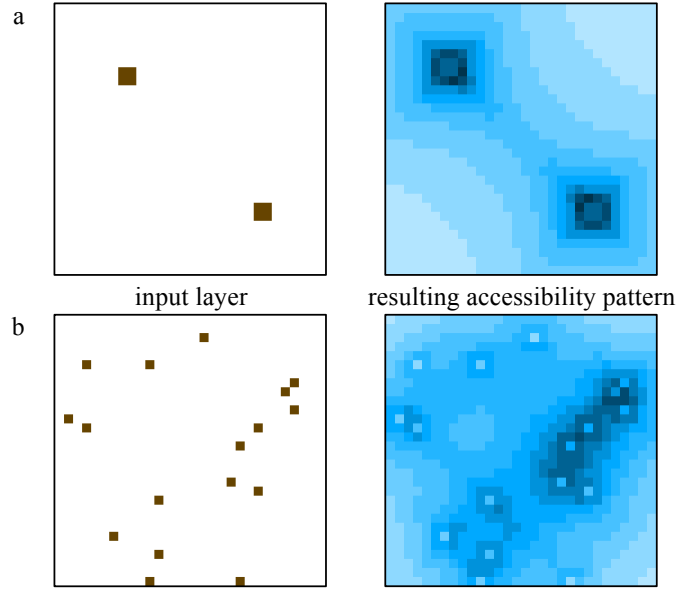
<sup>†</sup> j.barros@bbk.ac.uk

<sup>‡</sup> mauricio.polidori@gmail.com



$$AC_i = \sum (1/d_{ij}) \quad (1)$$

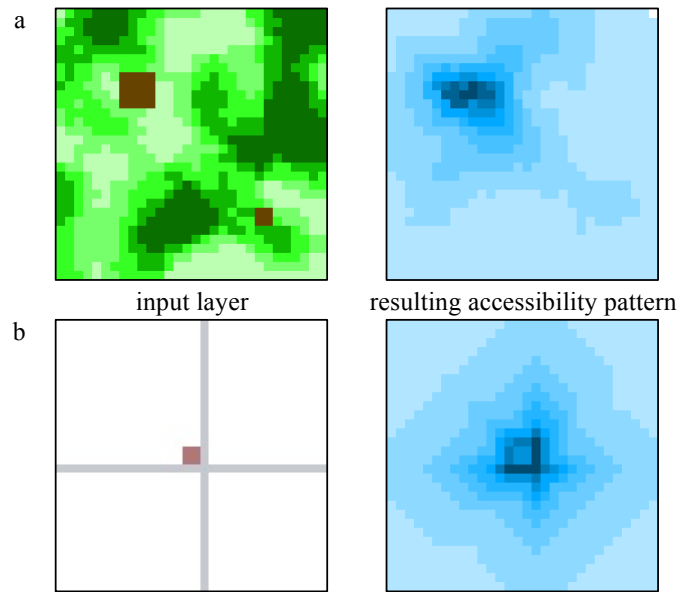
That can be read as: the cellular accessibility of cell  $i$  equals to the sum of the inverse of the distances between cells  $i$  and  $j$ , for each  $j$  cell that has attributes of attraction to urbanization. Figure 1 shows the resulting accessibility patterns for two scenarios: two urban areas, and multiple urban cells scattered by the grid.



**Figure 1:** (a) accessibility pattern (represented in tones of blue) was calculated between the two urban areas (in brown); (b) accessibility pattern based on multiple urban cells.

The measure of accessibility used in the model extends the one above by including different aspects of the space surrounding urban areas, including features of the natural environment (such as topography, swamps, lakes and streams) and human-made features such as the transportation system (roads, railroads and waterways) as well as land-use zoning policies. These are included by using a weighted accessibility measure, in which each cell receives a different weight based on the features it contains. Different weights are given to natural and human-made features.

The effects of environmental and human factors weighting on accessibility are shown in Figure 2, below.



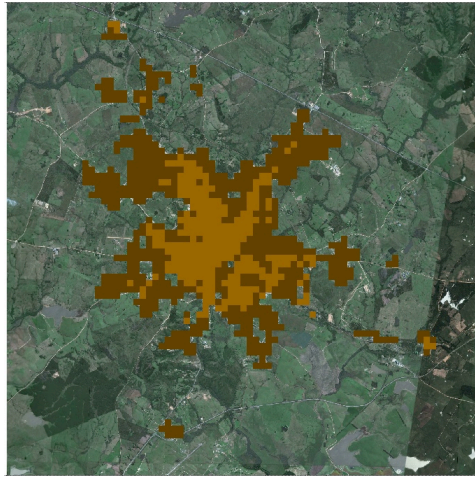
**Figure 2:** effects of weighting in cellular accessibility. (a) urban cells in brown and the intensity of environmental resistance in tones of green. (b) urban cells in brown and roads in grey.

As shown in Figure 2, the weighted cellular accessibility measure is able to capture the structure of the system, presenting variation in response to changes in the input data.

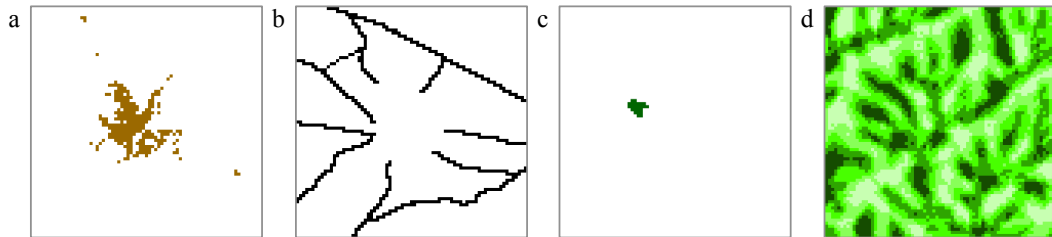
Mathematically, the cellular accessibility measure represents each cell's probability of ground conversion. The actual cells to be converted are selected randomly among the most accessible cells based on a random variable and urban expansion rate, both user-defined parameters.

### 3. Case study

The model was applied to the city of Bagé, a medium-sized city in the south of Brazil with 116.794 inhabitants. The study area was represented using a grid of square cells with a resolution of 200m. Figure 3 shows the city in the years 1974 and 2012, which have been set as the start and end times of the simulation. During the 38 years period, the city grew 326%, at an average rate of 3,2% per year. The input data used in the simulation is shown in Figure 4, and is composed by: a) the urban area in 1974; b) current road system; c) a military zone (not available for urbanisation); d) a grid derived from topography and hydrology via interpolation, assuming that better drained areas attract urbanization.

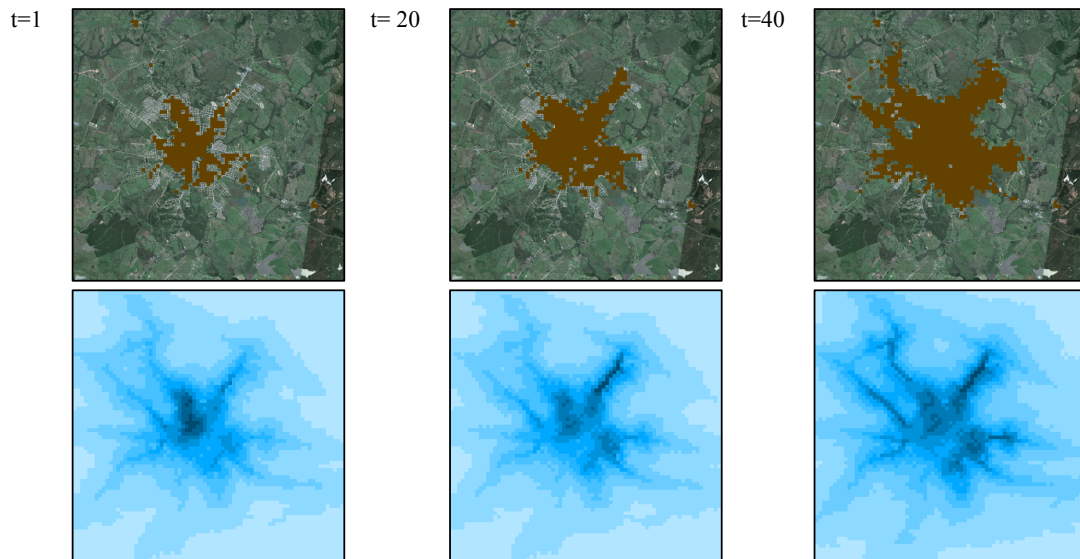


**Figure 3:** urbanized area of Bagé in 1974 (light brown, 262 cells) and 2012 (dark brown, 855 cells).



**Figure 4:** Input data for the model: (a) urban area in 1974, (b) road system, (c) military field, and (d) topography and hydrology grid, with areas more suitable to urbanization in light green.

Figure 5 shows the outputs of the urbanized area and cellular accessibility in three partial states of the simulation, representing iterations (t) 1, 20 and 40. For this simulation the following parameters were used: randomness = 50%, urban growth rate=3.2%.

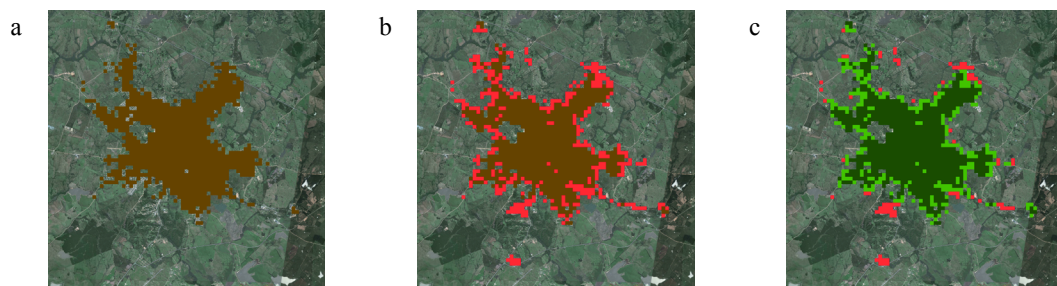


**Figure 5:** Outputs of the simulation, showing the urbanized area in brown and the corresponding weighted accessibility in tones of blue.

Results show visual similarities between the real and the simulated cities. The city of Bagé has an urban form predominantly compact in the centre and axial towards the city's borders and both of these characteristics could be replicated by the model. The accessibility distribution also shows a strong axial pattern in this scenario. The road system influence heightens the accessibility of distant places, facilitating the land conversion of these cells and generating the axial form of the city.

Figure 6 shows the comparison between the final state of the simulation and the city of Bagé in 2012. The image overlaps the satellite image of the city with the model's output. In order to aid comparison, the figure shows three versions of the model's output. On the first (Figure 6a), it shows the simulated urbanised area in brown, on the middle image (Figure 6b), the cells in which there is a match between simulated and real urban areas are represented in brown and the cells where there is no match are represented in red; in the last image (Figure 6c), matching cells are represented in dark green, non-matching cells in red and 'near-miss' cells in light green. Near-miss cells are a result of the calculation of fuzzy similarity between the two maps following Hagen's (2002) approach.

When analysing the results in Figure 6b, it is clear the non-matching cells (in red) are mainly concentrated in the borders of the city. On Figure 6c, which was calculated with fuzzy similarity, shows that the majority (76%) of the non-matching simulated cells (in red in Figure 6b) are located next to urbanised cells in reality. These are represented in light green in Figure 6c and indicate the formal structure of the city has been successfully captured.



**Figure 6:** a) urbanized area in brown; b) correct cells in brown and wrong cells in red; c) correct cells in dark green, wrong cells in red and near-miss cells in light green.

#### 4. Conclusions

The model was able to largely replicate the urban expansion pattern of the city of Bagé for a period of 38 years (from 1974 to 2012). Thus, the weighted accessibility measure seems to be a suitable way to capture the impact of surrounding areas in urban expansion, integrating natural and human factors. However, in the case study, small settlements located in the outskirts of the city were not captured by the model. This is due to the pattern of the accessibility measure, which values are higher close to the urban areas and decrease slowly towards the borders of the work area. The next step of this research is to attempt to mitigate this effect.

#### 5. References

- Alberti, M.; Marzluff, J. M.; Schulenberger, E.; Bradley, G.; Ryan, C.; Zumbrunnen, C. (2003) Integrating humans into ecology: opportunities and challenges for studying urban ecosystems. **BioScience**, 53(12). 1169-1179.
- Bunn, A. G.; Urban, D. L.; Keitt, T. H. (2000) Landscape connectivity: a conservation application of graph theory. **Journal of Environmental Management**. 59. 265-278.
- Geurs, K.; Ritsema Van Eck, J. (2001) Accessibility measures: review and applications. **RIVM Report n. 408505 006**. Bilthoven, Holanda.

Hagen, A. (2002) Multi-method assessment of map similarity. **5th AGILE Conference on Geographic Information Science**.

Ingram, D. R. (1971) The concept of accessibility: a search for an operational form. **Regional Studies**, 5. 101-107.

Liu, J.; Dietz, T.; Carpenter, S.; Alberti, M.; Folke, C.; Moran, E.; Pell, A.; Deadman, P.; Kratz, T.; Lubchenco, J.; Ostrom, E. (2007) Complexity of coupled human and natural systems. **Science**, 317.

Nystuen, J. D. (1968) Identification of some fundamental spatial concepts, in Berry, B. J.; Marble, D. F. **Spatial Analysis: a reader in statistical geography**. New Jersey: Prentice Hall.

## **6. Acknowledgements**

This work was financially supported by CAPES Foundation, Ministry of Education of Brazil, in the form of a PhD scholarship provided to the author, process number BEX 0960/14-4.

## **7. Biography**

Marcus Saraiva is a PhD student interested in computational models of urban dynamics, more specifically cellular automata models extended with techniques of graph theory and artificial intelligence.

Joana Barros is a lecturer in GI Science at Birkbeck, University of London and Marcus' supervisor. Her areas of expertise are urban planning and modelling, more specifically agent-based and cellular automata models applied to urban systems and urbanisation in developing countries.

Mauricio Polidori is an academic in the Federal University of Pelotas / Brazil. His interests include: urban and regional planning, urban simulation models, urban morphology, and the natural environment.

# Group Behaviour Analysis of London Foot Patrol Police

Jianan Shen, Tao Cheng

SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental and Geomatic  
Engineering, University College London,  
Gower Street, London WC1E 6 BT, UK  
Nov 03, 2014

## Summary

The main objective of this research is to propose a method for group movement pattern generalisation and classification. To this end, DBSCAN is used on stay points for POI identification. Then, movement features are extracted and selected for the behaviour classification. A kernel-based Support Vector Machine method is developed to infer the working types of the officers based on the selected features depicting their movement histories. By analysing the geo-tagged police data, we demonstrate how this method can be used to reveal user information, especially interest information based on their POIs and spatial-temporal movement patterns.

**KEYWORDS:** Policing, Travel Pattern, Machine Learning, SVM Classification, POI

## 1. Introduction

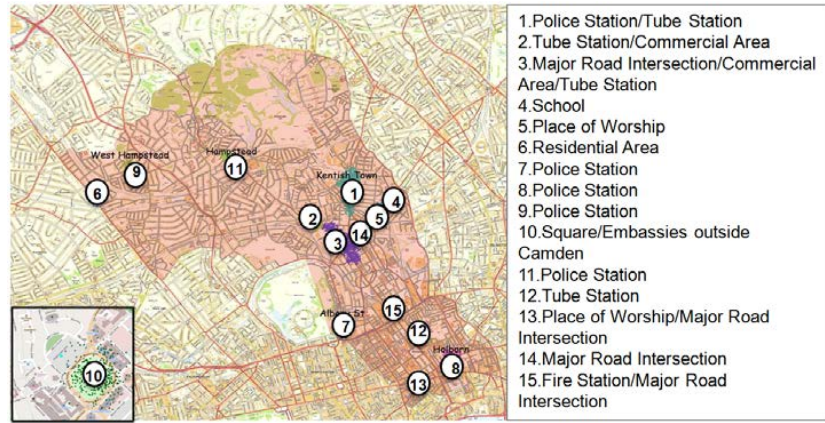
The more and more ubiquitously used GPS-integrated devices have generated increasingly large movement history data than ever before. Such movement history data have enabled researchers to analyse and visualise movement trajectories, home range and POI (Points of Interests) distribution, and mobility patterns. Police foot patrol activity, as a special kind of human movement phenomenon, can also be analysed in similar methods with certain modifications.

In this research, we intend to present a framework of patrol pattern analysis and classification of police working type, which is capable of 1) extracting police POIs and traveling sequences from GPS trajectory data; 2) selecting and summarising individual and group movement features for description and comparison of officers undertaking different works; 3) SVM classification of different types of officers based on the selected moving history features.

## 2. Case Study and Data

The case study takes place in the Camden Borough of London (Figure 1). Five major police stations are located (Places No. 1, 7, 8, 9 and 11 in Figure 1) in this region, namely, West Hampstead, Hampstead, Kentish Town, Albany Street and Holborn.

Automatic Personnel Location Systems (APLS) data set provided by the Metropolitan Police recorded officers' location stamps collected by GPS-integrated radio sets portable on every officer in patrol. The length of the data is 84 days. The 241525 records in the dataset recorded information such as call signs, device IDs, as well as the locations and times of all 745 officers active in that period. Usually, the sampling rate of the system is one update every 10 minutes.

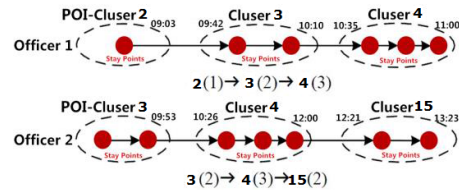


**Figure 1** The POIs identified in Camden, including 10 non-police-station POIs (1 of them locates outside Camden) and 5 police stations.

### 3. Methodology and Results

#### 3.1. POIs and Traveling Sequences Extraction

In discovering POIs, the location and time of officers' stopping behaviour is of more interests than the moving factors (Palma, 2008). To this end, DBSCAN is used for the clustering of stay points where officers stop or move slowly for more than 20 minutes. The massive activities and stops observed nearby police stations are common reflection of police daily routine and office works, and hence disturbed the searching for the POIs during the patrol activities. Therefore, outliers and records generated within 400m radius from the police stations are removed before the density based clustering. Only trajectory moving out of the radius will be considered as a start of a journey and when the time interval between two records exceed 2 hours, the series will be separated as two trips. By adjusting the parameters according to the quality of the data set and the method suggested by Zhou (2012), 10 non-police-station POIs are discovered. For instance, POI No. 2 represents a tube station, POI No.3 is considered to be a major road intersection.

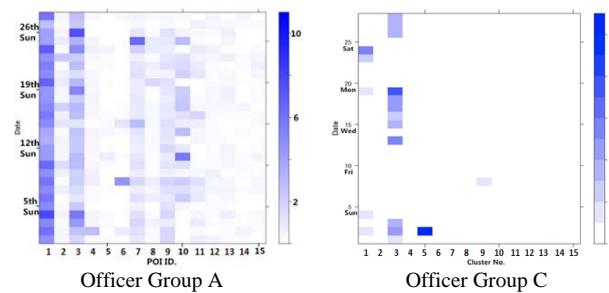


**Figure 2** Individual movement history expressed by sequences of visited POIs

By identifying building information on Ordinance Survey Maps and communicating with the police. The main topics of these POIs are identified as schools, places of worships, tube stations and major road intersections. The movement histories of the officers can then be simplified and expressed by a sequence of common POIs the officers visited and the time they arrive and leave each POI attached in the sequence (Figure 2). Similarly, it can also be expressed as a series of topics an officer went through during his/her patrol journey. For example, the journey of Officer 2 in Figure 2 can be summarised as "Commercial Area> School > Fire Station" and the attached time information about how long s/he stayed in each POI. The information of time series, POI topics and trajectories will be further processed to support the classification works.

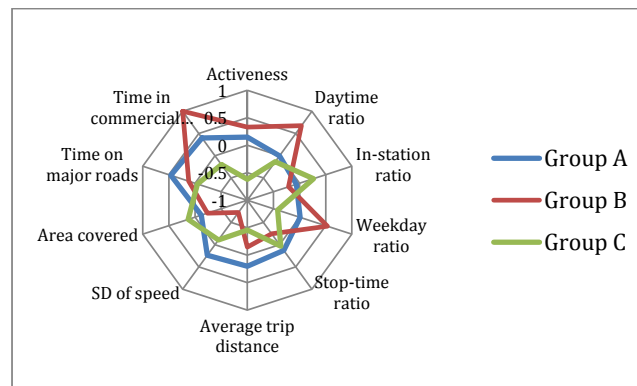
#### 3.2 Individual and Group Movement Features

After acquiring the semantic meaning of discovered POIs in previous step, individual and group movement can be depicted and summarised with 14 features that are selected or extracted from the raw GPS log as well as the summarised POI visiting history. The features include: mean and standard deviation of speed, mean and standard deviation of patrol covering area, proportion of stopping time, proportion of time spent in police stations and POIs of different topic types, proportion of time spent in weekends and nights and the activeness expressed by the number of records generated by the officer in unit time and so on.



**Figure 3** The comparison of POI visiting frequency between two groups consisting of two different types of officers in one month

Figure 4 shows 10 example features of the 3 groups. The values of features are z-score normalised for further classification works. It can be seen in the chart that the selected features of the 3 groups vary greatly. For example Group B has strong daytime activeness proportion and seldom work at night while the standard deviation of Group A is much larger than other groups.



**Figure 4** Radar chart comparison of 10 example features depicting the patrol behaviour of different groups

### 3.3. SVM Classification of Officer works using RBF Kernel

Many researches proposed SVM and other machine learning methods analysing online user behaviors. Some researchers developed similar methods on the analysis of travelling behaviour in real life, such as Baraglia (2013).

By using the parameter selection method developed by Schölkopf (2002), RBF kernel is used in the SVM classification of officer working types based on the individual movement features extracted in previous section. The model is trained with officer movement data of which the working types have already been labeled and is used to classify the data generated in 10 days of the next month. The accuracies of this attempt are logged in Table 1.



Table 1 Classification accuracy by training sets of different sizes

Size of training sets	Training Error	Accuracy of classification
10 days	13.52%	68.89%
20 days	15.00%	72.34%
30 days	14.83%	73.42%

This classification is based on manually-selected and unoptimised feature sets. This method will be further improved with optimum feature selection and using more advanced kernels.

#### 4. Summary and future work

In this research, a method for identifying user (officer) type based on travel (patrol) history is proposed. The framework include density-based POI clustering, feature selection and validation, as well as machine learning classification. The Camden APLS data enabled the study that others cannot proceed with due to the lack of modern GPS-enabled policing equipments. The method revealed the movement features of different officers in space and in time. It can also be used in other time series geo-tagged data for automatic movement pattern generalisation, traveler interest and routine mining, similarity analysis and so on.

Further works may include improving the performance of SVM in unbalanced data sets (the officer numbers in different types vary greatly) and comparing SVM with other methods such as KNN and Artificial Neural Networks.

#### 5. Acknowledgements

This work is part of the project - Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks ([www.ucl.ac.uk/cpc](http://www.ucl.ac.uk/cpc)), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) is highly appreciated.

#### 6. Biography

Jianan Shen is a PhD student in UCL SpaceTimeLab. He studied Information Engineering in NUDT in China (BEng, 2013). He is now working on the Crime, Policing and Citizenship Project sponsored by the EPSRC, UK. His research interests include movement pattern and trajectory analysis with Machine Learning approaches.

Tao Cheng is a Professor in GeoInformatics, and Director of SpceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimelab>), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

#### References

- Baraglia R, Muntean C I, Nardini F M and Silvestri F (2013). LearNext: learning to predict tourists movements. *In Proceedings of CIKM*, 751-756.
- Palma A (2008). A Clustering based Approach for Discovering Interesting Places in Trajectories. *Symposium on Applied Computing*, March 16-20, Fortaleza, Brazil. 863-868.

Schölkopf B and Smola A J (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press.

Zhou C (2004). Discovering Personal Gazetteers: An Interactive Clustering Approach. *Geographic Information Science*, 266-273.

# **Are we there yet?**

## **Exploring distance perception in urban environments with mobile Electroencephalography**

Katerina Skroumpelou<sup>1</sup>, Panagiotis Mavros<sup>2</sup> and Andrew Hudson Smith<sup>2</sup>

<sup>1</sup>Computer Networks Lab, School of Electrical and Computer Engineering, National Technical University of Athens

<sup>2</sup> Centre for Advanced Spatial Analysis, The Bartlett, University College London

May 20, 2015

### **Summary**

This paper explores the use of mobile Electroencephalography (EEG) in the study of environmental perception and the ways the perception of physical measurements of a space may affect individual walking behaviour. So far, the factor of an individual's affective state has not been taken into account in perceiving space. The hypothesis of this study is that that people perceive physical measurements differently. The question posed is to what extent distance and route length perceptions are affected by the psychological state, and whether these perceptions play a role on route planning. We propose the use of mobile EEG, a technology that permits such insights, to augment the traditional arsenal of questionnaires and self-reported measures of experience and mental representations of space.

**KEYWORDS:** mobile EEG; distance perception; Emotiv; urban mobility, walkability.

### **1. Introduction**

This paper explores the use of mobile Electroencephalography (EEG) in the study of environmental perception and the ways the perception of physical measurements of a space may affect individual walking behaviour. So far, the factor of an individual's affective state has not been taken into account in perceiving space. We set off with the hypothesis that people perceive physical measurements differently (Tolman, 1948, Pequet, 2002). Existing research has recorded a number of parameters that affect distance perception (Montello, 1997). The question posed is to what extent distance and route length perceptions are affected by the psychological state, and whether these perceptions play a role on route planning. We propose the use of mobile EEG, a technology that permits such insights, to augment the traditional arsenal of questionnaires and self-reported measures of experience and mental representations of space, in order to further our understanding of how distance perception is modulated by the static and dynamic characteristics of the environment.

### **2. Background**

The study of urban mobility behaviour is of great importance in the context of a rapidly urbanising population worldwide, together with the increasing efforts from authorities to encourage active transport, such as walking and cycling, along with the use of public transport (British Department for Transport, 2013), with benefits for urban planning, population health and wellbeing. Individual

perceptions of the environment affect the way we travel in the city, and also our trip-planning choices (Cadwallader, 1976). In this context, studying travel behaviour, route-choice and understanding which factors influence transport mode choice, is an important step towards better cities. As most daily trips are purposive (to work, home, school, shopping etc), one of the primary factors influencing behaviour is the distance between origin and destination. As the geographer Dan Montello notes, “[distance] is used to evaluate costs of traveling from one place to another, and it helps us utilize resources efficiently (time, money, food)” (Montello, 1997). Distance between our present location and our destination, or even the distance between forthcoming destinations in a trip chaining case (e.g. Garling, 1999) is subject to multiple cognitive biases that influence spatial decisions, including individual differences between people (Wolbers and Hegarty, 2010). In the context of urban mobility, perception of distance is an important factor influencing the spatial decision making techniques employed by pedestrians, as it affects the decision whether to walk towards a destination or seek an alternative means of transport.

Existing research on distance perception and walking behaviour has been focused on walkability, route and space complexity and cost functions, based on walking time, ease of navigation and route pleasantness (Saelens et al., 2003; Dewulf et al., 2012; British DfT, 2012). Modal choice procedures are influenced by the evaluation of time, comfort and wait-time annoyance. A less explored factor of distance perception and walking behaviours is the role of the affective state of mind of the pedestrian. Electroencephalography is a method of brain imaging that can provide us with such an insight. For this purpose, this study uses an electroencephalography (EEG) device, measuring brain activity, and a software that interprets this brain activity into emotional states of an individual when walking in different types of urban environments.

This choice of method looks into the subject of distance perception in urban mobility from a new perspective. The role of brain activity and emotions - as measured in real time with an EEG device - has been used before in urban setting experiments. In a study by Aspinall et al. (2013) the Emotiv EPOC headset was used to explore the differences between walking in urban versus natural environments (park). The study presented here, however, uses this technology to explore the effects of an individual's affective state on the perception of walking distance and time. Correlating the findings with the environmental attributes the participants come across aims to give a new insight on our knowledge of the topic.

### **3. Methods**

In order to estimate a distance between two locations, we rely either on external resources (e.g. a map) or our (internal) mental representation of space, also referred as ‘cognitive maps’ (for a discussion see Kitchin and Blades, 2002). In Spatial Cognition research, the acquisition of environmental knowledge from maps, virtual reality and real environments has been extensively studied. Learning modality, individual differences in spatial abilities or cognitive strategies used influence the detail and accuracy of resulting mental representations of space. Distance estimates between locations are a standard procedure to test the quality of mental representations of space, and several researchers have explored how fast and accurately such representations are established (Ishikawa and Montello, 2006) or how different forms of spatial experience (map, virtual reality or real experience) affect them (Richardson et al., 1999; Waller and Greenauer, 2007).

However, it is less clear how the emotional state of the individual affects their perception of travelled distance. To address this questions, a field experiment was conducted in the area of Fitzrovia,

in central London. For this exploratory study, a total of eight participants were asked to walk a designated route (Figure 3.1) that was divided into four segments. Each of the four route segments had different environmental characteristics (main, busy road, backstreet, shopping street, wide pavement, presence of trees).



Figure 3.1 The route in Fitzrovia. On the top is Euston Road, and on the right is Tottenham Court Road. The different colours indicate the four segments.

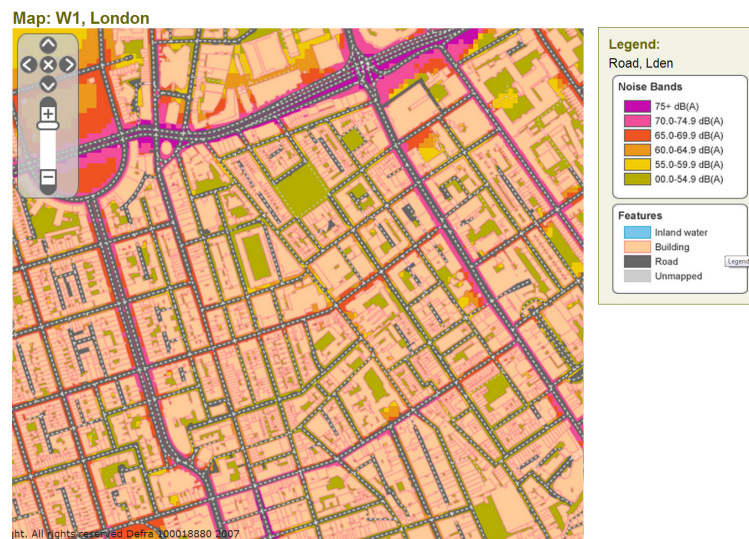


Figure 3.2 The noise levels for the area. (source: defra, 2012)

During the experiment, participants were instructed to walk normally, wearing an Emotiv EEG headset (Figure 3.3) and carrying a GPS recording device. After completing the route, participants were asked to evaluate various aspects of the route through a combination of interview and self-reported questionnaires.



Figure 3.3 The modified version of the Emotiv EEG headset that was used in the experiments, here worn by a participant.

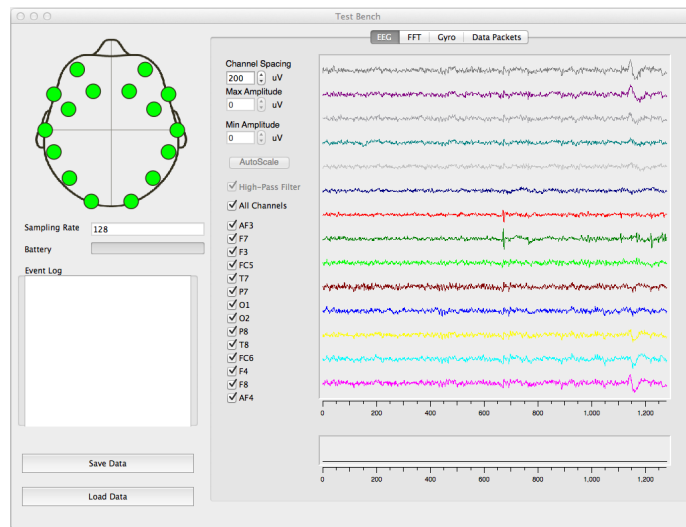


Figure 3.4 A screenshot of the testbench software that comes with Emotiv (shows raw EEG data).

The hypothesis of this research is that EEG can capture the effects of static as well as dynamic aspects of the environment on the individual. However, the brain activity of the individual is influenced by incidental events or situations that occur during the experiment, that are not directly relevant to the urban environment. The experimenter, in this case, has to annotate the route with observations of a number of events that occur during the experiment, like the confrontation of an obstacle (another pedestrian for example), or verbal communication.

In order to geo-annotate and timestamp with precision such events, a custom Android application was developed that enables the experimenter to make these annotations. The app, given the name “Logger” (Figure 3.5), is collecting data about the phone’s (the experimenters/subject’s) location, speed, GPS recording accuracy, the altitude, the current time in unix epoch time and in human readable

format, at a sampling rate of 1Hz. The user can collect behavioural data out of eight “event” options: “Instructions/Talk”, “Pause”, “Obstacle”, “Looks Around”, “Start Walking”, “Hesitates”, “Controlled Crossing”, “Uncontrolled Crossing”. The app is open-source, available online<sup>1</sup> and is currently being used in more studies, contributing to the toolkit for conducting neuro-behavioural research in natural environments.

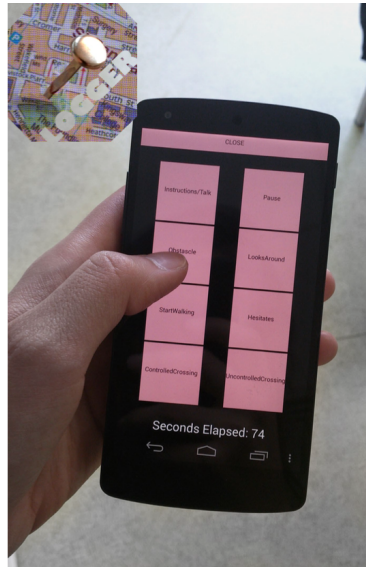


Figure 3.5 The Logger App.



<sup>1</sup><https://github.com/mandarini/Logger>

Figure 3.5 Events as recorded by the LoggerApp

4. Analysis

Analysis of participants’ neural, behavioural and self-reported data from the routes illustrate the potential of our approach. Raw EEG data were analysed with the software *Emotiv Affectiv suite*, which translates the raw EEG signal to four different mood readings: “Excitement”, “Frustration”, “Meditation” and “Engagement/Boredom”. Based on these, the route segments were ranked from most to least “Frustrating”, “Exciting” and “Boring”. Route segments were also ranked by participants according to 7 characteristics: *stressfulness, pleasantness, length, fatigue, scale, light and noise*.

Our Analysis was based on the following variables:

Independent Variables	Dependent Variables
Distance	Mental States
Space Syntax Measures	Excitement
Segment	Frustration
Street scale	Engagement
Street type	Meditation
Land Use	Self-reports
Junction (yes/ no)	Pleasant
Junction Type	Noisy
Location	Long

For each route, we had a series of continuous EEG data for the route (Figure 4.1) and a self-reported ranking for the route four segments. In the self-reports, the participants were asked to rank the four route segments according to a number of characteristics (Figure 4.2). Our intention was to group the experience of the route in these four segments that were each of distinct environmental characteristic. The annotation of the EEG data with the Logger App allowed us to segment the EEG data in these four parts as well, omitting the pauses, the instructions and the talking that took place during the route. For some parts of the analysis, we took the mean of the affective data for each of the route segments.



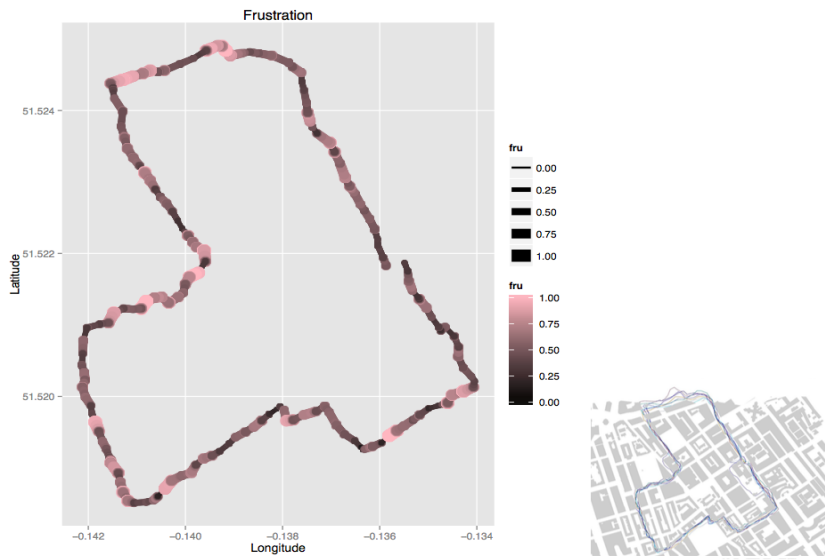


Figure 4.1 Frustration levels of a participant along the route

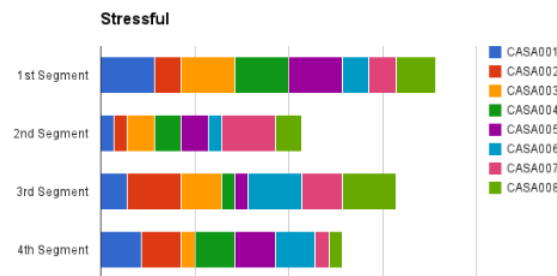


Figure 4.1 The 8 participants rank the four segments according to how stressful they were.

## 5. Results

Distance perception was the main focus of this study. As expected by the literature, participants could report well when asked to rank the segments from shortest to longest. Interestingly, participants were less accurate between the two of the segments that were of different environmental nature. Each one of them had a characteristic that caused participants to consider them unpleasant and frustrating, the one being busy and noisy, and the other having a complex route with turns.

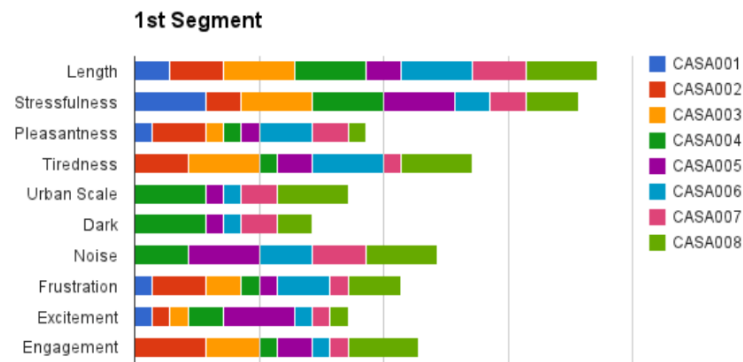


Figure 5.1 The stacked graphs of each characteristic for the first segment. We can see that positive feelings are the lowest here (as recorded by the EEG), the length and stress reports being the highest.

Analysis suggests that the segment that was characterised as tiring, stressful and evoked the least positive feelings (“excitement”) was also perceived as being the longest (Figure 5.1). There was also a trend between a route being characterised as tiring and long and other environmental attributes such as the lack of light and shops. We tested the Interaction between factors and some results have been observed, indicating a relationship between certain environmental parameters and the presence of negative perceptions and feelings. Comparing self-reported with direct measures of affect, the routes that were reported to be the shortest were the ones with the highest “excitement” levels, meaning the ones where positive feelings were dominant.

The Logger App, as we said, allowed us to record the moments where some elements appeared during the route, such as crossings and obstacles. Figure 5.2 illustrates this spatial diagram overlaid by the events as recorded by Logger.

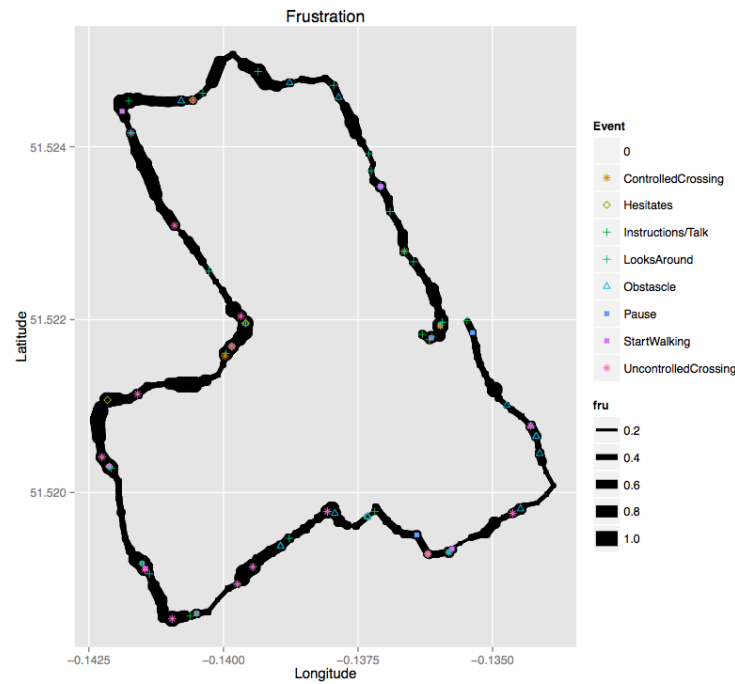


Figure 5.1 Frustration levels of a participant and all the events along the route.

“Frustration” levels were affected significantly by obstacles on the pathway, and influenced the perception of a route as long, tiring and unpleasant. The complexity of the route also played a significant role in characterizing a route tiring, but had little impact on “frustration” levels.

In order to explore the effects of subjective experience and individual incidents, we used a Neuroscience approach named “Event-Related Potential” (ERP) which studies brain activity in response to particular visual, auditory or other stimuli (Picton et al., 2000). In Neuroscience, the ERP method compares the neural signal during two different conditions (e.g. baseline vs stimulus) to determine whether and how a certain brain area is engaged in a particular task. In our research context, we were interested in the effects of a variety of incidental events, such as verbal communication, the confrontation of an obstacle, a busy (or uncontrolled) crossing. By geo-annotating and time-locking these “events” to the EEG data, we can explore how these correspond to different neural activity or emotion state. Two important differences with classic ERP studies and this research is that in a naturalistic experimental paradigm (urban walk) there is a variety of stimuli and, secondly, we have so far compared the emotional state levels (interpreted EEG) and not the raw signal. The frustration of the subjects peaks during an encounter with an obstacle and declines considerably after 30 seconds, close to the initial state before the obstacle. Some peaks of frustration appear also at crossing points, and at points where the environmental situation changes, like in the transitions from a busy street to a quiet street. As illustrated in Figure 5.2, highest values in frustration occur around obstacles and uncontrolled crossings. Around controlled crossings we also have high frustration peaks, with most frustration values being high.

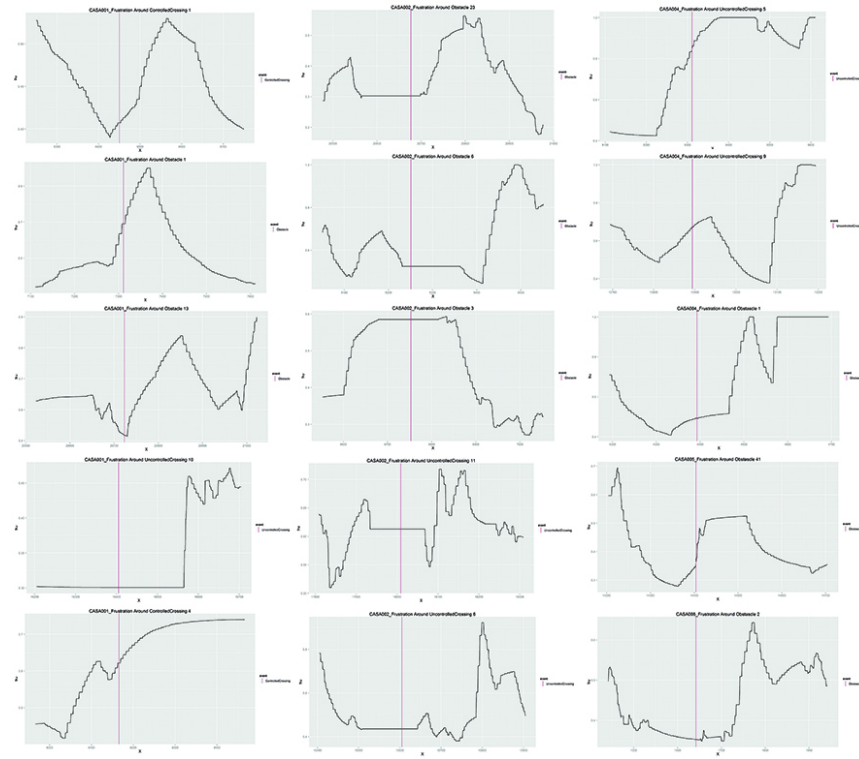


Figure 5.2 Frustration peaks after the encounter with an obstacle (vertical line) or around an uncontrolled crossing.

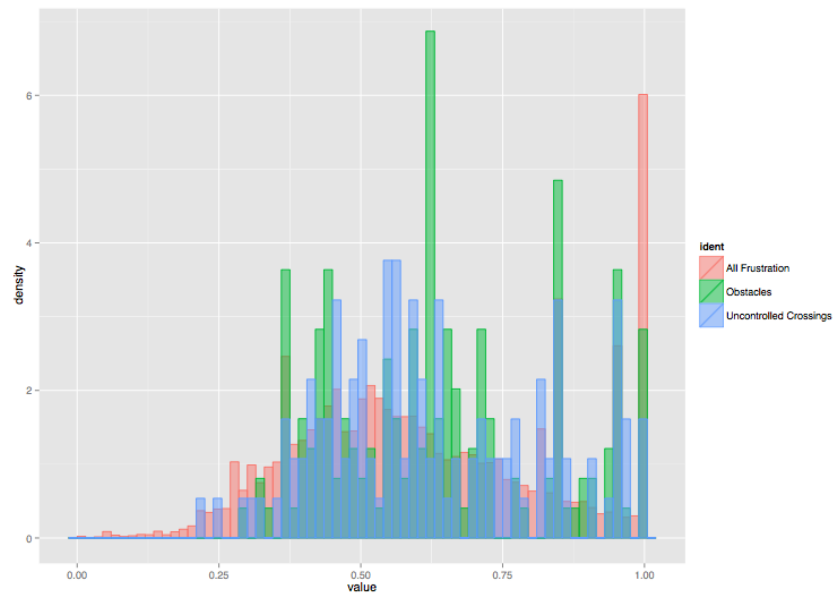


Figure 5.3 Distributions of the means of the recorded frustration values near obstacles and uncontrolled crossings as compared to the mean of all participants' frustration readings. There are more high frustration readings than the mean of general frustration

## **6. Discussion and Future Work**

The results of such studies contribute in our better understanding of the role of urban planning. Handy et al., 2002, insist on evidence that “a combination of urban design, land use patterns, and transportation systems that promotes walking and bicycling will help create active, healthier, and more livable communities”. Furthermore, as the American Planning Association eloquently notes, urban planning “improves the welfare of people and their communities by creating more convenient, equitable, healthful, efficient, and attractive places for present and future generations”, also acting as an enabler for the creation of strong communities. On the other hand, we continue to live in cities that urban planning has not met its stated purpose. The reasons are always complex, and design alone will not be enough, without the corresponding education, culture and customs. However, if this one arbitrary parameter of “feeling” and “perception” is attempted to be measured, a valuable insight can be offered to us, planners, of how people respond to the labeled “attractive” and “convenient” planning. This study produces some valuable results, however it can also serve as a case study of a wider subject, possibly pointing to new methods of evaluating the efficiency of urban design.

An innovative use of mobile EEG was used to investigate walking behaviour, the impact of environmental attributes on individuals’ state of mind and people spatial perceptions. Future research could explore how this technology might support innovative architecture and urban planning strategies. This study suggests that we should not support our planning based solely on physical measurements. This is because each individual perceives a different reality through the experience of the same space, accurately measured with the instrumentation of architects and planners

## **7. Acknowledgements**

We would like to thank the Future Cities Catapult for kindly providing the research equipment, as part of the Cities Unlocked collaboration. PM was financed with an academic scholarship by the A.G. Leventis Foundation.

## **8. Biography**

Katerina Skroumpelou is a PhD student at the School of Electrical and Computer Engineering at the National Technical University of Athens (NTUA). She is an Architectural Engineer of NTUA and holds an MRes on Advanced Spatial Analysis and Visualisation by the Centre for Advanced Spatial Analysis of UCL.

Panagiotis Mavros, is a PhD Candidate and his research is focused on the use of mobile EEG in the study of spatial cognition and behaviour. He was trained as an Architect Engineer at NTUA, and holds an MSc by Research in Digital Media and Culture by the University of Edinburgh.

Dr Andrew Hudson-Smith is Director of the Centre for Advanced Spatial Analysis (CASA) at The Bartlett, University College London. Andy is a Reader in Digital Urban Systems and Editor-in-Chief of Future Internet Journal, he is also an elected Fellow of the Royal Society of Arts, a member of the Greater London Authority Smart London Board and Course Founder of the MRes in Advanced Spatial Analysis and Visualisation and MSc in Smart Cities at University College London.

## 9. References

- American Planning Association. (2010). What is Planning?. Available: <https://www.planning.org/aboutplanning/whatisplanning.htm>. Last accessed March 12th, 2015.
- Aspinall, P., Mavros, P., Coyne, R., & Roe, J. (2013). The urban brain: analysing outdoor physical activity with mobile EEG. *British Journal of Sports Medicine*, 1, 1–7. doi:10.1136/bjsports-2012-091877
- British Government, Department for Environment, Food and Rural Affairs, Noise Mapping, 2012, available at <http://services.defra.gov.uk/wps/portal/noise>)
- British Government, Department for Transport, “Understanding and Valuing the Impacts of Transport Investment”, October 2013
- British Social Attitudes Survey 2011: Public Attitudes Towards Transport, Department for Transport, Copyright 2012
- Dewulf, B., Neutens, T., Van Dyck, D., De Bourdeaudhuij, I., & Van de Weghe, N. (2012). Correspondence between objective and perceived walking times to urban destinations: influence of physical activity, neighbourhood walkability, and socio-demographics. *International journal of health geographics*, 11, 43.
- Garling, T. (1989). THE ROLE OF COGNITIVE MAPS IN SPATIAL DECISIONS. *Journal of Environmental Psychology*, 9, 269–278.
- Handy, S. L., PhD, Marlon G Boarnet, PhD, Reid Ewing, PhD, Richard E Killingsworth, MPH. (2002). How the built environment affects physical activity. *American Journal of Preventive Medicine*. 23 (2), Pages 64–73.
- Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, 52(2), 93–129. doi:10.1016/j.cogpsych.2005.08.003
- Kitchin, R., & Blades, M. (2002). *The cognition of geographic space* (Vol. 4). IB Tauris.
- Meilinger, T., Frankenstein, J., & Bühlhoff, H. H. (2013). Learning to navigate: Experience versus maps. *Cognition*, 129(1), 24–30. doi:10.1016/j.cognition.2013.05.013
- Montello, D.R. (1997). The perception and cognition of environmental distance: Direct sources of information. In *Spatial Information Theory A Theoretical Basis for GIS. Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 297-311.
- Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G., Ritter, W., Ruchkin, D., Rugg, M. and Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology*, 37(02), 127-152.

- Richardson, A. E., Montello, D. R., & Hegarty, M. (1999). Spatial knowledge acquisition from maps and from navigation in real and virtual environments. *Memory & Cognition*, 27(4), 741–50. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10479831>
- Saelens, B.E., Sallis, J.F. & Frank, L.D. (2003). Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. *Annals of behavioral medicine: a publication of the Society of Behavioral Medicine*, 25(2), pp.80-91.
- Waller, D., & Greenauer, N. (2007). The role of body-based sensory information in the acquisition of enduring spatial representations. *Psychological Research*, 71(3), 322–32. doi:10.1007/s00426-006-0087-x

# Semantic and geometric enrichment of 3D geo-spatial building models with photo captions and illustration labels using template matching & SIFT

Jon Slade<sup>\*</sup>, Christopher B. Jones<sup>†</sup> and Paul L. Rosin<sup>‡</sup>

School of Computer Science & Informatics, Cardiff University, United Kingdom

March 12, 2015

## Summary

Whilst the number of 3D geo-spatial digital models of buildings with cultural heritage interest is burgeoning most lack semantic annotation that could be used to inform users of mobile and desktop applications about the architectural features and origins of the buildings. As part of an ongoing project this research describes methods (including template matching and SIFT) for enriching 3D models with generic annotation implied from images of building components and from labelled building plans and diagrams, and with the text from object-specific photo captions from social media. The work is part of a broader initiative aimed at annotating 3D models with elements of text from authoritative architectural guides.

**KEYWORDS:** 3D building model, semantics, enrichment, template matching, SIFT.

## 1. Introduction

3D geo-data is of significant utility in e.g. the fields of inter-visibility analysis, radio communications, visualization for urban planning, indoor navigation, augmented reality, and in cultural heritage applications relating to the built environment. The increasing prevalence of 3D city models has not however been matched by corresponding improvement in their semantic content – often the models lack any semantic content (Jones *et al.*, 2014) which limits their use for some geo-data applications. There is a requirement therefore to develop effective procedures to annotate 3D building models with descriptive attributes. In a cultural heritage context these could include the materials, origins, people and events associated with them.

This research is developing an approach to automate the process of semantic enrichment and is currently in its early stages. The central principle of the work is that texture maps on the building models can be matched to captioned images using computer vision techniques (including template matching and scale-invariant feature transform, or SIFT) allowing the captions to be linked to the corresponding parts of the building model. Recent work has been focussed on image and template matching steps, described below. Future work is then outlined.

With an emphasis upon buildings with cultural heritage the research builds upon methods that used captioned photos to annotate 3D models e.g. Simon and Seitz (2008) and Russell *et al.* (2013), and complements work such as Zhang *et al.* (2013) where 3D models were employed to enhance and

---

<sup>\*</sup> SladeJD@cardiff.ac.uk

<sup>†</sup> JonesCB2@cardiff.ac.uk

<sup>‡</sup> RosinPL@cardiff.ac.uk



annotate photos. Captioned photos might be sourced from social media sites such as *Flickr*.

The lack of well-captioned photos on social media for less well-visited buildings can be addressed through the use of captioned photos in cultural heritage guides – such captions may often be superior in their description of, for example, the architecture and history of the building. Such guides may also contain diagrams of the layout of buildings labelled with the names of rooms and associated spaces. It is this combination of captioned images as found in social media and authoritative guides, and annotated plans and diagrams from authoritative texts, as well as some 2D cartographic data that can be used to improve the semantic content of the building models.

The exploitation of captioned photos of buildings in order to facilitate virtual exploration of buildings was pioneered by *Photo Tourism* (Snavely *et al.*, 2008) in which the computer vision feature matching SIFT method was used to match photos of buildings, compute camera pose and generate point clouds representing the 3D geometry of buildings (referred to as structure from motion). Consequently users could view captioned photos of a building within the recreated 3D scene, in addition to which new photos could be matched and tagged with pre-existing tag content. Simon and Seitz (2008) meanwhile used a related approach to annotate 3D point clouds with *Flickr* photo captions.

Subsequently Russell *et al.* (2013) achieved finer granularity in the annotation of building components, linking text in *Wikipedia* articles about particular buildings to locations in 3D models. Again the models were created from sets of *Flickr* images tagged with the respective building name, with annotation making use of Google *Image* search based on text from the *Wikipedia* entry. As with these other methods Russell *et al.*'s approach used building models comprised of point clouds rather than explicit structure. Our work differs in that we use structured 3D building models, such as those in Trimble's *3D Warehouse*<sup>§</sup>. A number of geo-data applications, such as navigation, can require the ability to refer to individual spatial objects at a fine level of granularity - it is argued here that only *structured* 3D building models could easily provide such capability. Moreover, the generic object annotation which acts as an aide to the latter caption-extraction step and which is described below, requires structural models.

## 2. Methods

### 2.1. Annotating 3D Building Models with Generic Objects

Annotation of rooms, doors, windows, arches, clocks etc may be assisted by the identification of their geometry within the building model. Such geometric representation in a building model may also be of use within applications which guide users around a building. Incidentally, Mayer and Reznik (2005) describe similar approaches within photogrammetry.

Initial work in our study has focused on a template matching approach (Kroon, 2011) whereby a template such as a tightly cropped photo of a generic window design is matched to an image of the building via normalised cross-correlation and non-maximum suppression. An alternative to template matching would be the bag of visual words (BoVW) approach (Sivic and Zisserman, 2003).

Figure 1a shows the result of template matching using a cropped window template (from another building) taken from a Google *Image* search and a texture map from the building model. Since diagrams are a suitable source of templates, we have investigated their use in template matching – Figure 1b shows the result of template matching between the same template and texture map but after both have been converted in to line drawings using the method of Kang *et al.* (2007). Kang *et al.*'s approach has been used since it is effective in extracting long, curvilinear lines while minimising the amount of clutter and noise in the output. The matches in both figures are represented by the bounding boxes in **green**. Note the false positives in Figure 1a – whilst using the line drawing

---

<sup>§</sup> <https://3dwarehouse.sketchup.com>

approach (Figure 1b) has removed these and has detected four out of five windows, the central window on the second storey has not been detected, indicating that further refinement of the method could be conducted.



**Figure 1a Template Matching** with Photos  
Source data: template 89x168 pixels, image 800x904 pixels. Normalised cross correlation non-maximum suppression threshold 0.73



**Figure 1b Template Matching** with Lines  
Source data: template 95x196 pixels, image 800x904 pixels. Normalised cross correlation non-maximum suppression threshold 0.62 \*

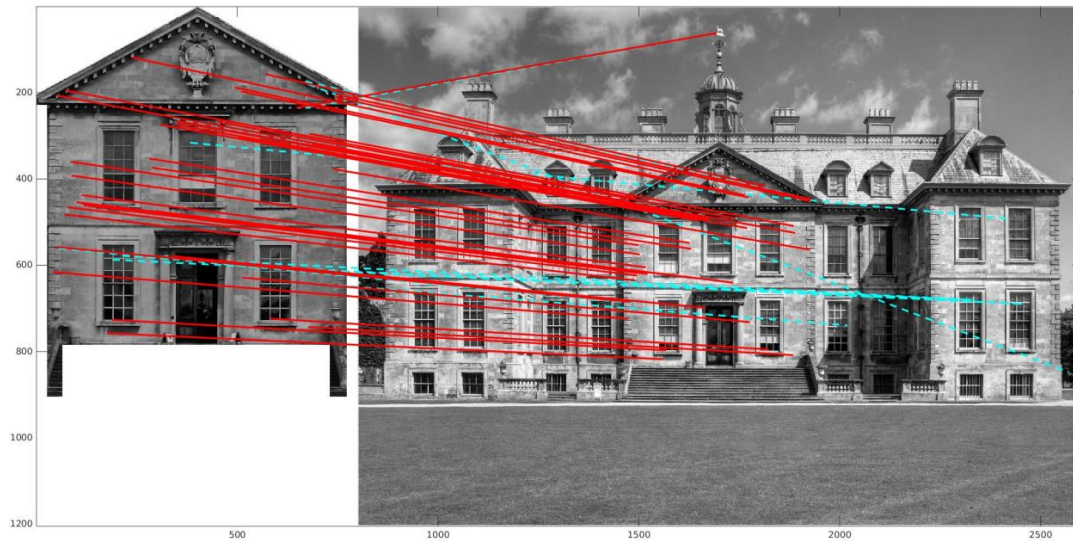
\* Blurring has been carried out in advance of matching in an attempt to consolidate scattered disjoint components of what should be solid lines, and to extend the influence of the highly localised features i.e. lines, thus increasing the tolerance of the matching to geometric differences between the template and the texture map image. The image was also padded by 100 pixels to remove erroneous detections at its edges.

## 2.2. Building Plan Exploitation

The text from supplementary building plans can be added to the models, following a process of spatially matching the ground or side projection in the plans with the geometry of the footprint of the model. It is envisaged that such spatial joining, via GIS conflation (Samal *et al.*, 2004), will require both geometric and semantic matching. Should there be ambiguity in the match then various machine learning techniques such as probabilistic mutual information (Walter and Fritsch, 1999) and Bayesian methods (Jones *et al.*, 1999) could be used. Incidentally, whilst most conflation works on vector data, the work of Smart *et al.* (2011) presented matching methods that entailed rasterising building geometry which lacked complete, contiguous structure before matching the result with a rasterized digital map.

## 2.3. Feature Matching Methods and Linking Captions to Models

Image feature matching methods such as SIFT are now widely used within computer vision (Mikolajczyk *et al.*, 2005). In this research these feature matching methods support the object detection phase, enabling the matching of captioned photos. SIFT can facilitate matching by identifying quantitative vector descriptors for distinctive local regions or key points in an image, where the similarity measures for those descriptors are then used to inform matches. Methods such as RANSAC (Fischler and Bolles, 1981) can be used to reduce outliers which may occur – this is achieved through the generation of a fundamental matrix for the perspective transformation between the two images where those matches which are inconsistent with the transform can then be removed.



**Figure 2a** Key point matching



**Figure 2b** Convex hull on building model of matching inlier key points

(a) Key point matching between a *Flickr* image of Belton House (on the right) and the texture map imagery of a 3D model of Belton House (on the left). Inlier matches are coloured as **red** lines and outliers as dashed **cyan** lines

(b) The 3D texture-mapped model of Belton House in which the convex hull region of the matching inlier key points is highlighted in **red**.

Figures 2a and 2b shows the result of SIFT-based matching (with RANSAC outlier-removal) between a captioned photo from *Flickr* of the heritage English mansion Belton House in Lincolnshire, with a corresponding texture map from the surface of a 3D model of the same building taken from the *3D Warehouse*. Note that there is a strong cluster of matching key points on the gable front, pedimented bay above the main entrance steps (Figure 2a) i.e. the inlier matched key point pairs, represented in the figure with **red** lines. Other, false matching outlier key point pairs are highlighted with dashed **cyan** lines in Figure 2a. The matched region on the 3D model is highlighted in **red** in Figure 2b and is based on the convex hull of the matching inlier key points. The caption is linked, initially, to this region.

Future work may refine this matching region, perhaps using region growing from the inlier key points whereby adjacent pixels from the captioned image could be transformed to the texture map image using the estimated fundamental matrix, and retained if the (dense) SIFT descriptors of the corresponding pixels match.

## 2.4. Linking to Rich Text Descriptions

Russell *et al.* (2013) demonstrated the linking of descriptive texts to 3D building models, in their case from *Wikipedia*, and our work plans to build upon that by linking the models to authoritative texts. Based on their examples however, their methods are quite selective (relying on the existence of architectural components in just a *Wikipedia* entry) and can therefore fail to match interesting descriptions to respective building components. Depending as they do on the availability of well-captioned crowdsourced images on social media, their approach and the field in general would benefit from improved methods for matching texts from sources other than on social media – in turn this might pave the way for semantic enrichment of models of less popular buildings.

One such approach might involve natural language processing to detect and interpret the often vague spatial relations found in descriptions of the locations of real-world objects (Kordjamshidi *et al.*, 2011; Mani *et al.*, 2010). For example a door might be described as being in the east wall of a church – in turn identification of the respective wall of the 3D building model and hence the generically labelled geometric object within that wall, may be achieved. Moreover, terms such as right and left, and architectural descriptions such as arch or lintel may improve object location determination. Captioned photos may also be exploited to assist in the geometric annotation process.

## 3. Concluding Remarks

We have outlined an approach for semantic and geometric enrichment of existing 3D building models, exploiting captioned social media photos and labelled ground plans obtained from illustrated guides. Our methods differ from the current state of the art in that we employ structured 3D models, rather than 3D point clouds, and focus on enhancing geometry and generic annotation of objects within these models.

The latter, achieved through computer vision methods and GIS conflation, facilitates linking text descriptions of particular object types to corresponding components in the 3D model. This avoids the dependence of existing methods on the need for large numbers of captioned photos. Finally, we pave the way for linking rich textual descriptions in authoritative guides to corresponding geometric objects.

## 4. Acknowledgements

Jon Slade is a PhD candidate funded by an EPSRC Industrial CASE studentship with Ordnance Survey, Great Britain.



## 5. Biography

Jon Slade holds an MSc in GIScience from UCL, with a BSc in Geology from the University of Bristol. Previous roles comprise 13 years as a consultant in the commercial IT sector including work as a GIS Analyst Developer at civil engineering and mobile telecoms firms.

Chris Jones is Professor of Geographical Information Systems in Cardiff University. His research interests include multi-scale spatial database design and integration, map generalisation and geographical information retrieval with particular regard to spatio-textual indexing methods, gazetteers, vernacular place names and spatial search engine architectures.

Paul L. Rosin is Professor of Computer Vision in Cardiff University. Previous posts include Brunel University, Joint Research Centre, Italy and Curtin University of Technology, Australia. His research interests include computer vision, remote sensing, mesh processing, non-photorealistic rendering, and the analysis of shape in art and architecture.

## References

- FISCHLER, M. A. AND BOLLES, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* [online], 24(6): 726-740. Available from: [http://dl.acm.org/ft\\_gateway.cfm?id=358692](http://dl.acm.org/ft_gateway.cfm?id=358692) [Accessed 11 June 2014].
- JONES, C. B., ROSIN, P. L. AND SLADE, J., 2014. Semantic and geometric enrichment of 3D geo-spatial models with captioned photos and labelled illustrations [online]. In: I. Press, (Ed). *Workshop on Vision and Language (VL4)*, Dublin. Available from: <http://www.aclweb.org/anthology/W/W14/W14-54.pdf#page=74> [Accessed 18 August 2014].
- JONES, C. B., WARE, J. M. AND MILLER, D. R., 1999. A Probabilistic Approach to Environmental Change Detection with Area-Class Map Data. [online] In: P. Agouris and A. Stefanidis (Eds.) *Integrated Spatial Databases*. Springer Berlin Heidelberg. pp. 8. Available from: [http://link.springer.com/content/pdf/10.1007%2F3-540-46621-5\\_8.pdf](http://link.springer.com/content/pdf/10.1007%2F3-540-46621-5_8.pdf) [Accessed 11 June 2014].
- KANG, H., LEE, S. AND CHUI, C. K., 2007. Coherent line drawing [online]. In: M. Agrawala and O. Deussen, (Eds). *5th International Symposium on Non-photorealistic Animation and Rendering (NPAR)*, San Diego, USA, 04-05 August 2007 1274878. ACM. Available from: [http://dl.acm.org/ft\\_gateway.cfm?id=1274878](http://dl.acm.org/ft_gateway.cfm?id=1274878) [Accessed 05 November 2014].
- KORDJAMSHIDI, P., VAN OTTERLO, M. AND MOENS, M.-F., 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing* [online], 8(3): 1-36. Available from: [http://dl.acm.org/ft\\_gateway.cfm?id=2050105](http://dl.acm.org/ft_gateway.cfm?id=2050105) [Accessed 11 June 2014].
- KROON, D.-J., 2011. *MATLAB Central - File Exchange - Fast/Robust Template Matching* [online]. MathWorks Inc. Available from: [http://www.mathworks.co.uk/matlabcentral/fileexchange/24925-fast-robust-template-matching/content/template\\_matching.m](http://www.mathworks.co.uk/matlabcentral/fileexchange/24925-fast-robust-template-matching/content/template_matching.m) [Accessed 01 September].
- MANI, I., DORAN, C., HARRIS, D., HITZEMAN, J., QUIMBY, R., RICHER, J., WELLNER, B., MARDIS, S. AND CLANCY, S., 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation* [online], 44(3): 263-280. Available from: <http://link.springer.com/content/pdf/10.1007%2Fs10579-010-9121-0.pdf> [Accessed 11 June 2014].
- MAYER, H. AND REZNIK, S., 2005. Building facade interpretation from image sequences [online]. In: S. U, R. F and H. S, (Eds). *ISPRS Workshop on Object Extraction for 3D City Models, Road Databases, and Traffic Monitoring - Concepts, Algorithms, and Evaluation (CMRT)*, Vienna, Austria, 29-30 August 2005. Available from: [http://www.isprs.org/proceedings/XXXVI/3-W24/papers/CMRT05\\_Mayer\\_Reznik.pdf](http://www.isprs.org/proceedings/XXXVI/3-W24/papers/CMRT05_Mayer_Reznik.pdf)

- [Accessed 02 September 2014].
- MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T. AND GOOL, L. V., 2005. A Comparison of Affine Region Detectors. *International Journal of Computer Vision* [online], 65(1-2): 43-72. Available from: <http://link.springer.com/content/pdf/10.1007%2Fs11263-005-3848-x.pdf> [Accessed 14 July 2014].
- RUSSELL, B. C., MARTIN-BRUALLA, R., BUTLER, D. J., SEITZ, S. M. AND ZETTLEMOYER, L., 2013. 3D Wikipedia: using online text to automatically label and navigate reconstructed geometry. *ACM Transactions on Graphics (TOG)* [online], 32(6): 193. Available from: [http://dl.acm.org/ft\\_gateway.cfm?id=1141964](http://dl.acm.org/ft_gateway.cfm?id=1141964) [Accessed 20 June 2014].
- SAMAL, A., SETH, S. AND CUETO, K., 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science* [online], 18(5): 459-489. Available from: <http://www.tandfonline.com/doi/full/10.1080/13658810410001658076> [Accessed 11 June 2014].
- SIMON, I. AND SEITZ, S. M., 2008. Scene Segmentation Using the Wisdom of Crowds [online]. In: D. Forsyth, P. Torr and A. Zisserman, (Eds). *10th European Conference on Computer Vision (ECCV)*, Marseille, France, 13-16 October 2008. Springer Berlin Heidelberg. Available from: <http://grail.cs.washington.edu/pub/papers/simon08ss.pdf> [Accessed 11 June 2014].
- SIVIC, J. AND ZISSERMAN, A., 2003. Video Google: a text retrieval approach to object matching in videos [online]. In: B. Werner, (Ed). *9th IEEE International Conference on Computer Vision (ICCV)*, Nice, France, 13-16 October. IEEE Computer Society. Available from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1238663> [Accessed 11 June 2014].
- SMART, P. D., QUINN, J. A. AND JONES, C. B., 2011. City model enrichment. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 66(2): 223-234. Available from: <http://www.sciencedirect.com/science/article/pii/S0924271610001231> [Accessed 11 June 2014].
- SNAVELY, N., SEITZ, S. M. AND SZELISKI, R., 2008. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision* [online], 80(2): 189-210. Available from: <http://link.springer.com/content/pdf/10.1007%2Fs11263-007-0107-3.pdf> [Accessed 11 June 2014].
- WALTER, V. AND FRITSCH, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science* [online], 13(5): 445-473. Available from: <http://www.tandfonline.com/doi/pdf/10.1080/136588199241157> [Accessed 11 June 2014].
- ZHANG, C., GAO, J., WANG, O., GEORGEL, P., YANG, R., DAVIS, J., FRAHM, J.-M. AND POLLEFEYS, M., 2013. Personal Photograph Enhancement Using Internet Photo Collections. *IEEE Transactions on Visualization and Computer Graphics* [online], 20(2): 262-275. Available from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6509872> [Accessed 11 June 2014].

# Assessing the impact of seasonal population fluctuation on regional flood risk management

Alan Smith<sup>\*1</sup>, Andy Newing<sup>2</sup>, Niall Quinn<sup>3</sup>, David Martin<sup>1</sup> and Samantha Cockings<sup>1</sup>

<sup>1</sup>Geography and Environment, University of Southampton, UK

<sup>2</sup>School of Geography, University of Leeds, UK

<sup>3</sup>School of Geographical Sciences, University of Bristol, UK

03 November 2014

## Summary

This paper focuses on the integration of population and environmental models to address the effect of seasonally varying populations on exposure to flood risk. A spatiotemporal population modelling tool, Population24/7, has been combined with LISFLOOD-FP inundation model outputs for a study area centred on St Austell, Cornwall, UK. Results indicate seasonal cycles in populations and their exposure to flood hazard which are not accounted for in traditional population datasets or flood hazard analyses and which provide potential enhancements to current practice.

**KEYWORDS:** Spatiotemporal population modelling, flood risk, Population24/7, LISFLOOD-FP, seasonality.

## 1. Introduction

Previous research applying high resolution spatiotemporal population modelling to flood risks has shown large variations in population exposure over time and space (e.g. Smith *et al.* 2014). A major refinement in this approach has been the inclusion of seasonally varying overnight visitor population estimates developed by Newing *et al.* (2013). These have been integrated within the flexible Population24/7 data framework (Martin *et al.* forthcoming) which can be used to produce spatiotemporal gridded population estimates using variable kernel density estimation methods.

This paper demonstrates analysis of seasonal variations in population exposure to flood risk using a local case study. It combines spatiotemporal population estimates with an extract from the national Environment Agency (EA) flood map and bespoke LISFLOOD-FP inundation modelling. The integration of seasonal tourist population estimates represents an advance in modelling spatiotemporal populations for applications such as population hazard exposure.

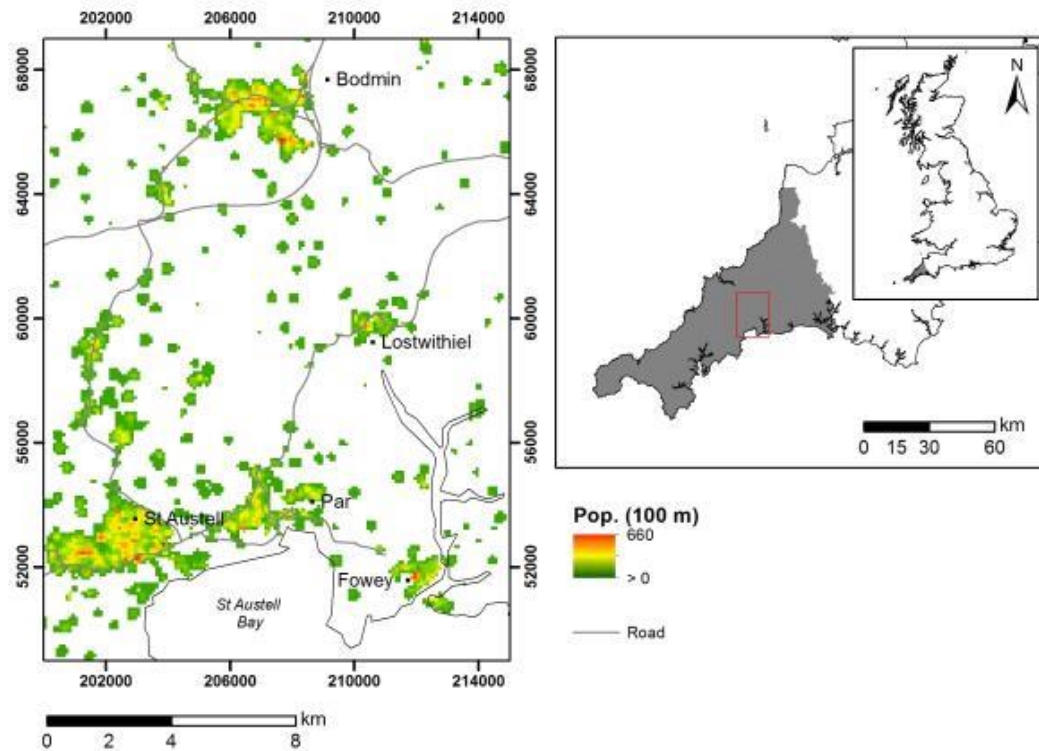
## 2. Case study: St Austell, Cornwall

This study is based on a  $15 \times 20$  km study area centred on St Austell Bay, Cornwall, UK (Figure 1). Coastal resorts within the study area, such as Par, Polkeris and Fowey, experience considerable seasonal fluctuations in population driven by an influx of domestic overnight visitors.

---

\* Alan.Smith@soton.ac.uk

The study area is subject to fluvial, tidal and surface water flooding. The warning system on the River Par provides less than two hours' notice of flooding (EA 2012a). The 'tide-locking' of local watercourses during high tides prevents drainage at coastal outlets and poses an additional risk of fluvial flooding. Tidal flood risk dominates the east of the study area. The Par area contains the highest number of properties at risk from current and predicted future flooding in Cornwall (EA 2012b).



**Figure 1** St Austell study area outlined in red, showing location within Cornwall (shaded grey) and Great Britain insets. An example 100 metre gridded population distribution provided for contextual purposes.

### 3. Methods and data

Hourly population estimates at 100 metre resolution have been produced for a 'typical' weekday in January, May and August 2010 using the SurfaceBuilder247 software tool (<http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1811/read>). These scenarios demonstrate the considerable variation seen in estimated seasonal visitor numbers within the case study area, reflecting the low, fringe and peak tourist seasons respectively.

Population is redistributed from origin (residential locations) to destination (e.g. locations of work, study, leisure) centroids. Population redistribution is constrained by a dasymetric background mask which includes the road transport network. The occupation of destination centroids is governed by a temporal profile specific to each site (e.g. a school occupied during school hours in term-time). The modelling framework is described in greater detail by Martin *et al.* (forthcoming). This particular application integrates novel overnight visitor population estimates within the modelling framework (Section 3.1) and combines these with bespoke flood inundation modelling using LISFLOOD-FP (Section 3.2).



### 3.1. Seasonal population fluctuation

Tourist visitor populations have a tendency to cluster in both space and time. In coastal areas, such as St Austell Bay, a concentration of visitor accommodation, attractions and other facilities generate spatial clusters of visitors, with numbers known to fluctuate at different times of the year, driven by the weather, local and national events, the institutional calendar and the operating season at accommodation sites and major attractions. In contrast to residential and workplace populations, very little is known about the spatial or temporal distribution of overnight visitors below the local authority district level.

We make use of a novel dataset estimating the seasonal and spatial distribution of visitors based on the provision and utilisation of tourist accommodation (Newing *et al.* 2013). Overnight visitor population estimates (Table 1) were built from the ‘bottom-up’ using local data collection and taking individual accommodation ‘units’ (e.g. a hotel room, self-catering cottage or camping pitch) as the building block, aggregated to the unit postcode or census Output Area (OA) level to form visitor ‘origins’. Visitors staying with friends and relatives are distributed across the existing housing stock, whilst major holiday parks and camping and caravanning sites generate spatial and temporal clusters of overnight visitor populations in areas which may have few or no usual residents.

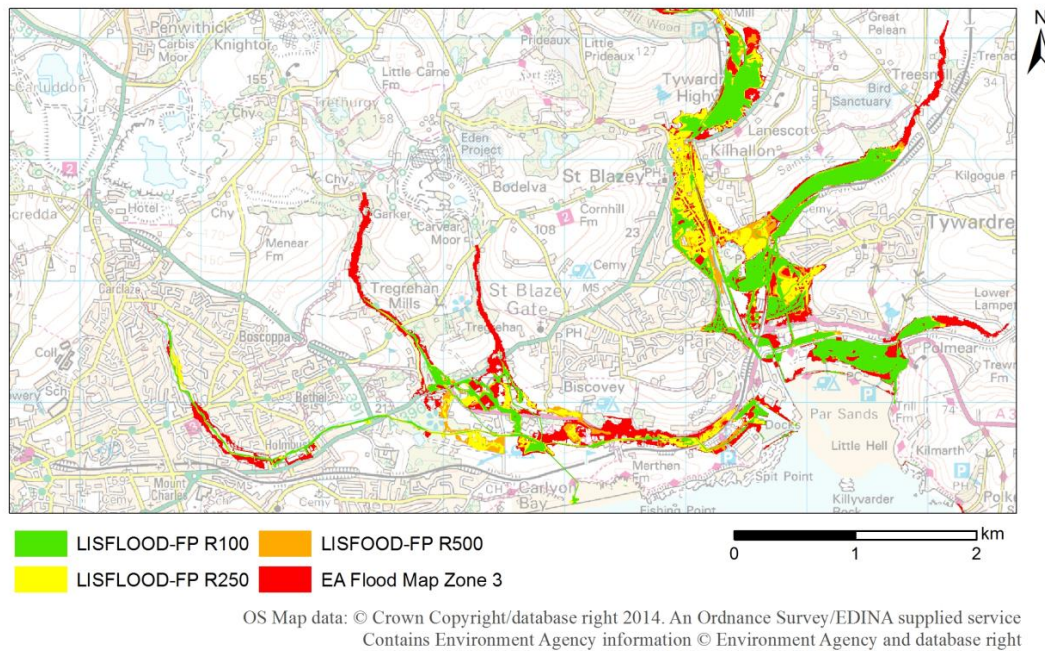
**Table 1** Overnight visitor estimates within the St Austell study area

Month (season)	Overnight visitor estimate
January 2010 (Low)	1,049
May 2010 (Fringe)	6,269
August 2010 (Peak)	12,389

In common with the approach used for residential populations, visitor populations are redistributed from their overnight origins to daytime locations such as major attractions, the transport network and leisure locations which may not traditionally be thought of as clusters of population in the same way as workplaces, hospitals and retail centres. Given the coastal and estuarine nature of the study area, some of these locations may also be at flood risk.

### 3.2. Flood inundation modelling (LISFLOOD-FP)

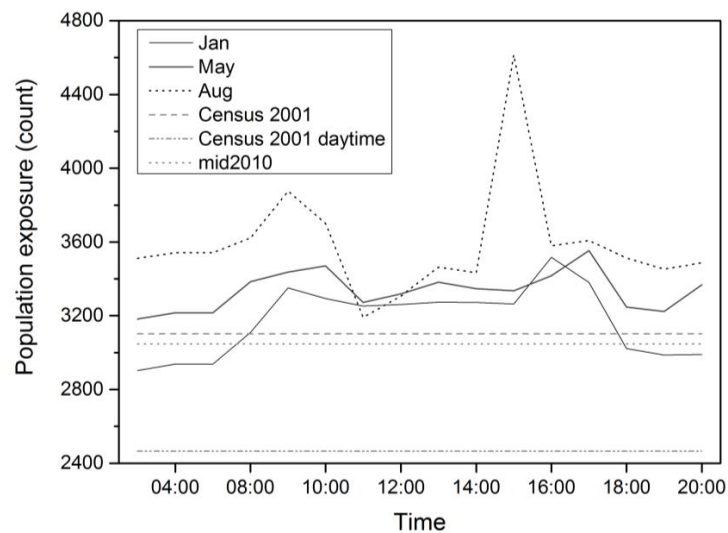
Three flood scenarios representing return periods (R) of 100, 250 and 500 years have been created using LISFLOOD-FP, a raster based flood inundation model, for an  $8 \times 4$  km subsection of the study area (Figure 2). The return period represents the likelihood of an event of a given magnitude occurring. In addition, the EA’s flood map zone three (FMZ3) represents the amalgamation of return periods for events of 100 (fluvial) and 200 (tidal) years. FMZ3 contains a greater extent of inland fluvial flooding whereas the bespoke LISFLOOD-FP outputs specifically account for defences and other structures.



**Figure 2** Comparison of LISFLOOD-FP and EA flood inundations for the selected area within the study area covering St Austell and Par.

#### 4. Results and discussion

This section presents the integration and analysis of spatiotemporal seasonal population estimates and flood inundation models. Figure 3 shows hourly spatiotemporal population estimates representing a ‘typical’ working weekday within each seasonal scenario. It shows the population exposure to the EA’s FMZ3 for the whole study area. It has been compared with static exposure estimates from rasterised census outputs representing: the baseline 2001 Census population at OA level (highest resolution available), 2001 Census daytime population at OA (only available for 2001) and the 2010 mid-year estimate (closest to target date but only available LSOA level).



**Figure 3** Flood exposure estimates from the EA Flood Map Zone 3 for the St Austell study area

The seasonal spatiotemporal variation is illustrated for weekday midday (12:00) and midnight (00:00) population estimates for two of the seasons modelled (January and August) (Figures 4 and 5). In both examples there is a general redistribution of the usually resident day-time population from the night-time locations to the main population centres (analogous with the main workplace locations), reflecting a daily transition from the surrounding rural areas into urban locations. The overnight visitor population increases by over 1000% January-August (Table 1) and is concentrated at leisure activity locations during the day.

Figure 6 shows modelled population results with OS map extracts and aerial imagery for 1 km national grid squares, providing detailed examples for an August weekday at 12:00 and 00:00. Daytime locations with example visitor leisure attractions are shown in (A) and (B). Night-time caravan and camping locations occupied by overnight visitors in August are shown in (C) and (D). The overnight visitors at these locations are completely missing from the traditional census datasets.

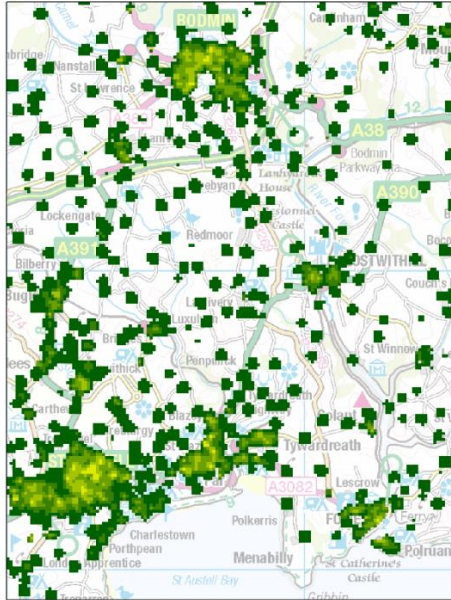
Population exposure to the varying flood risks, by season, residents and visitors for the St Austell study area is quantified in Table 2. This application has demonstrated what Martin *et al.* (forthcoming) term the ‘modifiable spatiotemporal areal unit problem’ whereby even the most detailed spatial data may be inadequate to support time-sensitive analyses. In this case, population exposure is highly dependent on time of day, season of the year and varying extent of flood inundation polygons.

**Table 2** Daytime usually resident and visitor population exposure to three LISFLOOD-FP inundations scenarios (R = return period) and EA flood map zone three for January, May and August (increasing levels of inundation left to right).

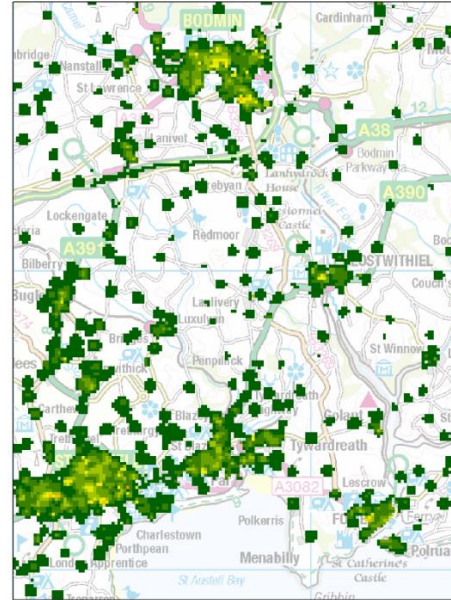
Population	LISFLOOD-FP R100	LISFLOOD-FP R250	LISFLOOD-FP R500	EA Flood Map
Residents 12:00 Jan	542	939	1069	1725
Visitors 12:00 Jan	2	5	7	15
<i>Total</i>	<i>544</i>	<i>944</i>	<i>1076</i>	<i>1740</i>
Residents 12:00 May	546	994	1139	1729
Visitors 12:00 May	34	108	131	114
<i>Total</i>	<i>580</i>	<i>1102</i>	<i>1270</i>	<i>1843</i>
Residents 12:00 Aug	498	1019	1178	1741
Visitors 12:00 Aug	65	206	249	212
<i>Total</i>	<i>563</i>	<i>1225</i>	<i>1427</i>	<i>1953</i>



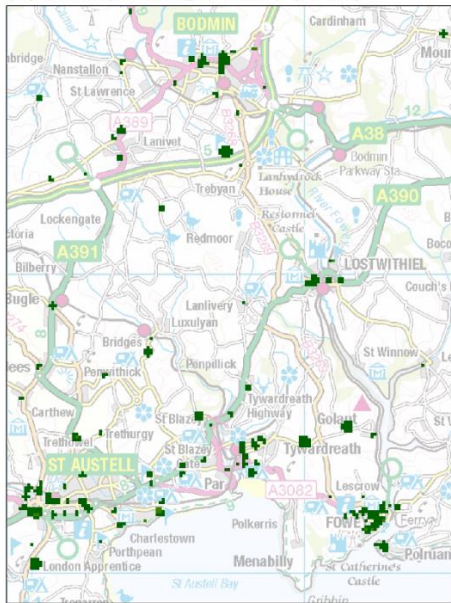
(A) January residents (00:00)



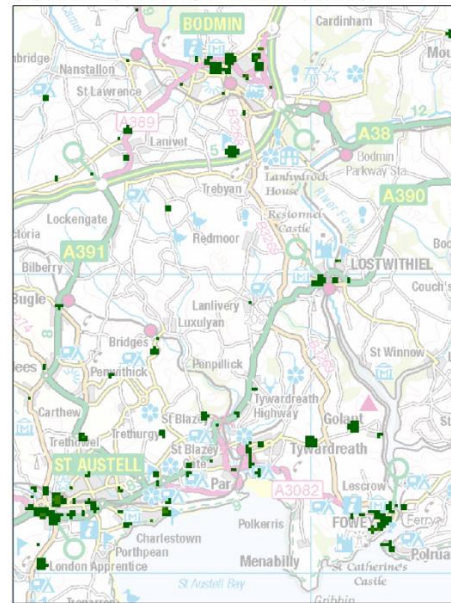
(B) January residents (12:00)



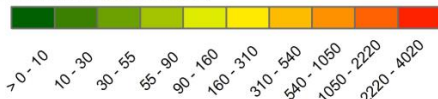
(C) January overnight visitors (00:00)



(D) January overnight visitors (12:00)



Population (100 metres)

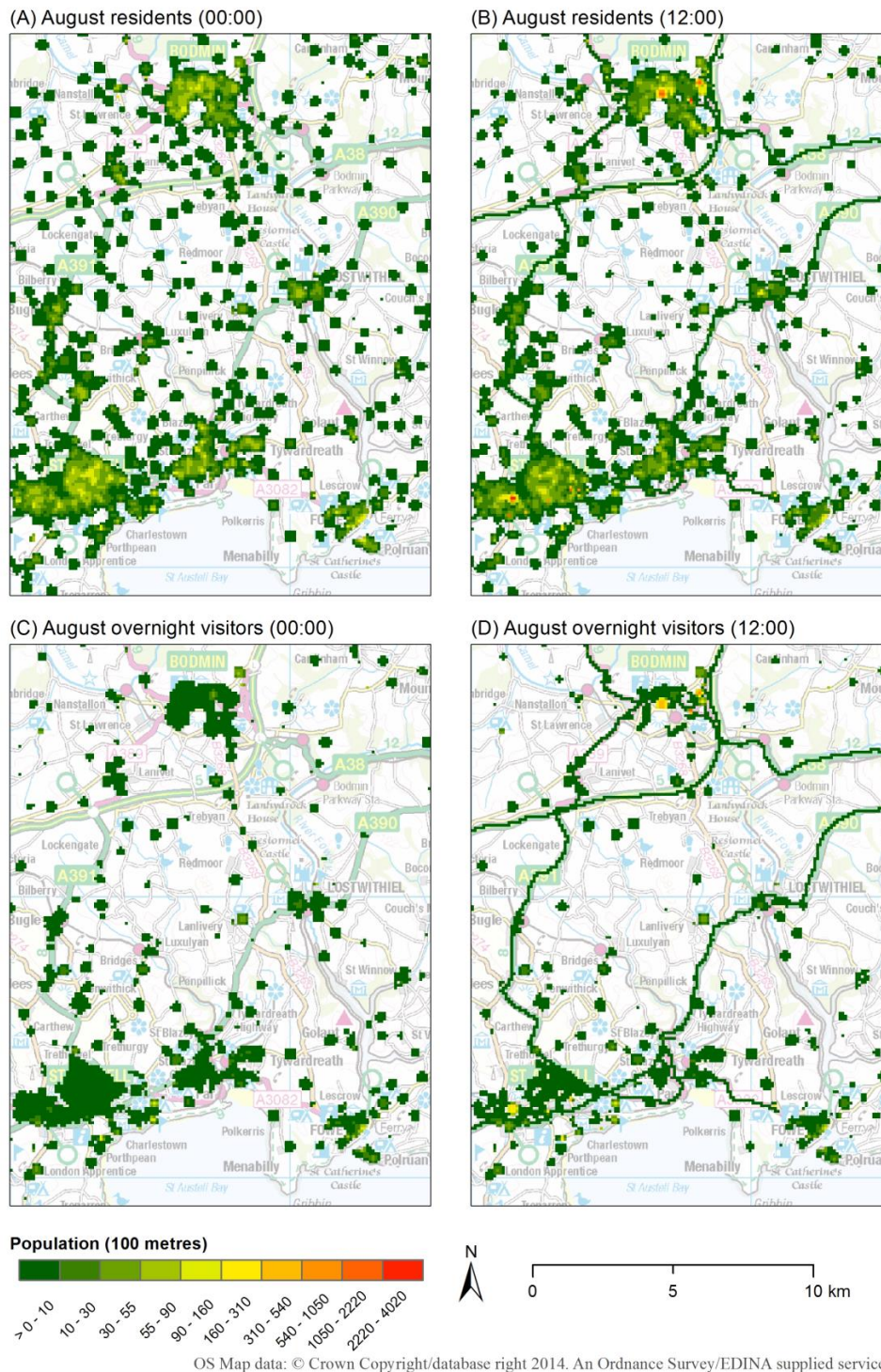


0 5 10 km

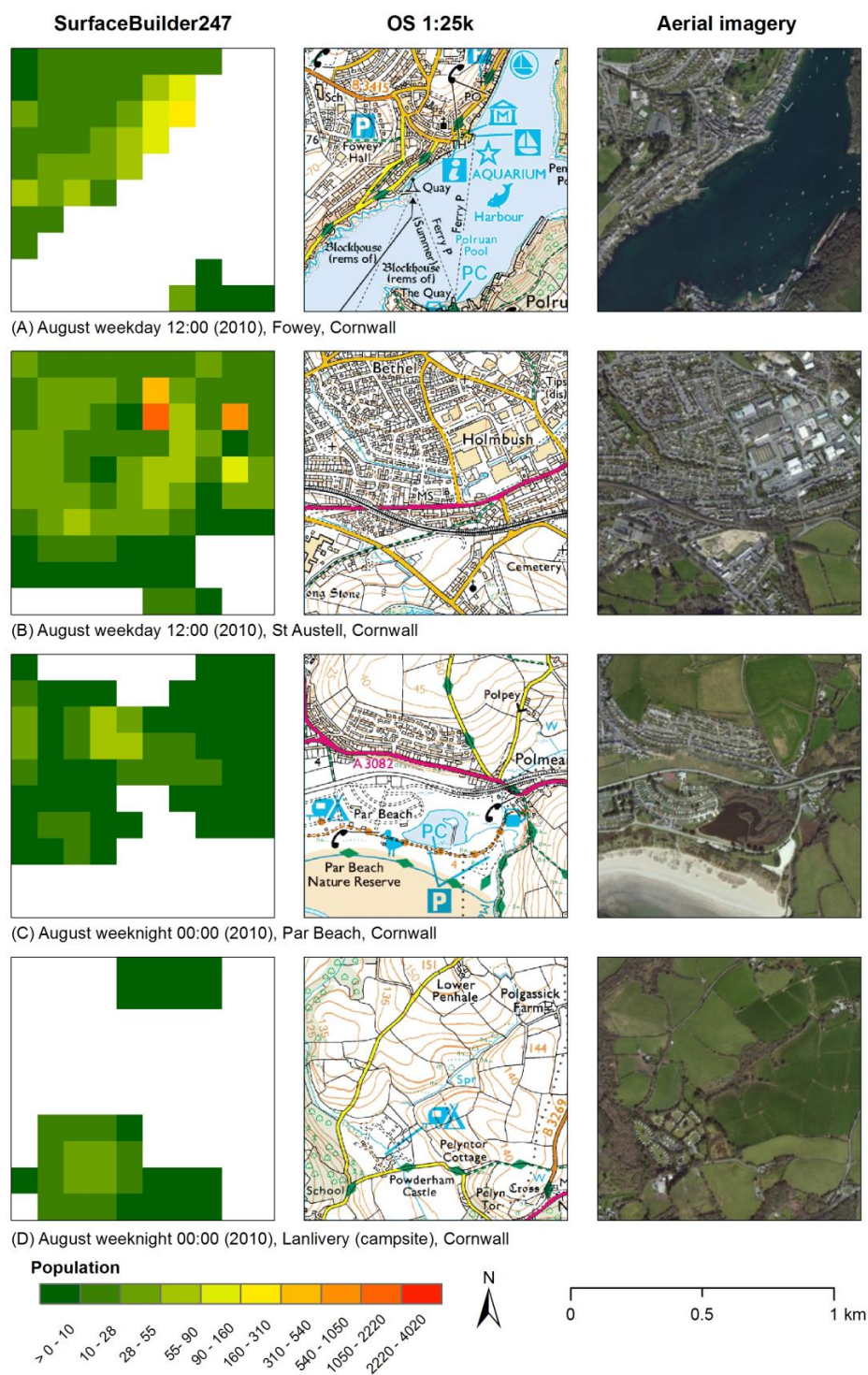
OS Map data: © Crown Copyright/database right 2014. An Ordnance Survey/EDINA supplied service

**Figure 4** Modelled seasonal population outputs (100 m) for the St Austell study area for a January weekday. (A) Usually resident night-time (00:00) population, (B) Usually resident daytime (12:00) population, (C) Overnight visitor night-time (00:00) population and (D) Overnight visitor daytime (12:00) population.





**Figure 5** Modelled seasonal population outputs (100 m) for the St Austell study area for an August weekday. (A) Usually resident night-time (00:00) population, (B) Usually resident daytime (12:00) population, (C) Overnight visitor night-time (00:00) population and (D) Overnight visitor daytime (12:00) population.



OS Map data: © Crown Copyright/database right 2014. An Ordnance Survey/EDINA supplied service  
Aerial Imagery: Esri, DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo

**Figure 6** A detailed comparison of SurfaceBuilder247 (100 metre resolution) results within the St Austell study area with 1:25000 scale Ordnance Survey (OS) background mapping and aerial imagery for selected 1 km national grid squares. (A) and (B): August weekday ‘daytime’ population. (C) and (D): August weekday ‘night-time’ population.

## 5. Conclusions

The spatio-temporal modelling framework adopted here has facilitated the inclusion of a highly important seasonally varying tourist population into estimates of population exposure to flood risk. This not only enhances current insights into high resolution spatio-temporal population movements but also demonstrates the large potential impact of temporary populations on assessing flood risk and hazard exposure. Flood risk to people is variable and depends on many factors other than the usually resident population base. Such insights are simply not possible using static or traditional datasets in isolation. This approach exemplifies one possible integration of population and physical models for both environmental and wider applications.

## 6. Acknowledgements

This research is supported by a University of Southampton ESRC doctoral training award.  
Data: Flood map data provided by the Environment Agency (April 2014).

## 7. Biographies

*Alan Smith is an ESRC doctoral researcher in Geography and Environment at the University of Southampton. His research interests include spatio-temporal population modelling and the integration with environmental models.*

*Andy Newing is a lecturer in Retail Geography at the University of Leeds. His research interests include spatio-temporal population estimation for applications related to service provision and delivery.*

*Niall Quinn is a post-doctoral researcher in Geography at Bristol University. His research interests are focussed around uncertainties in flood risk prediction.*

*David Martin is a Professor of Geography at the University of Southampton. His research interests are focused on social science applications of geographical information systems. He developed the SURPOP and SurfaceBuilder methodologies and software.*

*Samantha Cockings is an Associate Professor in Geography at the University of Southampton. Her research interests include automated zone design, space-time modelling of populations and links between environment and health.*

## References

- EA (2012a). *West Cornwall Catchment Flood Management Plan*. Exeter: Environment Agency.
- EA (2012b). *East Cornwall Catchment Flood Management Plan*. Exeter: Environment Agency.
- Martin, D., Cockings, S., & Leung, S. (forthcoming). Developing a flexible framework for spatiotemporal population modelling. *Annals of the Association of American Geographers*.
- Newing, A., Clarke, G., & Clarke, M. (2013). Visitor expenditure estimation for grocery store location planning: A case study of Cornwall. *International Review of Retail, Distribution and Consumer Research*, 23(3), 221-244, doi:DOI:10.1080/09593969.2012.759612.
- Smith, A. D., Martin, D., & Cockings, S. (2014). Spatio-Temporal Population Modelling for Enhanced Assessment of Urban Exposure to Flood Risk. *Applied Spatial Analysis and Policy*, 1-19, doi:10.1007/s12061-014-9110-6. Online first.



# Creating a spatio-temporal “Data Feed” API for a large and diverse library of historical statistics for areas within Britain

Humphrey Southall<sup>\*1</sup> and Michael Stoner<sup>†1</sup>

<sup>1</sup>Department of Geography, University of Portsmouth

June 7, 2015

## Summary

The GB Historical GIS holds 14m. diverse statistical data values in a uniform structure linked to a geospatial ontology of reporting units and a domain ontology of statistical concepts. This paper describes the addition of a Linked Data API enabling programmatic access to this “big data” structure and discusses topical and spatial sub-setting.

**KEYWORDS:** Social statistics, Open Linked Data, Datacube Vocabulary, Historical GIS

## 1. Introduction

Numerous examples, notably Google’s search engine, have shown the vast power of even quite simple analytic tools when accessing really large amounts of unstructured text. Similar approaches fail with statistics because individual data values lack intrinsic meaning, and the frameworks which give statistics meaning often divide them into silos, making automated analysis possibly only within individual silos, not across them. For example, the UK Data Service’s catalogue holds consistent information about very many “studies”, enabling users to download particular datasets. Their internal structure was hopefully documented by dataset creators to enable users to unpack them, but to the repository system data and documentation are essentially binary large objects, so you can’t “analyse the Data Service” *in toto*.

Although the organisation of data archives into studies and datasets has changed little over forty years, their development of new metadata frameworks enables radically new approaches if unencumbered by legacy holdings. Preparations for the 2011 census included much discussion of a “data feed” Applications Programming Interface, leading to a Census Web Services working group; and, eventually, to the Office of National Statistics launching an experimental API (<https://www.ons.gov.uk/ons/apiservice/web/apiservice>) and the Data Service launching InFuse, an API-based replacement for Casweb. However, ONS’s API is constrained by the organisation of the c. 8 billion 2011 data values into about 400 “Tables”, a paper-derived concept now constraining the almost purely digital. The InFuse API remains private and undocumented.

We describe here a more integrated approach, influenced by these projects but, perhaps surprisingly, less constrained by legacy data: our Great Britain Historical GIS relies mainly on data we computerise ourselves from old paper reports. Two more closely related systems are the Irish government’s census gateway (<http://data.cso.ie/index.html>; Maali et al, 2012), although our historical work must address long periods of time, diverse reporting geographies and diverse categorisations of similar things; and the Dutch CEDAR project (Meroño-Peñuela, 2014).

---

<sup>\*</sup> Humphrey.Southall@port.ac.uk

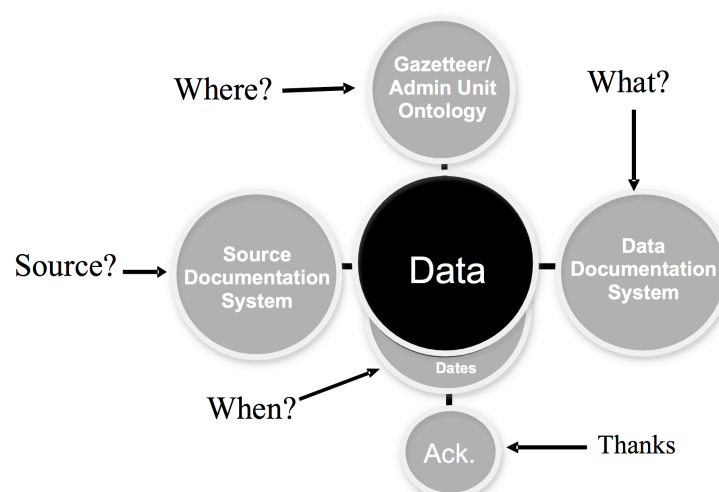
<sup>†</sup> Michael.Stoner@port.ac.uk



## 2. Holding a large and diverse library of statistical data in one table

Our data library currently comprises 14,099,469 data values including 8,525,126 numbers drawn from all British censuses 1801-2001 down to parish level; 3,202,448 death counts categorised by district, cause, age and gender from vital registration reports 1851-1910; 70,580 vote counts covering every candidate in every constituency in every parliamentary election 1833-2008; and 90,958 counts from farm censuses 1866-1971: a fair summary of the quantitative history of Britain's localities. Each value forms one row in the central "data table" within our Postgres database, including 2,100,844 copies of the number zero, 1,136,484 copies of one, etc.

Any particular zero is given meaning by other data table columns, mostly identifiers given meaning by sub-systems: a date; an ID defining geographical coverage; a source identifier; an "acknowledgment ID" identifying transcribers; and a "cell reference" recording *what* the number measures by locating it within an n-dimensional hypercube, or "nCube".



**Figure 1** Simplified structure of Great Britain Historical GIS

This last is based on the Aggregate Data Extension to the Data Documentation Initiative (DDI) standard (Southall, 2011; <http://www.ddialliance.org/>) and enables *Vision of Britain* to automatically select the most appropriate way to visualise any given dataset (Southall, 2008). In brief, the dimensions of an nCube are defined by the variables and categories in the underlying microdata, such as age group, gender and occupation; nCubes and variables are organised within a hierarchy of topics; and labels and explanatory text can be held for all these entities. We have shown that an absolutely fixed set of database tables can not only hold any amount of data, but also cover an endlessly expanding geographical scope and set of topics.

## 3. Defining the API

We have also implemented a Linked Data API accessing the PastPlace gazetteer, returning information about both "places", such as Portsmouth, and administrative units such as Portsmouth Registration District and Portsmouth County Borough, using separate systems of numerical place and unit identifiers, and consequently two sets of Uniform Resource Identifiers (URIs) (Southall, 2012). The API can be searched using a place name or by specifying a bounding box.

Our new PastPlace DataCube API is based on the World Wide Web Consortium's Datacube Vocabulary (Cyganiak *et al*, 2014) and is implemented as a separate application running within

Tomcat using Apache Jena, an open source Semantic Web framework for Java (<https://jena.apache.org/>). Jena serialises RDF graphs into different output formats including RDF/XML, Turtle, and Notation 3.

The sample below contains just one actual data value, 53,058, the total population of Portsmouth Registration District (unit 10154984) in 1841. That value appears half way down, and dates and location are identified concisely so most output captures data semantics. Where possible we link to externally defined vocabularies, such as Dublin Core (e.g. dc.publisher) but for now much is self-defined (Kramer *et al*, 2012). Current prototype API access for this data value is [http://testapi.pastplace.org/datacube?units=10154984&cellref=TOT\\_POP:now&yearfrom=1841&year=1841](http://testapi.pastplace.org/datacube?units=10154984&cellref=TOT_POP:now&yearfrom=1841&year=1841)

```
@prefix sdmx-subject: <http://purl.org/linked-data/sdmx/2009/subject#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix obs: <http://obs.gbghgis.geog.port.ac.uk/uri/#> .
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix gbghgis: <http://gbghgis.geog.port.ac.uk/> .
@prefix admingeo: <http://data.ordnancesurvey.co.uk/ontology/admingeo/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<http://dataset.gbghgis.geog.port.ac.uk/uri/#TOT_POP:now>
  a qb:dataset ;
  gbghgis:hgisMeaningDDI "TOT_POP:now" ;
  dc:publisher "gbghgis" ;
  dc:subject <subjects> ;
  dc:title "TOT_POP:now" ;
  qb:slice <http://gbghgis.geog.port.ac.uk/TOT_POP:now/1841> .

obs:Observation-18283158
  a qb:Observation ;
  gbghgis:ref-auo "10154984" ;
  gbghgis:ref-auo-type "PR_DIST" ;
  gbghgis:ref-period "1841" ;
  qb:dataset "http://dataset.gbghgis.geog.port.ac.uk/uri/#TOT_POP:now" ;
  sdmx-measure:obsValue "53058"^^<http://www.w3.org/2001/XMLSchema#decimal> .

<http://dataset.gbghgis.geog.port.ac.uk/uri/#refArea>
  a rdf:Property ;
  rdf:PropertyConcept <http://dataset.gbghgis.geog.port.ac.uk/uri/#refArea> ;
  rdfs:label "reference area" ;
  rdfs:range
    "http://data.ordnancesurvey.co.uk/ontology/admingeo/UnitaryAuthority" ;
  rdfs:subPropertyOf "http://data.ordnancesurvey.co.uk/ontology/admingeo/refArea" ;
  qb:DimensionProperty <http://dataset.gbghgis.geog.port.ac.uk/uri/#refArea> ;
  qb:concept "http://purl.org/linked-
data/sdmx/2009/concept#UnitaryAuthority" .

<http://gbghgis.geog.port.ac.uk/TOT_POP:now/1841>
  a qb:slice ;
  gbghgis:hgisMeaningDDI "TOT_POP:now" ;
  gbghgis:ref-period "1841" ;
  qb:Observation "http://obs.gbghgis.geog.port.ac.uk/uri/#Observation-18283158"
;
  qb:sliceStructure "http://gbghgis.geog.port.ac.uk/sliceByAUO" .
<subjects> sdmx-subject: "1.1" .
```

**Figure 2** Sample output from prototype PastPlace datacube API

#### 4. Subsetting mechanisms

The underlying database structure makes dumping out data values straightforward, but current software just outputs the whole data table in the above format – “big data” with a vengeance. The challenge is to provide useful sub-setting mechanisms, and here we will draw on two distinct mechanisms already developed for the Vision of Britain download sub-system ([www.VisionOfBritain.org.uk/data](http://www.VisionOfBritain.org.uk/data)).

The first and spatial strategy starts with the client specifying a point coordinate and a broad statistical theme. The system returns a list of specific reporting units whose boundary polygons cover the point, and associated nCubes within the theme. The client can seek additional information about both units and nCubes, then extract data values for selected unit/nCube combinations, so obtaining local time series.

The second strategy starts with the client reaching an nCube by moving down the topic hierarchy or searching by keyword, the results being relevance-ranked based on where and how frequently the search term appears in the various metadata elements linked to the nCube. The system returns the “unit types”, effectively GIS coverages, for which data exist within the nCube; the dates for which data exist; and the number of data values for each type/date combination. Extracting all data for specific combinations essentially populates a statistical map.

## **5. Conclusion**

Current funding does not extend to developing client software. However, we get many data requests from academics, the media and government, with current projects for the EU and Greater London. We hope that providing programmatic access will let them start exploring the analytic potentials of a “big” and genuinely integrated dataset spanning all Britain’s localities over two hundred years.

## **6. Acknowledgements**

This research is part of an Arts and Humanities Research Council Big Data Project (award AH/L01002X/1).

## **7. Biography**

Humphrey Southall is Professor of Historical Geography and Michael Stoner is Senior Research Associate at the University of Portsmouth.

## **References**

- Cyganiak R, Reynolds D and Tennison J (2014) The RDF Data Cube Vocabulary. World Wide Web Consortium, Cambridge MA (<http://www.w3.org/TR/vocab-data-cube/>)
- Kramer S, Leahey A, Southall H R, Vampras J and Wackerow J (2012) *Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model*. Working Paper. Data Documentation Initiative, Ann Arbor, Michigan.
- Maali F, Cyganiak R, and Peristeras V (2012) A Publishing Pipeline for Linked Government Data, in Semperl E et al (eds) *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference*, 778-92. Springer, Berlin.
- Meroño-Peñuela A, Guéret C, Ashkpour A and Schlobach S (2014) CEDAR: The Dutch Historical Censuses as Linked Open Data. *Semantic Web Journal* (under review but online: <http://www.semantic-web-journal.net/system/files/swj878.pdf>).
- Southall H R (2008). Visualization, data sharing and metadata, in Dodge M, McDerby M and Turner M (eds) *Geographical Visualization: Concepts, Tools and Applications*, 259-75. Wiley, Chichester.
- Southall H R (2011) Rebuilding the Great Britain Historical GIS, Part 1: building an indefinitely scalable statistical database. *Historical Methods*, 44 (3), 149-59.
- Southall H R (2012) Rebuilding the Great Britain Historical GIS, part 2: a geo-spatial ontology of administrative units. *Historical Methods*, 45 (3), 119-34.

# GIS, Big Data and Lessons from John Snow

Doug Specht<sup>\*1</sup>

<sup>1</sup>University of Westminster, Communication and Media Research Institute.

January 07, 2015

## Summary

This paper examines the work of Snow within the ultra-modern context of big data and GIS, and questions the results that may be born from GIS and Big Data alone. Arguing that while GIS and spatial research have a great potential for unearthing trends, caution must be taken to ensure we do not generate dangerously misleading information about geographical and sociological connections.

The paper concludes that when reflecting on contemporary GIS and Big Data practice, that we should look to the work behind Snow's map, rather than being besotted by his famous geographic visualisation.

**KEYWORDS:** Big Data, GIS, John Snow, data visualisation, data analysis

The Big Data age is unquestionably here. In recent years, the volume of data collected and stored by business and government organizations has snowballed. Driven by reduced costs of storage and ever increasing analysis capabilities Big Data has become a big industry. The benefits of Big Data have also been widely reported, the McKinsey Global Institute (MGI) has cited examples of the transformative effect of Big Data from sectors as dispersed as health care to retail to manufacturing to political campaigns (Manyika *et al.*, 2011). Research conducted at the Massachusetts Institute of Technology shows that companies that use "data-directed decision making" enjoy a 5% – 6% increase in productivity, and that mass analysis of mobile phone calling patterns can help detect flu outbreaks (Boyd, & Crawford, 2012; Kirkpatrick, 2013). Big Data has also been put to use in humanitarian efforts, researchers from Sweden's Karolinska Institute and Columbia University have used data from Digicel, Haiti's largest cell phone provider, to determine the movement of displaced populations after the 2010 earthquake, aiding in the distribution of resources; Tweets in Indonesia have been analysed to predict how people fare with food price volatility, allowing pre-emptive measures to be put in place to reduce scarcity shock. And social media output analysis in the United States and Ireland have shown to be good early indicators of spikes in unemployment (Kirkpatrick, 2013).

It is the combination of this rich data source, collected through mobile devices, coupled with the Big Data Paradigm - which promises to turn ever larger and imperfect, complex, often unstructured data into actionable information, within a discourse of increased speed, efficiency and inclusivity (Hilbert, 2013; Burns, 2014) - that presents exciting opportunities for geographers, GIS analysts and the GI industry as a whole. By 2020, more than 70 percent of mobile phones are expected to have GPS capability, up from 20 percent in 2010 (Hilbert, 2013), leading to a massive increase to the flood of spatially located data already generated every day. The Big Data Paradigm suggests that the more data we have, the better our predictive modelling and analytics will be. Caution, however, must be taken in our rush to exploit these new vast pools of data and information for predictive analysis.

The evidence for a more nuanced approach to combining Big Data and GIS comes from a surprising point in mapping history, namely the unpicking of the mythology surrounding John Snow and his infamous 19<sup>th</sup> Century Cholera maps. Through an exploration of John Snow's methods and the

---

\* [doug@specht.co.uk](mailto:doug@specht.co.uk)

application of the principles of critical, participatory and feminist GIS, this paper seeks to ground our understanding of the value of Big Data in GIS and remind us of the underlying principles of GIS analysis.

Snow's study of the "Broad Street outbreak" has long been heralded as the start of spatioanalytical research and is oft cited as a fundamental example of epidemiology and medical geography (Koch & Denike, 2009). Our fascination with the map however, has somewhat distorted our understanding of the methods employed in its creation, and the conclusions that were drawn at the time. Two misconceptions persist around Snow's maps that have implications for how we merge GIS and Big Data. Firstly, that it was his maps that led Snow to reach the conclusion that Cholera was water born. And Secondly, that his maps provided good evidence for this conclusion.

When Snow first presented his hypothesis to the parish officials that the water pump maybe the source of the Broad Street outbreak he neither presented a map, nor did he allude to the idea that a map had been instrumental in his discovery (Bordy *et al.*, 2000). Moreover, the first edition of *On the mode of communication of cholera*, published in 1849, contained no maps, it was not until 1854 that his spot map was first published, possibly due to the influence of Shapter, whom Snow had cited in *OMCC*'s second edition. It would appear then that Snow had developed and tested his hypothesis well before he drew his map. This is not an unlikely scenario given that he was already engaged in an ambitious study of cholera in South London. It was likely these earlier studies that led him to conclude that a sharp localised outbreak pointed to a contaminated pump rather than, as commonly reported, an induction arrived at primarily from the geographical facts of the case (Brody *et al.*, 2000). Snow's map then did not give rise to the insight, but was the tool used to confirm and illustrate an already held hypothesis and conclusion.

More widely known about Snow's work is that, despite his body of evidence, he had a hard time convincing those around him that the water pump was the source of contamination, leading to the emotive myth of him striding into Broad Street and breaking off the water pumps handle. Snow's contemporaries, his readers and the parish had been unconvinced by his arguments. It was not so much his theory to which they objected, but rather his lack of detailed consideration of other potential sources of contagion (Koch & Denike, 2009). It was this lack of support for his ideas that led Snow to draw his maps, to prove and illustrate his theory. Simply plotting deaths on a map, however, did not lead others to reach the same conclusions, nor the immediate, unquestioning adoption of his theory (Brody *et al.*, 2000). "*On examining the map given by Dr Snow, it would clearly appear that the centre of the outburst was a spot in Broad-street, close to which is the accused pump; and that cases were scattered all round this nearly in a circle, becoming less numerous as the exterior of the circle is approached. This certainly looks more like the effect of an atmospheric cause than any other*" was the conclusion reached by Edmund A Parkes in his review of *On the mode of communication of cholera* (Parkes, 1855: 458 cited in Brody *et al.*, 2000). Indeed spot maps such as Snow's had previously been used by both contagionists and anticontagionists to advance their stance in Yellow Fever research as early as 1798 (Brody *et al.* 2000). Snow's map alone was not enough to convince either his contemporaries or other parties as to his, albeit correct, theory.

What can we learn from Snow and his approach in the age of GIS and Big Data, and how to we integrate this in our work? The trap of examining Snow's work is in positioning him as a mapmaker without an appreciation of Snow the thinker. Data are meaningful because of how someone collects, interprets, and forms arguments with it. Data are not neutral. The goal of Snow's maps was not to produce data, but to filter data, to reframe his hypothesis. Snow's original theorem was discounted, his map was misinterpreted, but together they were able to support each other, leading to the saving of many lives. The Big Data age is moving us into a new epistemology in which knowledge politics are deeply embedded in what counts as data, how those data are represented and the proposes for which those representations may be used. Evidence has shown that we will continue to see marginal increases in predictive performance even to a massive scale (Junqué de Fortuny, 2013), but these increases are useless if we are representing this data through a epistemology that embodies an unquestioned form of legitimacy and power, such as that often given to map-based artefacts, and which leads to wide

misinterpretation or which is based on unsound hypothesis. Big Data, plotted in the form of a highly sophisticated and accurate map, may easily seduce us into concluding that we have learned something.

Big Data should be considered as less a physical entity and more as a paradigm shift. It won't solve issues on its own, but it allows us to think differently. Big Data thinking opens our view to non-traditional data for predictive analytics, but we must be careful as we embrace this technological idea not to become besotted by it. Key to the integration of Big Data as a meaningful tool of analysis is ensuring that Big Data is viewed as much as a social innovation as a technical one. The future of GIS isn't about pulling mass amounts of data towards the centre like a black hole, it's about pushing for simple, purposeful and considered data collection leading to better collaboration, analysis and understanding. While our understanding of Snow's map production and methodology might be skewed by myth and legend, the truth of the story is still a shining example of how to blend mapped data, visualisation and theory. The Broad Street maps and Snow's work demonstrate that the mere act arranging data graphically in space does not yield new understanding without the support of a solid, considered and researched theory, and that our theories will be widely dismissed if we have not explored all hypotheses and represented them appropriately, regardless of the size of our Big Data set.

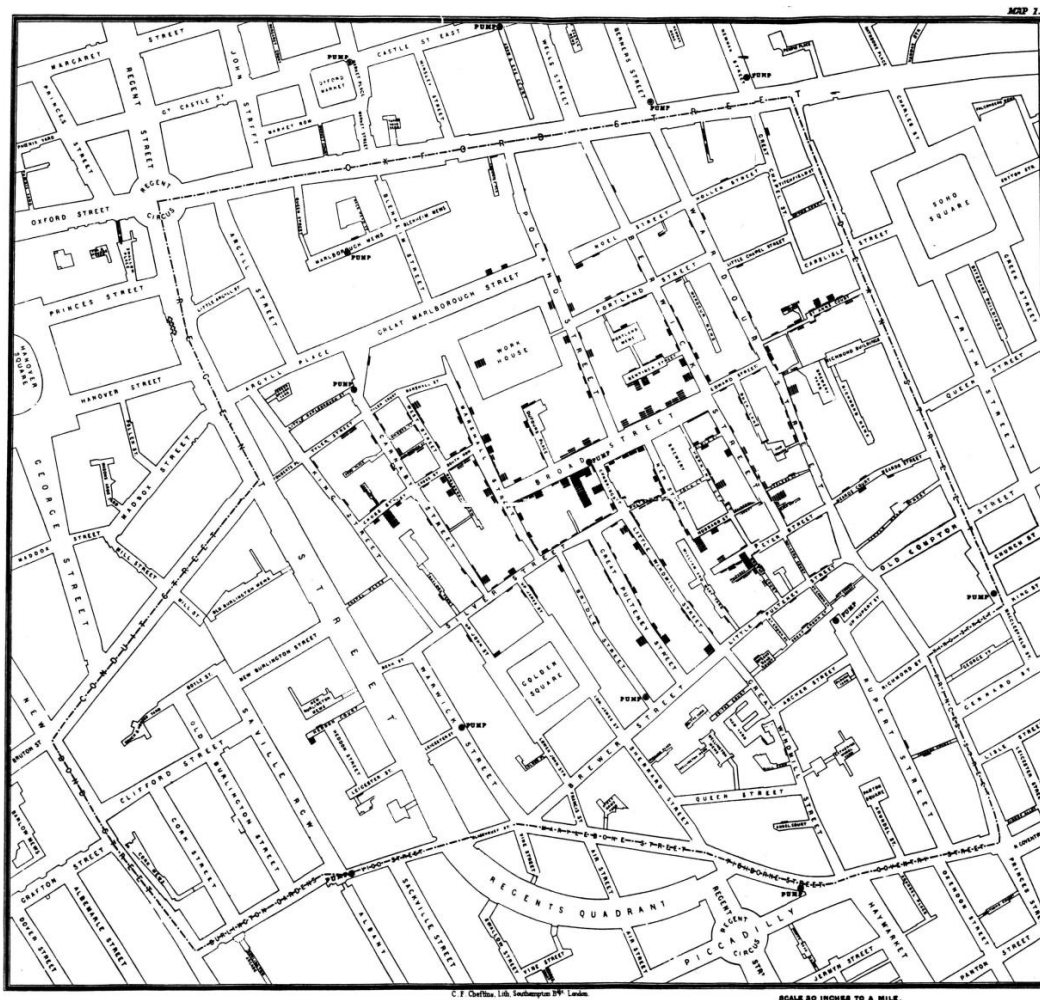


Figure 1: Original map made by John Snow in 1854. Cholera cases are highlighted in black. Published by C.F. Cheffins, Lith, Southampton Buildings, London, England, 1854 in Snow, John. On the Mode of Communication of Cholera, 2nd Ed, John Churchill, New Burlington Street, London, England, 1855. This image is in the public domain due to its age

## References

- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662-679.
- Brody, H., Rip, M. R., Vinten-Johansen, P., Paneth, N., & Rachman, S. (2000). Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, 356(9223), 64-68.
- Burns, R. (2014). Rethinking big data in digital humanitarianism: practices, epistemologies, and social relations. *GeoJournal*, 1-14.
- Hilbert, M. (2013). Big data for development: From information-to knowledge societies. Available at SSRN 2205145.
- Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive Modelling With Big Data: Is Bigger Really Better?. *Big Data*, 1(4), 215-226.
- Kirkpatrick, R. (2013). Big Data for Development. *Big Data*, 1(1), 3-4
- Koch, T., & Denike, K. (2009). Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854. *Social science & medicine*, 69(8), 1246-1251.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

## Biography

Doug Specht is a Doctoral Researcher at the University of Westminster where he is exploring how digital media and GIS are used in legitimizing and codifying local knowledge within the context of International Development. He is also the Director of VOZ, a PGIS platform that supports human and environmental rights.

# The Impact of Task Workflow Design on VGI Citizen Science Platforms

James Sprinks<sup>\*1</sup>, Jeremy Morley<sup>†1</sup>, Steven Bamford<sup>‡2</sup> and Robert Houghton<sup>§3</sup>

<sup>1</sup>Nottingham Geospatial Institute, University of Nottingham, UK

<sup>2</sup>School of Physics and Astronomy, University of Nottingham, UK

<sup>3</sup>Human Factors Research Group, University of Nottingham, UK

November 4, 2014

## Summary

Citizen science platforms allow non-scientists to take part in scientific research across a range of disciplines, and often involve the collection of volunteered geographic information from remotely sensed imagery. What these systems ask of volunteers varies considerably in terms of task type, level of user judgement required and user freedom. This work studied the Zooniverse's Planet Four project and investigated the effect of task workflow design on user engagement and outputs. Results show participants found the more guided, less-autonomous interface more frustrating, while the less complex, repetitive interface resulted in greater data coverage.

**KEYWORDS:** Citizen Science, Volunteered Geographic Information, Planetary Science

## 1. Introduction

Citizen science, or “public participation in scientific research” (Hand, 2010), can be described as research conducted, in whole or in part, by amateur or nonprofessional participants often through crowd-sourcing techniques. It increasingly utilises virtual citizen science (VCS) platforms (Reed *et al.*, 2012) that gather volunteered geographical information (VGI) from remotely sensed imagery, both of the Earth and other solar system bodies, through a website. As citizen science is a relatively new area of work, and while there has been research into human-computer interaction (HCI) design and functionality (Prestopnik & Crowston, 2012), there has been relatively little attention paid specifically to human factors issues regarding the collection of VGI. This comprises a significant research gap, given that the success of a citizen science venture is directly related to its ability to attract and retain users, both to gather the large amount of data required, and to ensure the utility of the data collected (Prather *et al.*, 2013). In this study we make a first step in considering how virtual citizen science systems can be better designed for the needs of the volunteer, exploring whether manipulating task flow would affect both the information collected, as well as the volunteers' experience of user the interface.

Some studies have considered motivation amongst citizen science volunteers (Raddick *et al.*, 2010, Reed *et al.*, 2013) but have not considered in any depth the form of work itself. This may be considered remiss as over thirty years of human factors research has identified a relationship between motivation, satisfaction and work design. Hackman & Oldham (1975) developed the “Job Diagnostic Survey” in order to better understand jobs and how they could be re-designed to improve motivation

---

\* James.sprinks@nottingham.ac.uk

† Jeremy.morley@nottingham.ac.uk

‡ Steven.bamford@nottingham.ac.uk

§ Robert.houghton@nottingham.ac.uk



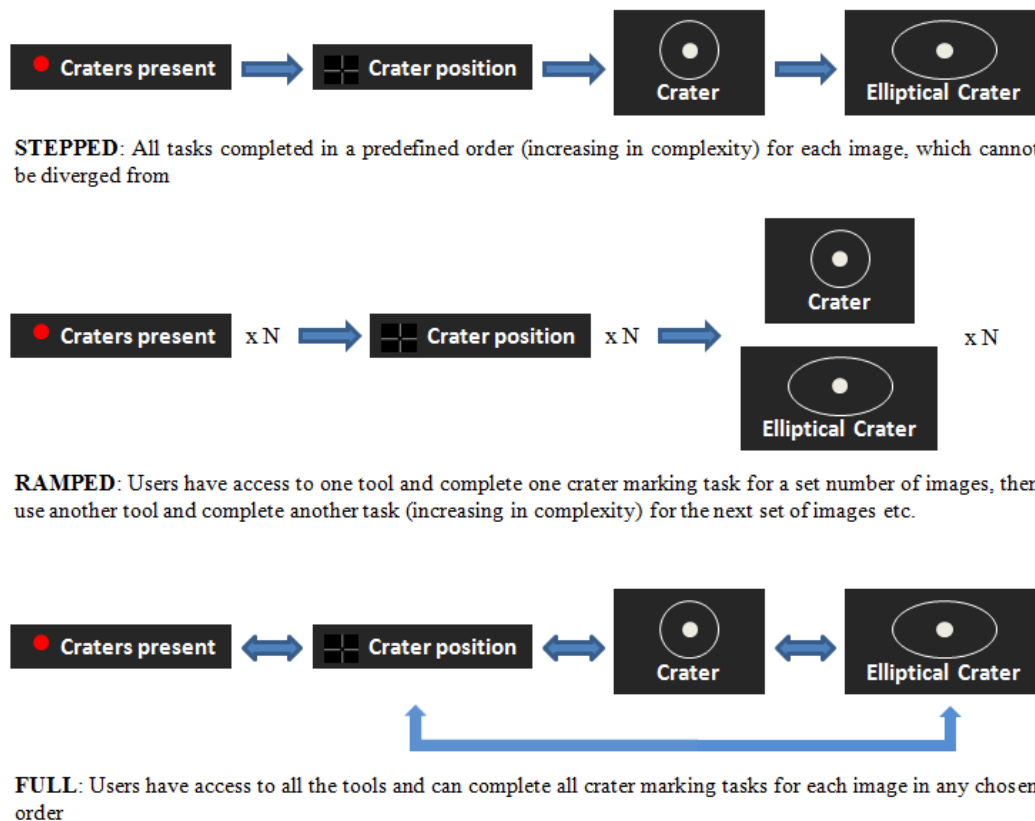
and productivity. Factors such as task variety, complexity and autonomy were identified as key to this process, all of which can be influenced in VCS design.

This paper describes the current active Zooniverse site Planet Four, a project that allows volunteers to mark the positions and directions of seasonal fans on the planetary surface, presenting the results of three different iterations of the site that differ in task workflow design, in terms of user experience/satisfaction and scientific output. Finally the impact of task workflow design on these results, and the implications for VCS platforms and other online mechanisms, are discussed.

## 2. Methodology

In order to investigate the effect of task workflow design on user experience and VGI output, a new version of the Zooniverse's Planet Four project has been developed. The new site allows users to mark craters on images of the Martian surface. A lab study has been carried out to both consider task workflow factors and also act as a technical test, identifying any general functionality and usability issues before a public launch.

The platform has been developed to include three different interfaces for marking craters that vary in task type, number of tasks available to the user and user freedom. The crater-marking task has been split into four tools per crater: to indicate if any craters are present; to mark the crater centre; to draw a circle around the centre; and to adjust a second radius to make an ellipse. Figure 1 shows the tools available to each user and when for each interface, namely stepped, ramped and full:



**Figure 1** Flow Diagram of Tools Available to User for each Interface

Thirty participants took part in the lab study between January and March 2014. There were no specific prerequisites for participation. Each participant used each interface in a random order, and afterwards completed a questionnaire asking them to share their views across themes including *design & usability*, *tasks & tools* and *imagery*. At the end of each section, ‘free text’ boxes were available for participants to give additional comments and opinions.

### 3. Experimental Results

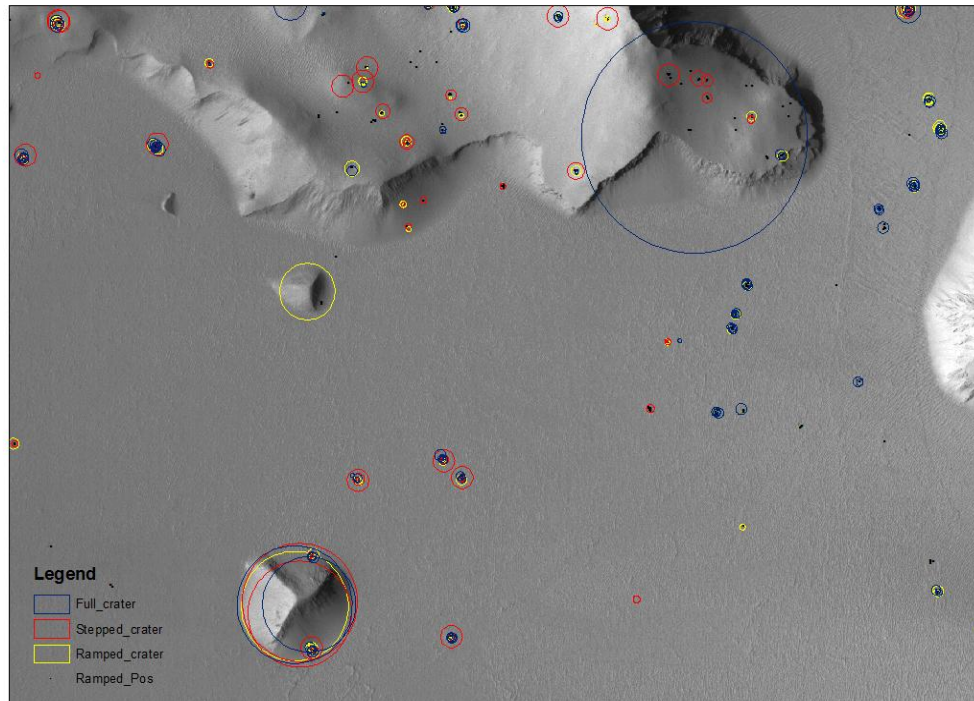
#### 3.1. Participant Questionnaire Results

Table 1 shows a breakdown of the number of comments made by participants regarding each interface, sorted into four different topics. Usability comments were concerned with the general usability and mechanics of each interface; accuracy comments focussed on how accurately craters could be marked; tool issue comments were specifically about the tools provided to mark craters; and imagery comments discussed the remotely sensed imagery displayed.

**Table 1** Numbers of Responses by Comment Topic and Interface

Topic	Full Interface	Stepped Interface	Ramped Interface
Usability	3	9	4
Accuracy	6	5	7
Tool Issues	7	6	6
Imagery	3	2	2

#### 3.2. Crater Marking Results



**Figure 2** Participant Markings using each Interface, Full (blue), Stepped (red) & Ramped (yellow)

Table 2 shows the crater marking results for each interface, in terms of the number of crater clusters identified (craters marked by more than one participant), the average standard deviation of the cluster centre position and the average standard deviation of the cluster crater diameter in terms of screen pixels (i.e. the level of agreement between participants).

**Table 2** Crater Marking Results for each Interface

	Full Interface	Stepped Interface	Ramped Interface
No. of Crater Clusters	182	185	298
Stand. Dev. of Position	$2.64 \pm 0.35$	$2.41 \pm 0.46$	$6.28 \pm 2.69$
Stand. Dev. of Diameter	$6.52 \pm 0.57$	$7.70 \pm 0.57$	$5.10 \pm 0.44$

While participants using the full and stepped interface have identified similar numbers of craters (182 and 185 respectively), it is clear that the ramped interface has resulted in more craters being marked (298, ~61% greater). Similarly, the standard deviation of the central positions is comparable between the full and stepped interface ( $2.64 \pm 0.35$  and  $2.41 \pm 0.46$  pixels respectively), whilst the ramped interface has an average standard deviation of  $6.28 \pm 2.69$  pixels. Regarding the standard deviation of crater diameter, the stepped interface is highest at  $7.70 \pm 0.57$  pixels, followed by the full at  $6.52 \pm 0.57$  and ramped at  $5.10 \pm 0.44$ .

#### 4. Discussion and Conclusions

This study found that altering the task workflow design of the interface can have an effect both on the user experience and on the resulting VGI data. Regarding the topics of accuracy, tools and imagery, participant comments are comparable in number and predominantly concern the difficulty of marking small craters across each interface. This is perhaps as expected as this issue is more related to the imagery displayed (which is constant in this study) rather than the interface used. However when considering usability, participant comments were much greater in number for the stepped interface and predominantly negative regarding the restriction of choice, as explained by participant S19:

*“I don't like to be forced to use a certain task order, and I couldn't go back or switch tools...”*

This is in agreement with Hackman & Oldham's findings which suggest that both variety and autonomy are important in ensuring greater job satisfaction.

When considering VGI output, again task workflow design had an effect. The ramped interface resulted in a much higher number of crater clusters being identified, but less agreement in their central position. This is an important result, as reducing the number of null returns (images with no markings) would in turn reduce the time spent on data reduction by the science team, however the greater range of marked position would require extra consideration.

When considering task workflow design, future citizen science platforms involving VGI and remotely sensed imagery will need to perform a balancing act, weighing up the importance of user satisfaction, the data needs of the science case and the resources that can be committed both in terms of time and data reduction, more than likely on a case-by-case basis.

#### 5. Acknowledgements

The first author is supported by the Horizon Centre for Doctoral Training at the University of Nottingham (RCUK Grant No. EP/G037574/1) and by the RCUK's Horizon Digital Economy Research Institute (RCUK Grant No. EP/G065802/1).

## 6. Biography

*James Sprinks* is a PhD candidate at the University of Nottingham's Horizon Doctoral Training Centre. He is researching how citizen science platforms, and the 'user' tasks associated with them, can be designed to ensure that the data generated is scientifically robust, while maintaining a user experience that builds the community.

*Jeremy Morley* is the Geospatial Science theme leader at the Nottingham Geospatial Institute at the University of Nottingham. His research has included the quality of VGI data; presentation of GIS data online; and planetary mapping, particularly of Mars.

*Dr Robert Houghton* is a Horizon Transitional Fellow and a member of the Human Factors Research Group at the University of Nottingham. He is interested in exploring ways in which Crowdsourcing and Human Computation can be optimised to improve the quality and utility of contributions made by citizen science.

*Dr Steven Bamford* is a STFC Advanced Fellow and a member of the Centre for Astronomy and Particle Theory at the University of Nottingham. His research focuses on galaxy morphology, structure and environment. He is involved several collaborative projects that involve the coordination and utilisation of Citizen Science.

## References

- Hackman J R and Oldham G R (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60(2), 159-170.
- Hand E (2010). Citizen Science: People Power. *Nature*, 466(7307), 685-687.
- Prather E E, Cormier S, Wallace C S, Lintott C, Raddick M J, Smith A (2013). Measuring the Conceptual Understandings of Citizen Scientists Participating in Zooniverse Projects: A First Approach. *Astronomy Education Review*, 12(1), 010109.
- Prestopnik N R and Crowston K (2012). Citizen Science Assemblages: Understanding the Technologies that Support Crowdsourced Science. *Proceedings of the 2012 iConference*, 168-176.
- Raddick M J, Bracey G, Gay P L, Lintott C J, Murray P, Schawinski K, Szalay A S, Vandenberg J (2010). Galaxy Zoo: Exploring Motivations of Citizen Science Volunteers. *Astronomy Education Review*, 9, 010103-1.
- Reed J, Raddick M J, Lardner A, Carney K (2013). An Exploratory Factor Analysis of Motivations for Participating in the Zooniverse, a Collection of Virtual Citizen Science Projects. *Proceedings of the 2013 46<sup>th</sup> Hawaii International Conference on System Sciences*, 610-619.
- Reed J, Rodriguez W, Rickoff A (2012). A Framework for Defining and Describing Key Design Features of Virtual Citizen Science Projects. *Proceedings of the 2012 iConference*, 623-625.

# Assessing the risk landslides pose to road and rail networks

Taalab K P<sup>\*1</sup> and Cheng T<sup>†1</sup>

<sup>1</sup> SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, Chadwick Building, Gower Street, London WC1E 6BT, United Kingdom

November 5, 2014

## Summary

Road and rail networks are critical infrastructure, vital for ensuring the flow of essential goods and services necessary to maintain a country's economic and national security. Landslides are a natural hazards which can seriously affect these networks, so in order to plan mitigation strategies, calculate losses and minimise casualties, it is necessary know the risk a posed by landslides. Using a case study of Piedmont, Italy, this study proposes an empirical modelling approach to the quantification of landslide risk using support vector machines and simple network analysis.

**KEYWORDS:** Landslide, Infrastructure, Support vector machine, Hazard, Risk

## 1. Introduction

Road and rail networks are critical infrastructure, vital for ensuring the flow of essential goods and services necessary to maintain a country's economic and national security. Landslides are a natural hazards which can seriously affect road and rail networks, so in order to plan mitigation strategies, calculate losses and minimise casualties, it is necessary know the risk a posed by landslides. The identification of risk can be deconstructed into a number of constituent parts. Varnes (1984) proposes that risk can be calculated using the formula:

$$Risk = (susceptibility \times trigger) \times (vulnerability \times exposure) \quad (1)$$

where  $susceptibility \times trigger = hazard$ . Hazard susceptibility is a relative measure of the spatial likelihood of the occurrence of landslides (Pourghasemi et al., 2013). It can be determined as a function of terrain attributes (e.g. slope, aspect) and environmental variables (e.g. geology, soil and land use). Susceptibility alone cannot give the probability of landslide occurrence. For this, it is necessary to model the relationship between events which trigger landslides and landslide occurrence. There are many potential triggers including precipitation, earthquakes and human activity, with heavy rainfall being the most common trigger (Cepeda et al., 2010). As an event in itself, a landslide does not pose any risk. The risk comes from the exposure of people or elements of the build environment. When assessing landslide risk at a European scale, Jaedicke et al. (2014) used population density and the density of road and rail networks as metrics to represent exposure. Exposure, and therefore risk, increases when landslides occur in areas of high population or infrastructure density.

Vulnerability is a measure of the potential degree of loss. It is a complex concept as, for example, a well-designed structure could be seen to be less vulnerable than a poorly designed structure, as it is less likely to be damaged in the event of a landslide. In economic terms, however, the well-designed structure may be significantly more valuable than the poorly designed structure, which would make any repairs more costly, meaning economic loss is higher if it is damaged. Vulnerability can also be

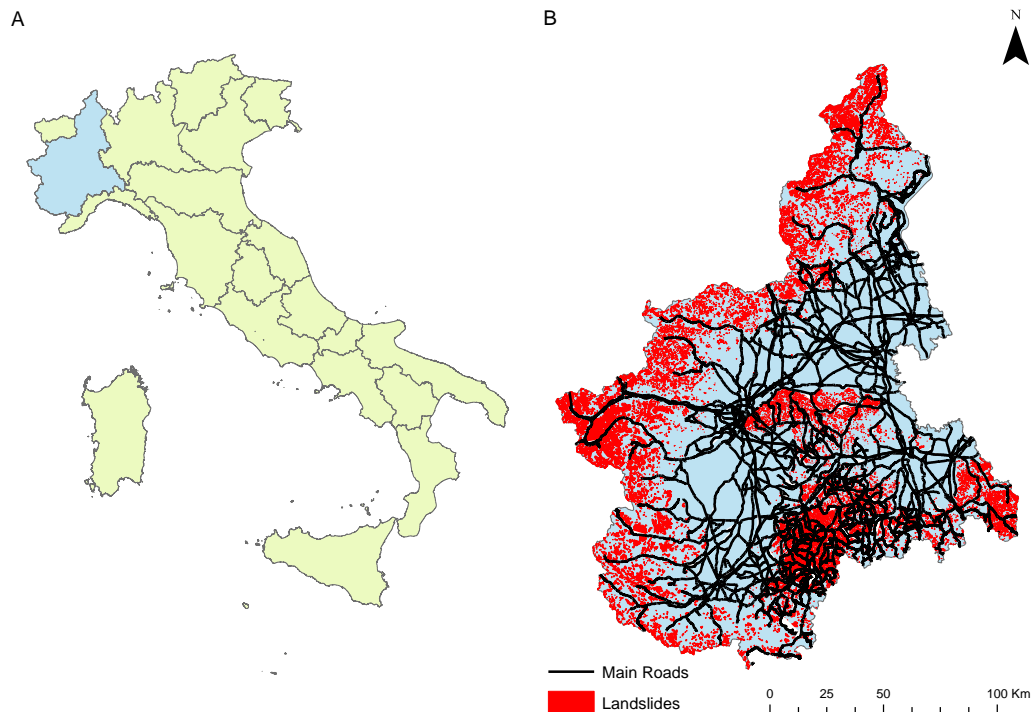
---

<sup>\*</sup> k.taalab@ucl.ac.uk

<sup>†</sup> tao.cheng@ucl.ac.uk

considered in terms of individuals, communities and regions, which makes this a difficult metric to measure, especially at the national or international scale.

This study aims to assess the risk that landslides pose to the road and rail network using a case study of Piedmont, Italy (Figure 1A). Italy has been identified as having the greatest extent of infrastructure which is exposed to landslide hazards in the whole of Europe. The Piedmont region is a particularly apt case study as it is within a ‘landslide hotspot’- an area of Europe where hazard and risk are greatest (Jaedicke et al., 2014). This study proposes some amendments to the Varnes (1984) risk model as it is applied to landslides. Firstly, that landslide susceptibility should be categorised in terms of the types of landslide that can occur. There are a number of different classes of landslide, based on mass movement characteristics. Each of these will require different mitigation strategies, as well as having different triggers (Cruden, & Varnes, 1996). Secondly, the vulnerability of the road and rail network can be considered in terms of the importance of the road or rail link to the network as a whole. As well as the density of the road and rail network, it is important to assess how critical the exposed parts of the network are. For example, parts of the network are used more frequently can be considered more vulnerable than those less frequently used.



**Figure 1** A) Location of Piedmont B) the location of previous landslides and main roads

## 2. Materials and Methods

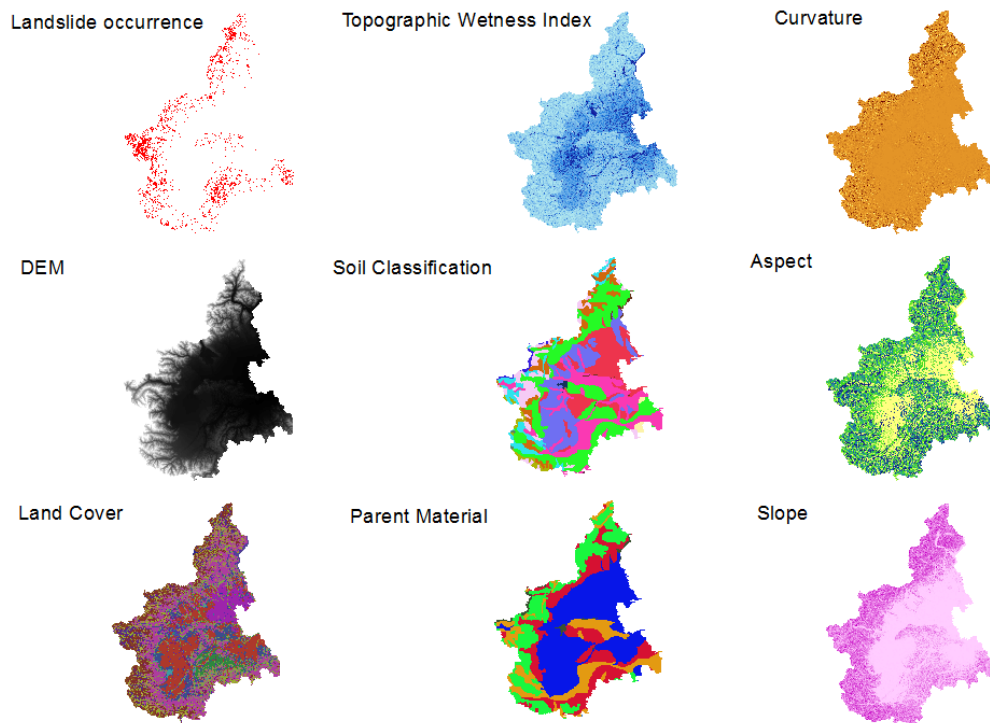
### 2.1 Input data

To develop empirical susceptibility maps, an inventory of previous landslide activity is required. This study uses SIFRAP (the landslide inventory of the Piedmont region) (Lanteri & Colombo, 2013). This dataset contains shapefile records showing the spatial extent of over 30000 previous landslides (Figure 1B). SIFRAP also categorises landslides, a summary of this classification is shown in Table 1.

**Table 1** SIFRAP landslide classification

Classification	Description
Crash / Rollover	The mass moves mainly in the air, for free fall, for jumps and rebounds to rolling, shattering into different elements of variable size, and is generally characterized by extremely quick motion
Expansion	An extension movement of cohesive soil or rock, combined with a general subsidence of the mass itself, which fracture and dismantles into several parts, above a soft material, not cohesive
Slow dripping	Movements are generally characterized by low speed and involving soils with high clay content and mostly low water content affecting not very steep slopes
Fast dripping	High speed, affecting mostly loose soils in the presence of significant water content. It is triggered as a result of heavy rainfall and usually involve the loose soil cover on steep slopes
Sliding rotational / translational	Movement along one or more surfaces, where the shear strength is exceeded, or within a zone characterized by relatively thin, intense shear deformation
DGPV	Very complex deformation which occurs through a mostly slow and progressive rock mass, without any appreciable continuous failure surfaces. The process deformation occurs extremely slowly

Other environmental variables used to determine landslide susceptibility are shown in Figure 2. The environmental metrics were selected based on the recommendations of previous empirical studies on landslide susceptibility (Dai & Lee, 2003; Wand & Sassa, 2006; Bui et al., 2013). The 100 m resolution digital elevation model (DEM) was used to derive slope, aspect, curvature and TWI.



**Figure 2** Environmental covariates used for modelling landslide susceptibility

In order to create empirical models of landslide susceptibility, it is necessary to spatially sample all environmental variables at both locations where landslides have occurred previously and at locations where landslides have not occurred previously. This created a dataset of 400000 samples which randomly divided into a training dataset of 300000 samples and a validation dataset of 100000 samples. As well as sampling the values of the environmental variables, samples where landslides were assigned a value of 1 and the landslide class was recorded. Samples where no landslide had occurred were given the value 0.

## 2.2 Support vector machines (SVM)

SVMs are based on statistical learning theory (Vapnik, 1998). Originally developed as a binary classifier, SVMs perform classification (and regression) by constructing N-dimensional hyperplane that optimally separates data into categories (Hearst et al., 1998). For example, if we wish to separate data into two classes, we would like to find a threshold which could discriminate between the two. The simplest example of this would be a straight line in two-dimensional space, or a hyperplane in higher dimensional space. The SVM tries to find the optimal separating hyperplane that gives the largest separation between classes.

Often there is a situation where it is not possible to separate the classes with a straight line or hyperplane. This is where SVMs employ a kernel function (sometimes known as the 'kernel trick'). It is possible to project data which is not linearly separable into higher dimensional space where the data can then be separated by a hyperplane. By using kernel mapping, SVMs can operate in an arbitrary number of dimensions, making it possible to find hyperplane separating solutions for even highly complex datasets (Ballabio & Sterlacchini 2012). Despite the kernel trick, for very complicated datasets, it is usually difficult and not particularly desirable to entirely separate two classes, as this can lead to overfitting. To allow some flexibility, SVMs have a cost parameter C that determines the trade-off between misclassification and enforcing strict (ridged) margins- creating soft



margins that allow some misclassification. Increasing the C parameter increases the penalty for misclassifying points, creating a more rigid model.

In high dimensional feature space, support vector regression uses an  $\varepsilon$ -insensitive loss function to perform linear regression, while reducing model complexity by seeking to minimise  $\|w\|^2$  by solving the optimisation problem shown in Equation 1 (Cherkassky & Ma, 2004)

$$\begin{aligned} & \text{minimise } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ & \text{subject to } \begin{cases} y_i - f(x_i, \omega) - b \leq \varepsilon + \xi_i^* \\ f(x_i, \omega) + b - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \quad (2)$$

Where  $\|w\|$  is the norm of the normal hyperplane,  $\xi_i^*$  and  $\xi_i$  are slack variable which measure the deviation of the training data beyond the  $\varepsilon$ -insensitive zone and C is the cost parameter which regulates the relationship between model complexity and error. Overall, SVMs have two parameters which need to be set; the cost parameter C which determines the amount of generalisation and kernel function, which can be linear, polynomial, or Radial Base Function. In this study the kernel function was set as a radial base function due to its robustness (Kavzoglu et al., 2014). The parameter C was determined using the carat package in R and the SVM models were built using the R package e1071 (Meyer et al., 2012).

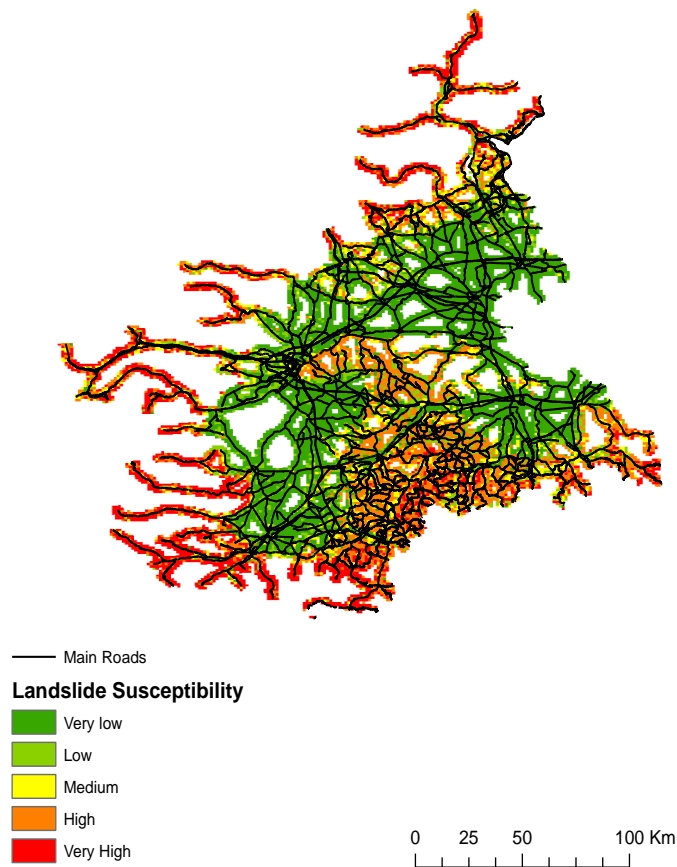
### 2.3 Network analysis

Simple network analysis uses the topology of the network to analyse the relative importance of each section of the network. This study proposes using betweenness centrality, which is a measure of how frequently a section of the network is part of the shortest path between points and is used to indicate the influence that a given section of the network will have on flow (in this case traffic flow).

In theory, the higher the centrality, the greater the effect on the network if this section of the network was to be closed, hence the greater the vulnerability in the risk model (Freeman, 1977). This study will used sDNA software for ArcGIS to derive centrality metrics for the road and rail network.

## 3. Results

Figure 3 shoes the initial result of the SVM landslide susceptibility modelling for Piemont. Given the distribution of pervious landslides, it is to be expected that the areas that show the highest susceptibility are in the west and south of the study area.



**Figure 3** Landslide susceptibility for the road network in Piedmont

#### 4. Future work

In order to quantify the risk landslides, there are a number of tasks which need to be completed

1. Validate the landslide susceptibility map (Figure 1) using test dataset
2. Create a landslide classification map using the SIFRAP data and SVM classification. The data used to train the model will be the environmental data which was sampled in areas where landslides have previously occurred. This will be split into training and validation datasets.
3. Create a normalised betweenness centrality classification for the road and rail network in Piedmont using sDNA software in ArcGIS. This can be used to assess the vulnerability of the network.
4. Using the landslide susceptibility maps as a base, attribute rainfall thresholds to landslide occurrence. Using SVM modelling to develop an empirical relationship between historic rainfall data (both antecedent rainfall and storm events) to identify landslide triggers (e.g. Li et al., 2010; Segoni et al., 2014)

## 5. References

- Ballabio C and Sterlacchini S (2012) Support vector machines for landslide susceptibility mapping: the Staffora River Basin case study, Italy. *Mathematical geosciences*, 44(1), 47-70.
- Bui D T, Pradhan B, Lofman O, Revhaug I and Dick Ø B (2013) Regional prediction of landslide hazard using probability analysis of intense rainfall in the Hoa Binh province, Vietnam. *Natural hazards*, 66(2), 707-730.
- Cepeda J, Höeg K and Nadim F (2010) Landslide-triggering rainfall thresholds: a conceptual framework. *Quarterly Journal of Engineering Geology and Hydrogeology*, 43(1), 69-84.
- Cherkassky V and Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1), 113-126.
- Cruden D M and Varnes D J (1996) Landslides: Investigation and Mitigation. Chapter 3-Landslide types and processes. *Transportation research board special report*, (247).
- Dai F C and Lee C F (2003) A spatiotemporal probabilistic modelling of storm-induced shallow landsliding using aerial photographs and logistic regression. *Earth surface processes and landforms*, 28(5), 527-545.
- Freeman L C (1977) A set of measures of centrality based on betweenness. *Sociometry*, 35-41.
- Hearst M A, Dumais S T, Osman E, Platt J and Scholkopf B (1998) Support vector machines. *Intelligent Systems and their Applications*, IEEE, 13(4), 18-28.
- Kavzoglu T, Sahin E K and Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides*, 11(3), 425-439
- Lanteri L and Colombo A (2013) The Integration Between Satellite Data and Conventional Monitoring System in Order to Update the Arpa Piemonte Landslide Inventory. In *Landslide Science and Practice* (pp. 135-140). Springer Berlin Heidelberg.
- Li C, Ma T and Zhu, X (2010). aiNet-and GIS-based regional prediction system for the spatial and temporal probability of rainfall-triggered landslides. *Natural hazards*, 52(1), 57-78.
- Meyer D, Dimitriadou E, Hornik K, Leisch F and Weingessel A (2012) *Package: e1071*. R package version 1.6.
- Pourghasemi H R, Jirandeh A G, Pradhan B, Xu C and Gokceoglu C (2013) Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. *Journal of Earth System Science*, 122(2), 349-369.
- sDNA version 1.13.4. Cardiff University. [www.cardiff.ac.uk/sdna](http://www.cardiff.ac.uk/sdna)
- Segoni S, Lagomarsino D, Fanti R, Moretti S and Casagli, N (2014) Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system. *Landslides*, 1-13.
- Varnes D J, The International Association of Engineering Geology Commission on Landslides and Other Mass Movements (1984) *Landslide hazard zonation: a review of principles and practice*. *Natural Hazards*, vol 3. United nations educational scientific and cultural organization [60 pages]

Vennari C, Gariano S L, Antronico L, Brunetti M T, Iovine G, Peruccacci S, Terranova O and Guzzetti F (2014) Rainfall thresholds for shallow landslide occurrence in Calabria, southern Italy. *Natural Hazards and Earth System Science*, 14(2), 317-330.

Vapnik, V (1998) Statistical Learning Theory. Wiley, New York, NY.

Wang H B and Sassa K (2006) Rainfall induced landslide hazard assessment using artificial neural networks. *Earth Surface Processes and Landforms*, 31(2), 235-247.

## **6. Acknowledgements**

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 603960

## **7. Biography**

Khaled Taalab is a research associate in the SpaceTimeLab at University College London. In 2013 he was awarded a PhD in digital soil mapping. He is currently working on a European FP7 project called InfraRISK which is developing models linking natural hazards, critical infrastructure and the environment. His research interests include environmental mapping, data mining and predictive modelling.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab (<http://www.ucl.ac.uk/spacetimelab>), at University College London. She has broad knowledge and experience in Geographic Information Sciences (GISc), from data acquisition, to information processing, management and analysis, with applications in environmental monitoring, natural resource management, health, transport and crime studies. She has over 140 publications.

# Comparing different spatial microsimulation frameworks

Tomintz M, Kosar B

Carinthia University of Applied Sciences  
Department of Geoinformation and Environmental Technologies

November 07, 2014

## Summary

Spatial microsimulation modelling was introduced in the 1980s but there is still a lack of open and easy to use frameworks. This obstacle was recognized by researchers and now there are frameworks available to access/download. The aim of this paper is to compare three different spatial microsimulation frameworks, as all of them have different algorithms implemented, that are available for free to the research community. The expected results are to find the best approach to model the Austrian smoking population for municipalities based on Austrian datasets and to list the pros and cons of the frameworks based on defined criteria.

**KEYWORDS:** comparing spatial microsimulation framework; smoking; simSALUD; Flexible Modelling Framework; spatial microsimulation with R

## 1. Introduction

Spatial microsimulation modelling is a novel method for creating small area estimates of data that are not available for small geographical scales. There are different approaches for microsimulation modelling but due to complexity, missing easy or free to use software frameworks, approaches can often not be used and therefore are implemented from scratch. There is an emerging need for comparing different spatial microsimulation methods (Clarke and Harding, 2013). Nowadays there are frameworks published for the community including handbooks and tutorials, so that people can use them.

The aim of this paper is to use and compare different spatial microsimulation frameworks, including FMF, Spatial Microsimulation with R and simSALUD, to find the best method for estimating the smoking population at municipality level for Austria. Criteria are defined for comparison, including preparing and import of input data files, speed of calculation, etc. The most interesting fact why doing this comparison is to see which method performs best for the existing input data.

## 2. Spatial microsimulation frameworks

For this paper, three spatial microsimulation frameworks are selected for comparison because they have the same aim but differ in the following facts: implemented algorithm, different programming and scripting languages respectively, the way of importing the input data and exporting the output data (simulated results), the access to the framework (installation, Web-based access), to name a few. The following subchapters describe each software framework in more detail based on the aforementioned criteria.

### **1.1. FMF (Flexible Modelling Framework)**

The flexible modelling framework (FMF) is a software application which was created by Kirk Harland and his members of the Multi Agent Systems and Simulation (MASS) research group (Harland et al. 2012). The FMF software framework has been developed in the Java programming language which presents a graphical user interface (GUI) for connecting spatial analysis and other tools to data. The FMF has initially been used for generating realistic populations of Leeds and is currently being used to examine trends and processes within retail markets. The framework itself requires one single flat (records with no structured relationships) survey file (e.g.: csv) as well as for each constraint one single aggregated data (e.g.: census) file. The used algorithm is an iterative optimisation algorithm simplified from simulated annealing and creates synthetic population for the areas that distributes individuals from the microdata in the population areas. No programming skills are required and the application can be run on every computer without installing the software. The application allows the users to validate the results using a various number of validation methods such as cell percentage error, percentage error, Standard absolute error, total absolute error, standardized root mean square error, etc.

The framework was also developed as a generic box of plugins which allows external users to create their own plugins to contribute it to the project for further development and testing. Subsequently other useful tools have been added, including a cluster hunting tool to identify clusters in geographical data.

### **1.2. Spatial Microsimulation with R**

Another approach to create a spatial microsimulation model was done by Robin Lovelace using the free software package R (Lovelace 2014). His decision of using the software R was the low-level language of R compared with other statistical programs based on a strong graphical user interface in combination with the great flexibility (many pre-made functions) for analysing and modelling the data. Bit high-level, compared to other general purpose languages such as C or Python. The model is based on the deterministic method to allocate individuals to areas called iterative proportional fitting (IPF). The tool can be used to either adapt the code on the input data of the user or to write it from scratch to learn the language R and the algorithm, as illustrated in Lovelace (2014). The tool requires as input data one individual-level dataset (csv format) and depending on the used function one aggregated area data or for each constraint one csv file. In general the framework requires to have some knowledge in the language R, at least for adopting the parameters for the own data. In contrast to the aforementioned framework, the software R offers packages to visualize the spatial data in R (Lovelace et al. 2014). Additionally the tool provides some integerisation methods to only allocate whole people to the result (Lovelace 2013).

### **1.3. simSALUD**

simSALUD is a spatial microsimulation framework which was created as part of the research project SALUD (SpatiAL microsimUlation for Decision support) at the Carinthia University of Applied Sciences (Tomintz et al. 2013). The application was designed as a web spatial microsimulation application and consists of three wizard-based modules: simulation, validation, visualization. Currently, the algorithm implemented within simSALUD is a deterministic reweighting approach after Ballas 2005, O'Donoghue 2013 with an extension to integerise (Ballas 2005) the model outputs. Based on a static microsimulation approach also this framework requires as input one non-spatial survey population file (csv) and for each constraint one single csv file containing demographic and socio-economic population data for small areas (e.g.: municipalities). As the framework FMF, simSALUD requires no programming skills and needs to have only an internet connection for using the application. All three steps are guided through a wizard and allow the user an easy-to-use handling. After running the model, also this framework provides the user to validate and verify the models robustness. Available validation methods are for example: total absolute error, percentage error, percentage error, simple regression, etc. One feature of the framework is that the outputs of the simulation can be either visualized in form of a map on the simSALUD web-based framework itself

or exported after the model run for further analyses in common geographic information software products.

### **3. Input data: case study “smokers”**

For this paper the case study “smokers” is chosen to do the comparison with the different software frameworks. The aim is to estimate the smoking population starting at the age of 15 at municipality level. Therefore two datasets are of interest: the Austrian Health Survey 2006/07 and the registered-based census 2011. One requirement for the modelling process is the availability of common variables in both datasets that predicts being a smoker best as possible. Therefore, statistical pre-analysis are performed on the Austrian Health Survey 2006/07 that holds more than 15.000 persons, i.e. chi-square and regression analysis. Finally four variables are selected as constraints for the spatial microsimulation modelling, i.e. age (eleven categories), sex (two categories), marital status (four categories) and last completed education (three categories).

The input data is accessible from the Statistics Austria noting that the micro-dataset (Austrian Health Survey 2006/07) can be requested free of charge but the registered-based census 2011 (demographic and socio-economic variables at municipality level) is with costs.

The advantage, however, is the common requirement of all three frameworks for the input data files, which are comma-separated values (csv) files. So there is no additional huge amount of data preparation when testing all three frameworks. The only difference is that the spatial microsimulation with R framework (see 1.2) requires all data in one file whereas the other two frameworks (see 1.1 and 1.3) require the input data in separate files separately for each file.

### **4. Expected results**

The case study for this paper will identify the simulated smoking population for Austria at municipality level. As there are different spatial microsimulation approaches available, different ones are getting tested to explore which one brings the best result based on the Austrian input data. The results are tested using statistical analysis. One main validation source is the Total Absolute Error (TAE) where the simulated constraints are compared with the registered-based data constraints, as this method is used within all three software frameworks. Further, the results from all three frameworks are getting mapped to identify if there are huge spatial variations.

Further, all three frameworks are getting compared based on the criteria mentioned in the introduction. During the analysis it is possible that further criteria are getting defined. This will show which spatial microsimulation approach works best for the Austrian datasets and also their pros and cons in terms of usage and simulation time.

### **5. Conclusion and future work**

This paper explores the best spatial microsimulation approach to model the smoking population in Austria among three approaches that are embedded in different frameworks. With it, the frameworks are tested based on certain criteria, including simulation speed, programming knowledge, etc. It is known that there is a lack of available open spatial microsimulation frameworks and therefore it is good to see that first attempts are being made. However, there is need for testing different approaches and frameworks under comparison of different criteria.

As the comparison is based on Austrian datasets only, the next planned step is to use cross-national datasets, e.g. from the UK. This dataset has then a different number of individuals from the health survey and a different number of areas that are modelled. The results are valuable for further

framework developments and to help people in their decision which framework fits best for their simulation aims.

## 6. Acknowledgements

This research project SALUD is funded by the Federal Ministry for Transport, Innovation and Technology (bmvit) and the Austrian Science Fund (FWF) (project number TRP280-G16). The registered based census data is accessed from “STATcube – Statistical Database of STATISTICS AUSTRIA”.

## 7. Biography

Melanie Tomintz is senior researcher and lecturer at the Carinthia University of Applied Sciences (CUAS), Department of Geoinformation and Environmental Technologies. Her main research interests are in the area of health-GIS (Geoinformation for health science), spatial simulation, spatial analysis and usability evaluation.

Bernhard Kosar is research assistant at the Carinthia University of Applied Sciences (CUAS), Department of Geoinformation and Environmental Technologies. During his internship at the Louisiana State University he gained deep knowledge in the fields of Geoinformation in combination with the development of Web-based systems, spatial databases and spatial simulations.

## References

- Ballas D, Clarke G P, Dorling D, Eyre H, Thomas B and Rossiter D (2005). SimBritain: a spatial microsimulation approach to population dynamics, *Population Space and Place*, 11:13–34.
- Clarke G and Harding A (2013). *Conclusions and the future of spatial microsimulation modelling*, in Spatial microsimulation: a reference guide for users, eds. Tanton, R and Kimberley, E, Springer.
- Harland K (2013). Microsimulation Model User Guide (Flexible Modelling Framework)', School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom.
- Harland K, Heppenstall A J, Smith D and Birkin M (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *JASSS*.15(1) 1.
- Lovelace R (2013). Supplementary information: a user manual for the integerisation of IPF weights using R, 1–18. Available online on arXiv or from the University of Leeds.
- Lovelace R (2014). Introducing spatial microsimulation with R: a practical. National Centre for Research Methods 08.
- Lovelace R and Cheshire J (2014). Introduction to visualising spatial data in R. NCRM Working Paper. EloGeo.
- Tomintz M N, Kosar B and Garcia-Barrios V M (2013). simSALUD – a Web-based spatial microsimulation application to support regional health planning in Austria, Conference paper, European Regional Science Association, Palermo, Italy.



# Exploring Digital Welfare data using GeoTools and Grids

Hodkinson, S.N., Turner, A.G.D.

School of Geography, University of Leeds

June 20, 2014

## Summary

As part of the Digital Welfare project [1] a Java GIS library called GeoTools [2] has been used to automate the production of numerous maps. The extended abstract outlines this work and provides some detail of the geographical analysis involved. The main source data are: client enquiry data obtained from the main advice giving agencies in Leeds Local Authority District (LAD); and, housing and council tax benefit claimant data for Leeds LAD. The primary data are spatially referenced by residential postcodes either at the postcode unit or postal sector level. The data have been explored and are being analysed for specific purposes that are outlined.

[1] <http://www.geog.leeds.ac.uk/people/a.turner/projects/DigitalWelfare/>

[2] <http://www.geotools.org/>

**KEYWORDS:** Java, GeoTools, Leeds, Welfare, Automation, Mapping, Postcode, Census, Advice, MAUP

## 1. Introduction

Java programs are being developed in order to reproduce (in a few simple steps) spatial generalisations and geographical maps from postcoded benefit claimant and Advice Leeds client data. The programs make use of GeoTools (an open source Java library that provides tools for geospatial data) and Grids (an open source Java library for manipulating 2D square celled raster data) and are similarly made available as open source (Turner, 2015).

Customised exploratory data analysis tools are being developed based on this work. The hope is that these will prove useful in both helping to understand the geography of benefits claimants and advice seekers and also in strategically reorganising services to support citizens in an era of seemingly increasingly constrained budgets. This work is being done in collaboration with Leeds City Council and Advice Leeds and is focussed at this stage on the Leeds Local Authority District (LAD).

Section 2 provides some more detailed context for this work. Section 3 focuses on the input data. Section 4 presents a selection of Advice Leeds client data geographical maps that illustrate and revisit the Modifiable Areal Unit Problem (MAUP). Section 5 is for discussion. Section 6 concludes and outlines some of the next steps we are planning to take.

## 2. Context

This research is operating on a pro-bono basis given support from the School of Geography at the University of Leeds. For a time it was partially funded as part of the Digital Welfare Project and RCUK Digital Economy Communities and Culture Network+ under the theme of Communities and Culture. The work is also based on research conducted over a number of years on other projects that evolved methodology and software, skills and knowledge.

There are a combination of reasons for using Java, GeoTools and Grids for this work. An important

one is familiarity and a desire for greater familiarity, but these are also capable, functional and reasonably stable and sustained technological developments. GeoTools is a widely used platform for web based geographical information systems that seeks to implement standards for interoperability. GeoTools is adaptable and can readily be extended in bespoke ways to develop custom interactive exploratory geographical data analysis tools. Additionally GeoTools and Grids are open source and whilst Java is not fully open source it is widely available. One of the main advantages of going down the free and open source software route is that the resulting mapping technology can be readily used and deployed on any IT infrastructure without the need for expensive software licenses.

The Digital Welfare Project aimed to look at the impacts of contemporary welfare policy changes on both service users and service providers in both welfare and general advice contexts. The project had a focus on critically exploring digitalisation and its effects on advice service users, advice service providers, and benefits claimants. What digitalisation refers to in this context is to do with how access to benefits and advice and the interface between service users and providers is becoming more online and computerised and less face to face and human. So, the Digital Welfare Project focussed on something different to what is presented here, but it provides very important context both to the work presented here and for additional work we are undertaking to try to better understand the distribution and demand for advice services, and provide useful information for service reorganisation.

The place of study is in and around Leeds which hosts GISRUUK 2015. Leeds is a metropolitan Local Authority District (LAD) area of around 552 km<sup>2</sup>. The Leeds LAD is one of five metropolitan LADs that comprise the county of West Yorkshire in the broader region of Yorkshire and the Humber in the North of England in the UK. Arguably, Leeds has a broader city region that extends beyond both the LAD boundary and the boundary of West Yorkshire. Our investigations have revealed that many citizens seeking advice from Advice Leeds are not resident within the Leeds LAD.

### **3. Input Data**

In the broader research we are undertaking we not only use data from Advice Leeds, but also data provided by Leeds City Council on welfare benefits claimants and residents in social housing. Although these data have names and addresses, the data we are supplied with has had this removed, but in most cases, the unit postcode is part of each record.

The maps shown in this extended abstract are based on the following: Advice Leeds client enquiry data at postcode unit level; postcode sector and unit postcode boundaries; 2011 UK Human Population Census boundaries (at Output Area, Lower Layer Super Output Area, Middle Layer Super Output Area levels) and the ONS Postcode Directory look up (ONSPD).

The Advice Leeds client enquiry data at postcode unit level includes data from Leeds Citizens Advice Bureau (CAB), Leeds Chapeltown CAB and Leeds City Council Welfare Rights Unit. Data from other Advice Leeds organisations are not included. The postcode sector and unit postcode boundaries were obtained via Edina Digimap. The census boundaries and the ONSPD were downloaded from the Office for National Statistics (ONS) website.

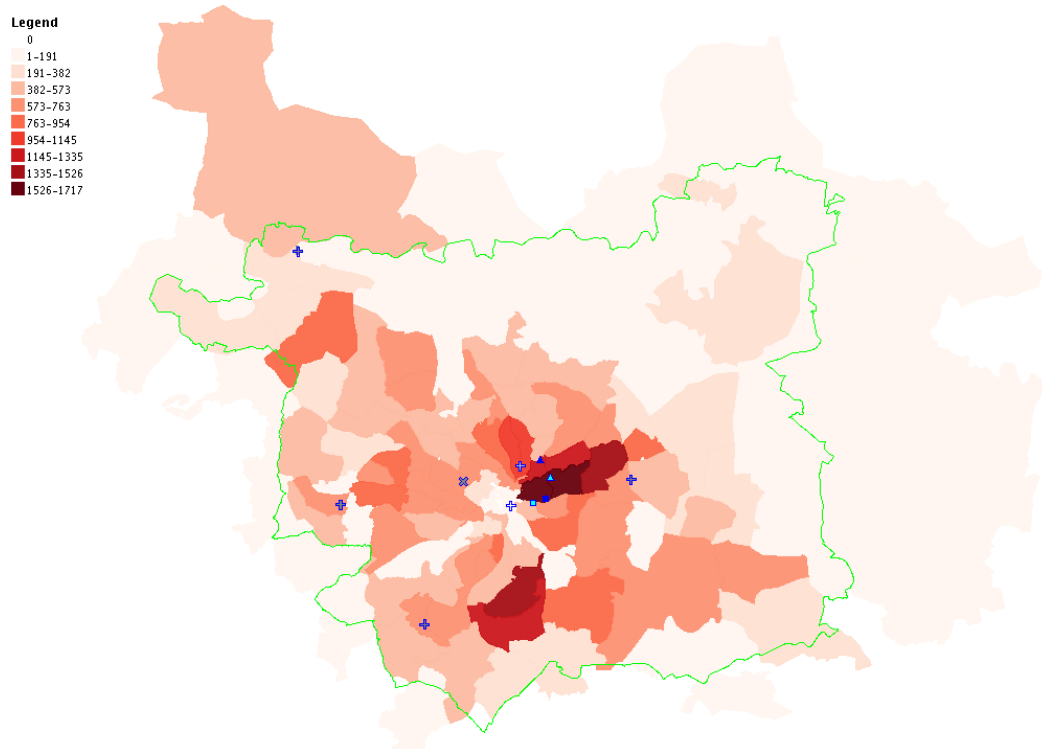
The Advice Leeds client data are sensitive even though the data have had names and full addresses of individuals removed. However, the aggregated and mapped data presented here are not considered to be sensitive and it is thought that these can be disseminated openly. The maps are more for illustrative purposes than for studying the distribution of Advice Leeds clients in detail. Readers familiar with the geography of Leeds might recognise some geographical patterns in the maps. In particular, it is likely that the Advice Leeds client distributions correlate in places with commonly used measures and indices of deprivation. It is intuitive that these things correlate and indeed we have considered developing new measures of deprivation based on data we have been provided with.

### **4. Initial Mapping**

Figures 1 to 16 are effectively maps of the same Advice Leeds postcode unit client count data for 2012. Values of zero are shown as white. The data have been classified using an equal interval

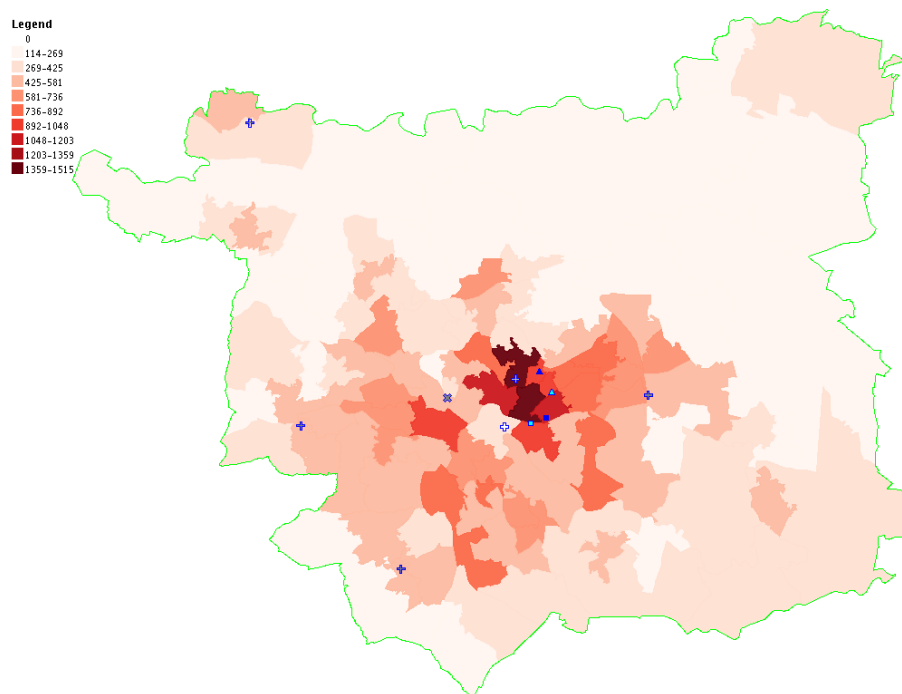
classification which divides up ranges into equal intervals based on the minimum and maximum values for each unit at each resolution. The classified data are displayed using 8 colours of red; with lower values coloured in lighter reds and the higher values coloured in darker reds. The maps shown in all the figures have the same shaped green outline which represents the Leeds LAD boundary. The data shown are for an annual time period and they represent unique clients according to client references in the data. (There may be clients represented multiple times if they are served by multiple parts of the Advice Leeds service in the time period). There are a number of reference points depicted on each of the maps. These are some of the Advice Leeds locations where citizens have been or were able to seek advice face to face.

To produce the maps, the unit postcode data were attributed to points and either these points were used in a point-in-polygon type method, or the ONSPD was used to look up which census geography a postcode was predominantly in, or the postcode was truncated into a postcode sector code. Whichever of these methods was used, the aggregated data are pretty much the same with the exception being at the boundaries where the unit areas overlap the boundary of the Leeds LAD.

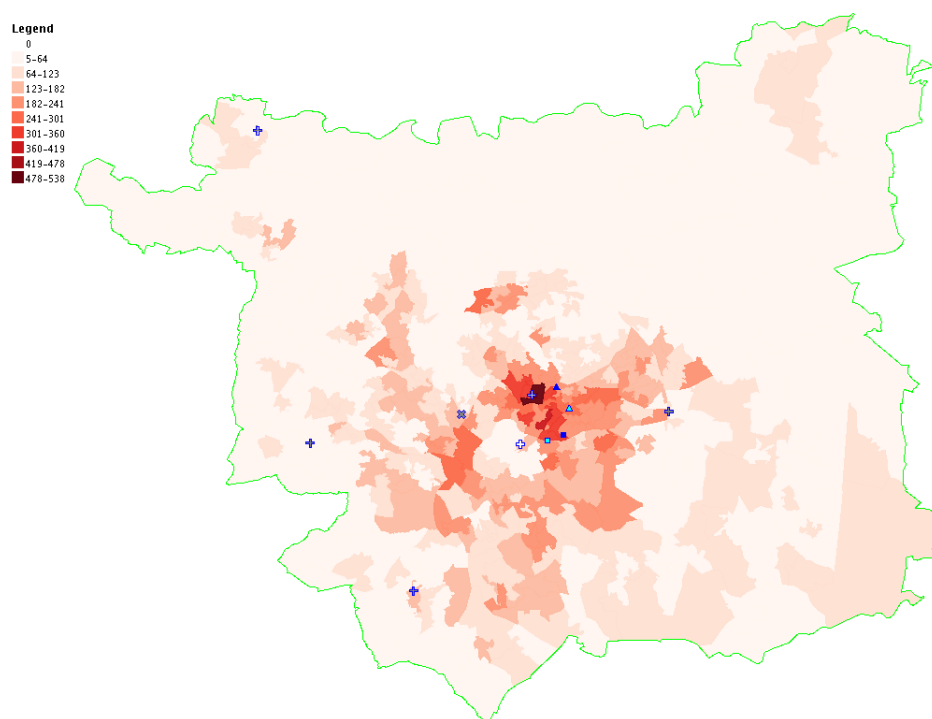


**Figure 1** Map of Advice Leeds Client Count at Postcode Sector Resolution

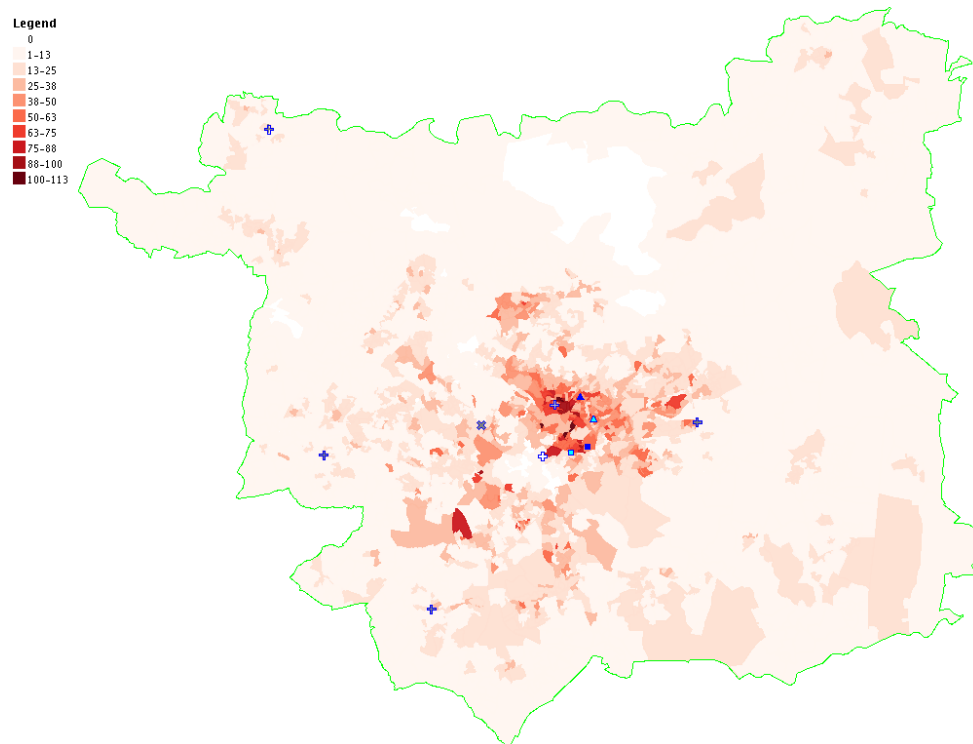
Figure 1 presents the client count data for all postcode sectors that intersect with the Leeds LAD. Figure 2 presents the client count data for Middle Layer Super Output Areas (MSOAs). Figure 3 presents the client count data for Lower Layer Super Output Areas (LSOAs). Figure 4 presents the client count data for Output Areas (OAs). Figure 5 presents the client count data for unit postcodes. The spatial resolution of the data is generally increasing in the maps shown from Figures 1 to 5. Figures 6 to 9 show square celled raster representations of the counts at 400, 200, 100 and 50 metre resolutions respectively. Although the pattern in the maps is similar, there are differences, and collectively the maps in Figures 1 to 9 illustrate various aspects of the so called Modifiable Areal Unit Problem (MAUP) detailed by Openshaw (1984). In short the MAUP cautions inference as patterns shown in maps are dependent on the resolution of the maps and the choice of boundaries, so the less regular the units are, the more concern there is with regard spatial bias.



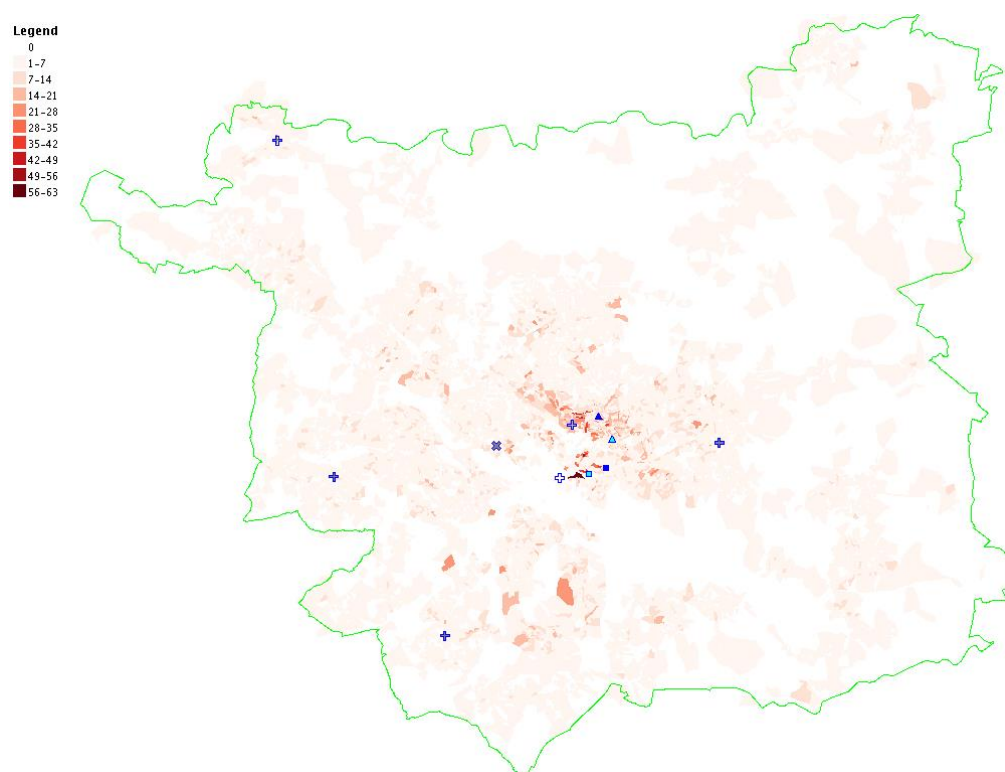
**Figure 2** Map of Advice Leeds Client Count at MSOA Resolution



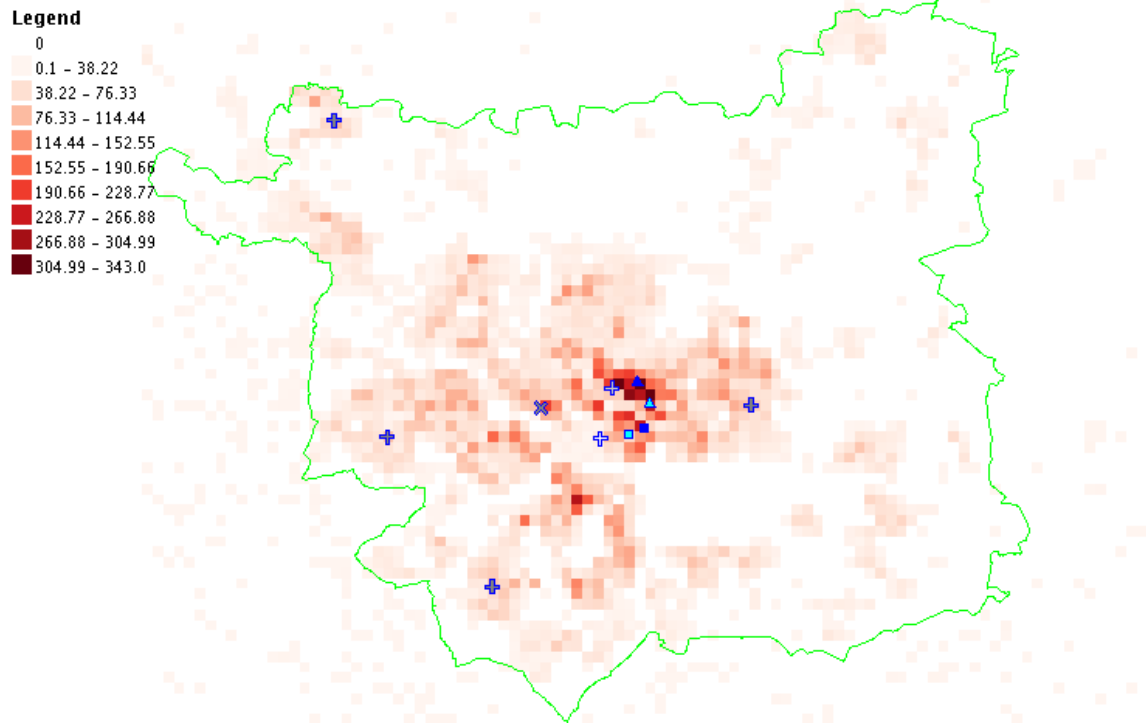
**Figure 3** Location of Lancaster University



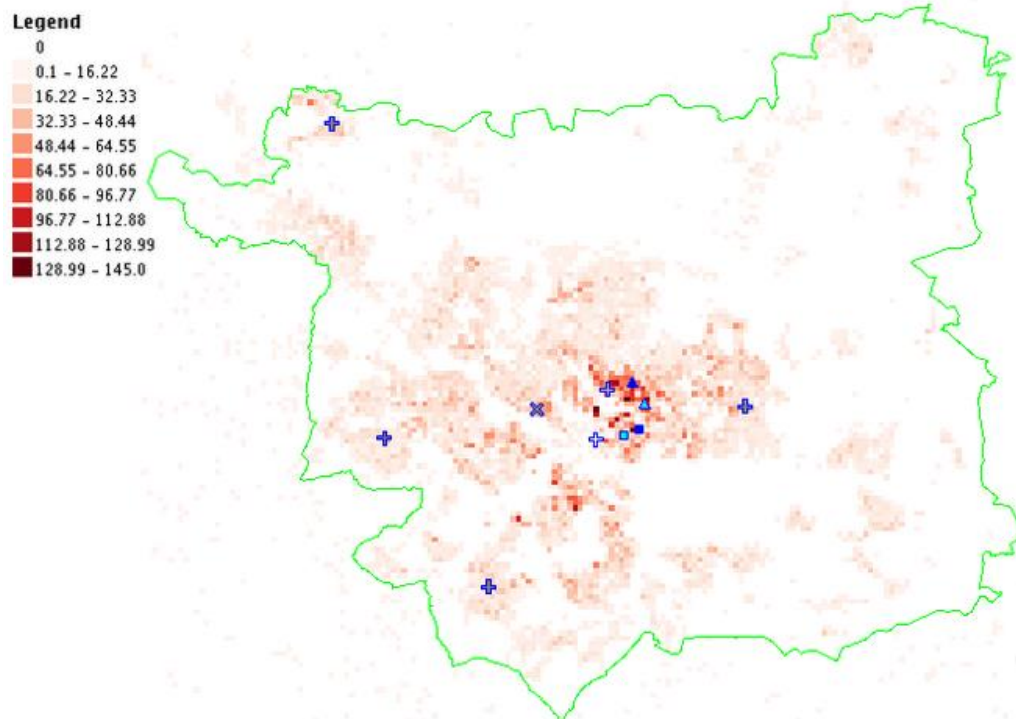
**Figure 4** Map of Advice Leeds Client Count at OA Resolution



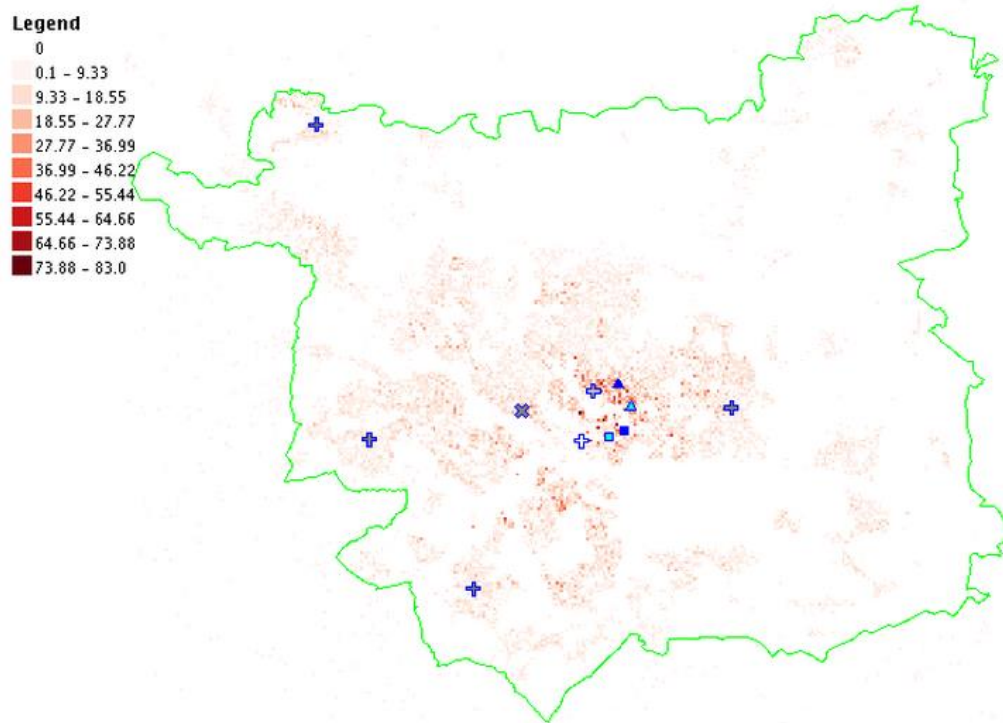
**Figure 5** Map of Advice Leeds Client Count at Postcode Unit Resolution



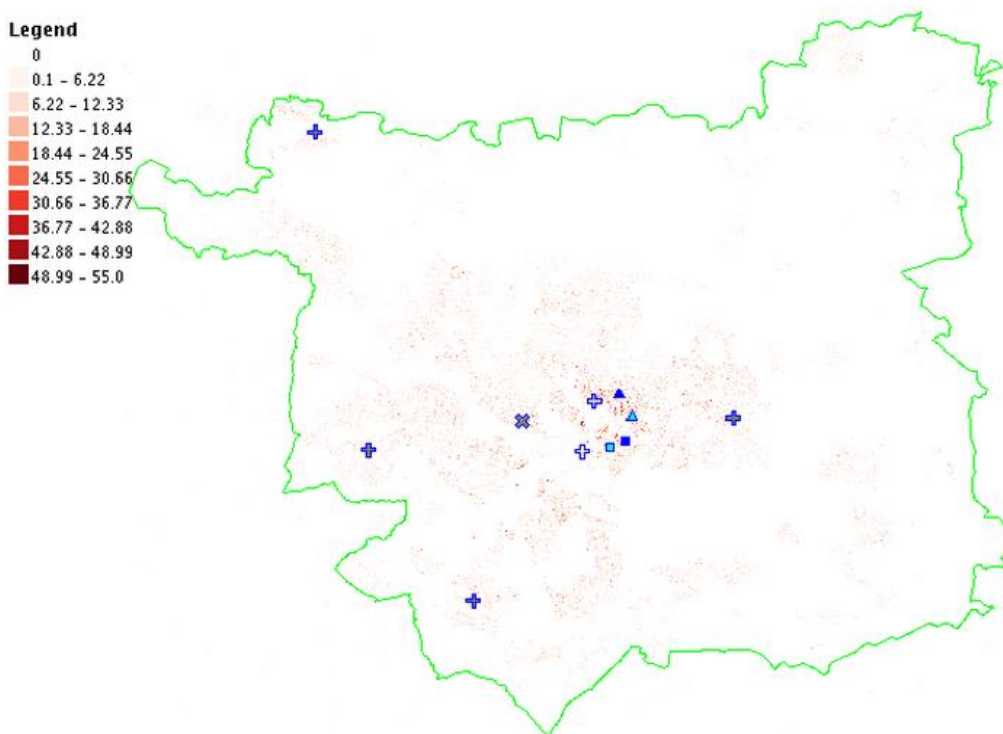
**Figure 6** Map of Advice Leeds Client Count at a 400 metre Grid Resolution



**Figure 7** Map of Advice Leeds Client Count at a 200 metre Grid Resolution



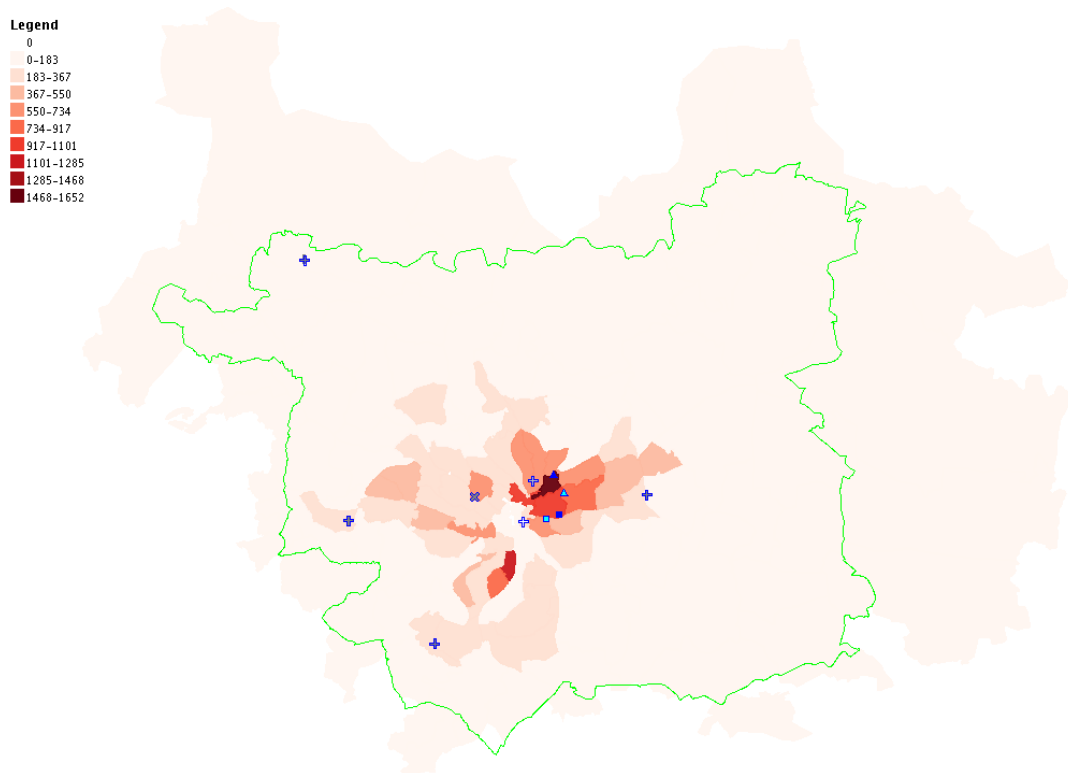
**Figure 8** Map of Advice Leeds Client Count at a 100 metre Grid Resolution



**Figure 9** Map of Advice Leeds Client Count at a 50 metre Grid Resolution

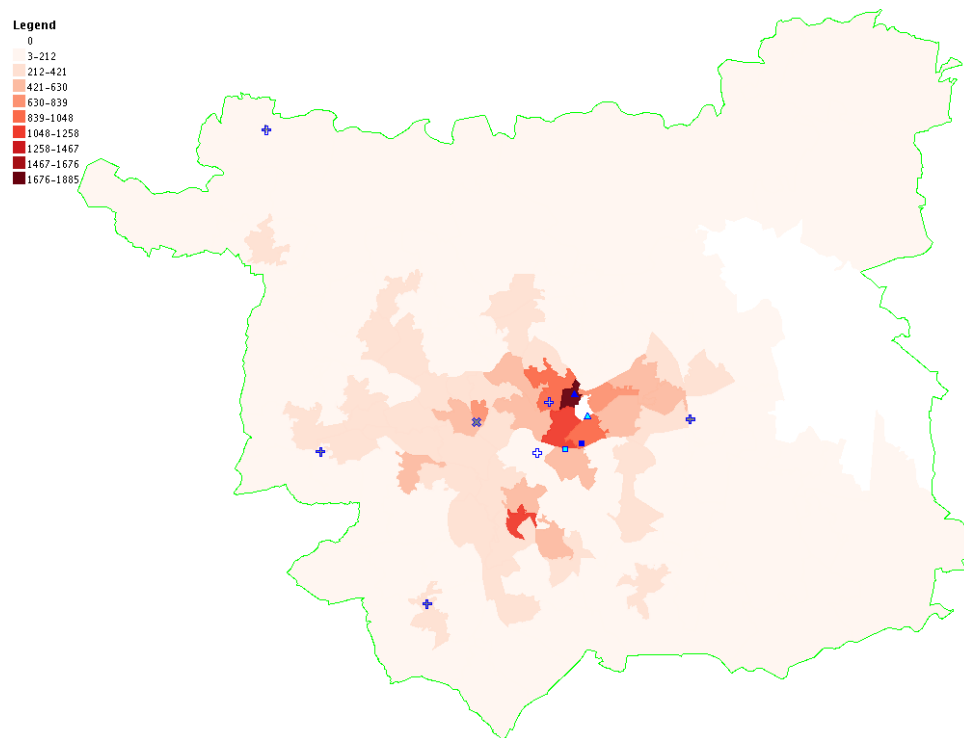
Figures 6 to 9 show raster data where all the cells are the same size. Because the cells are all the same size, the maps in these figures show the same pattern as densities. Figure 6 shows values for a grid with 400 by 400 metre square cells and has a largest client count value of 343. This value is between those of the LSOA and OA resolutions with largest values of 538 and 113 respectively (as shown in Figures 3 and 4). Figure 7 shows values for a 200 by 200 metre square cells and has a largest client count value of 145 which is more similar to the OA resolution. Figure 8 and 9 show values for a 100 by 100 metre and 50 by 50 metre grids and has a largest client count value of 83 and 55 respectively. These are above and below the unit postcode resolution maximum value show in Figure 5 to be 63.

For the postcode sectors, MSOAs, LSOAs, OAs and postcode units, the count data can generally be converted into density data by dividing by the area of each unit. Respectively, the density data are shown in Figures 10 to 14. Given a quick visual inspection, it should be clear that Figures 6 to 14 appear more similar in pattern than do Figures 1 to 9.

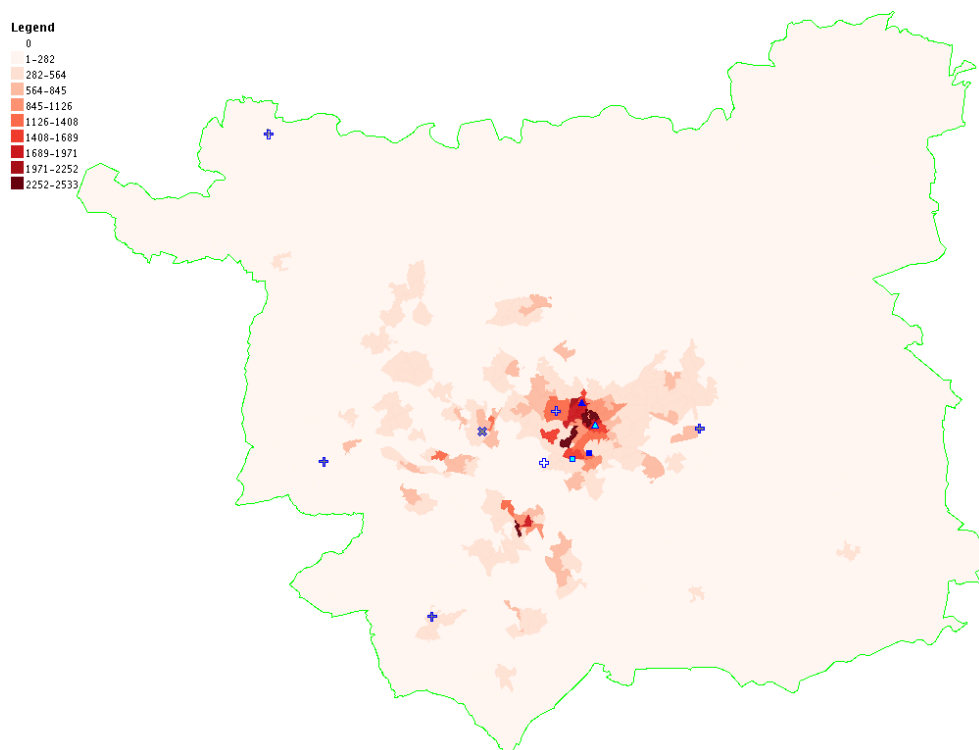


**Figure 10** Map of Advice Leeds Client Density at Postcode Sector Resolution

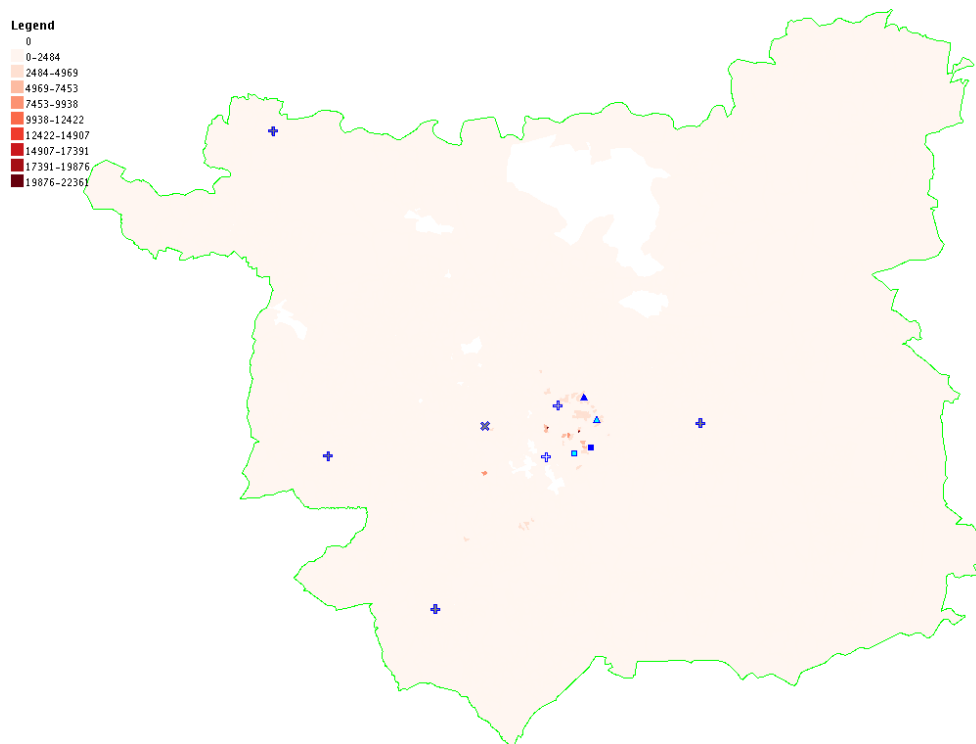




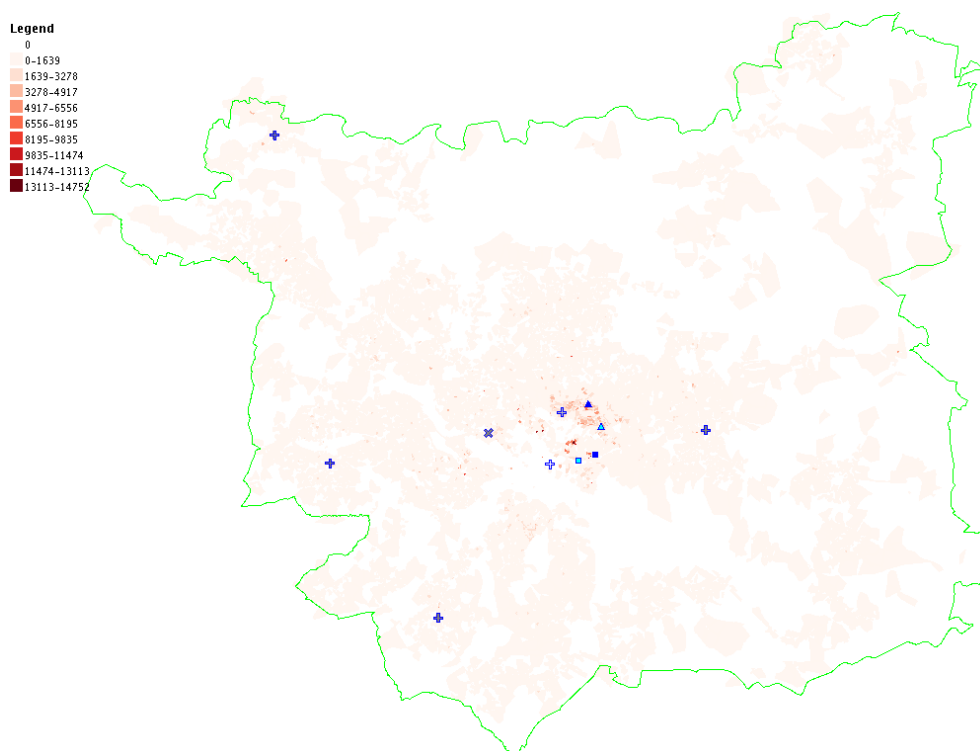
**Figure 11** Map of Advice Leeds Client Density at MSOA Resolution



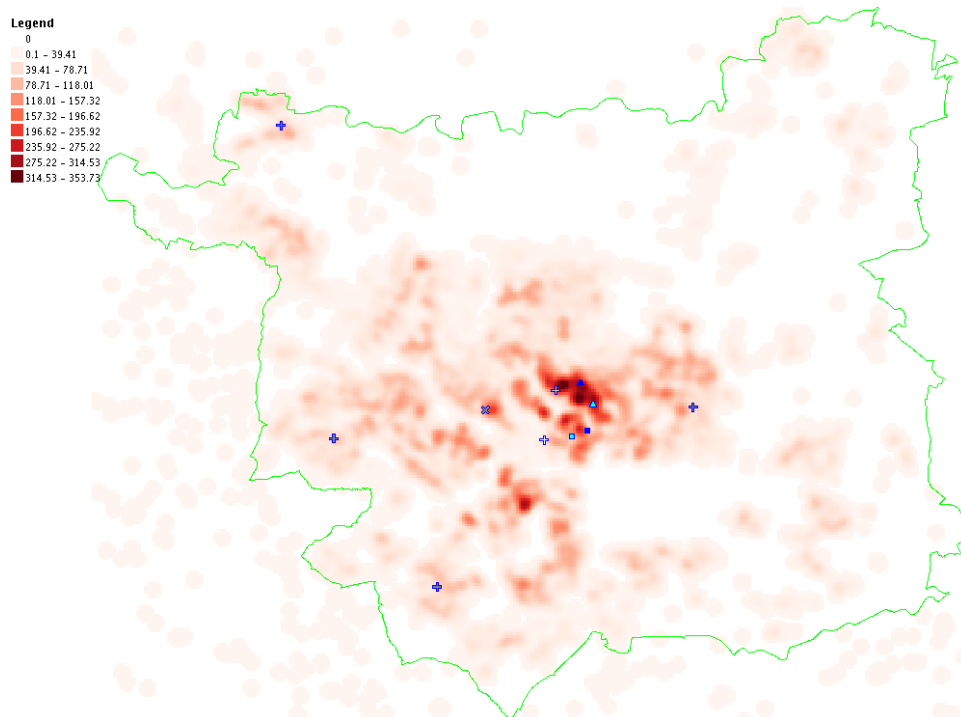
**Figure 12** Map of Advice Leeds Client Density at LSOA Resolution



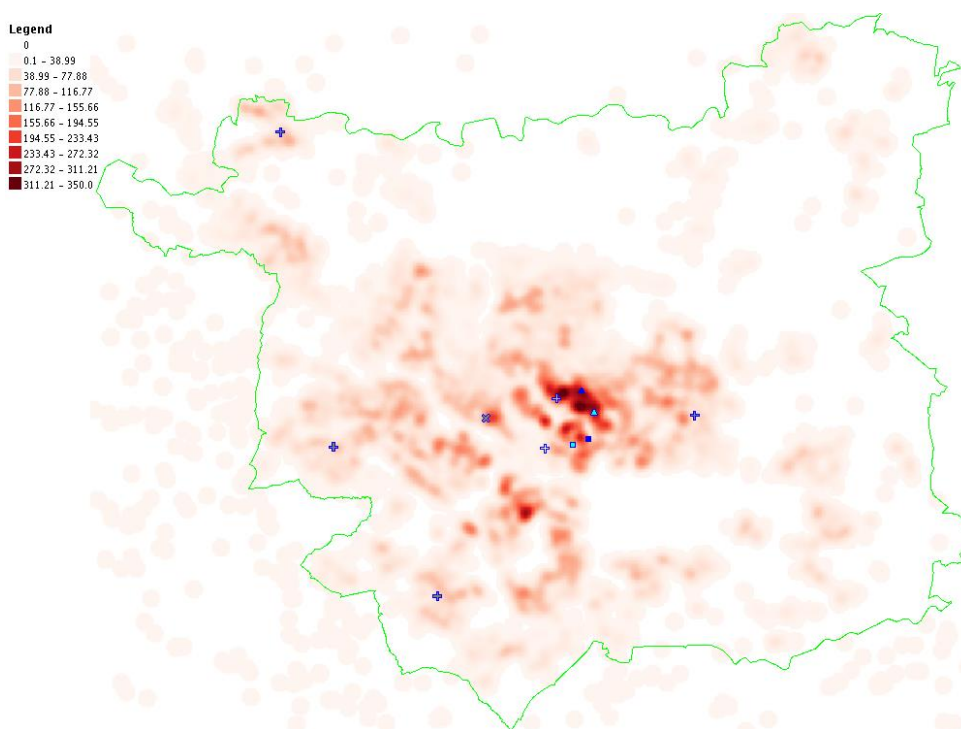
**Figure 13** Map of Advice Leeds Client Density at OA Resolution



**Figure 14** Map of Advice Leeds Client Density at Postcode Unit Resolution



**Figure 15** Map of Advice Leeds Client Density Generalised across a range of scales from 100 metres to 400 metres



**Figure 16** Map of Advice Leeds Client Densities Generalised across a range of scales from 50 metres to 400 metres

The maps reveal that the distribution of Advice Leeds clients is highly concentrated. In general they are concentrated in the more inner city areas - areas where large numbers of people reside, and with significant proportions of these being what we might term as 'deprived' and that we might expect to seek advice from Advice Leeds.

Detail is somewhat lost in the maps at high levels of spatial resolution as the areas become very small. However, the maps displayed could be generated much larger and with a zoomable slippy interactive maps, even the most detailed resolution can be explored at this scale to get a good idea of the distribution of Advice Leeds clients.

Generalising the raster data across a range of scales helps to reveal the distribution more clearly at scale. A detailed method for doing this is described by Turner (2000). Figure 15 shows the 100 metre resolution initial raster displayed in Figure 8. The values within a 400 metre radius being generalised back to each cell using a distance weighted function. Figure 16 shows the 50 metre resolution initial raster displayed in Figure 9. Similarly values within a 400 metre radius being generalised back to each cell using a distance weighted function. The pattern in Figures 15 and 16 are very similar. Both figures are produced to make the point that it is not necessary to go any smaller than a 100 metre initial resolution at the Leeds LAD scale for this kind of map in order to see the distribution clearly.

## **5. Discussion and conclusion**

For a discussion of the results and a conclusion please see the full paper which is available online via the following URL:

<http://bit.ly/GISRUK2015>

## **6. Acknowledgements**

This work has been and continues to be supported by the University of Leeds School of Geography. Some seed corn funding was provided by the ESRC funded NCRM Node Talisman which allowed collaborations with Advice Leeds to be started. Further funding was provided to continue this project under the EPSRC Connected Communities + Network as the Digital Welfare Project. This project would not be possible without the data and collaboration with Advice Leeds and Leeds City Council.

## **7. Biography**

Stuart Hodkinson is a Lecturer in Critical Urban Geography. His main research focus is on the 'new urban enclosures' with a specific interest in the politics, policies and day-to-day realities of housing privatisation, urban regeneration and state-led gentrification in the UK.

Andy Turner is a researcher specialising in computational geography and research ethics. In addition to working on research projects, he is involved in teaching and administration. Andy is a highly skilled Java programmer and has a background in mathematics and e-Research.

## **References**

- Openshaw, S. (1984) The modifiable areal unit problem. Norwich: Geo Books.  
<https://alexsingleton.files.wordpress.com/2014/09/38-maup-openshaw.pdf>
- Openshaw, S., Turner, A.G.D. (2001) 'Disaggregative Spatial Interpolation'. Paper presented at The GIS Research UK 9th Annual Conference, University of Glamorgan, Wales, UK.
- Turner, A.G.D. (2015) Digital Welfare Project Source Code Web Page:  
<http://www.geog.leeds.ac.uk/people/a.turner/src/andyt/java/projects/DigitalWelfare/>
- Turner, A.G.D. (2014) Digital Welfare Project Web Page:  
<http://www.geog.leeds.ac.uk/people/a.turner/projects/DigitalWelfare/>
- Turner, A.G.D. (2000) Density Data Generation for Spatial Data Mining Applications. Paper

presented at The 5th International Conference on GeoComputation, University of Greenwich, UK.

# Temporal profile of daily sales in retail stores in London

Syed Rakib Uddin<sup>1</sup>, Paul Longley<sup>2</sup>

Consumer Data Research Centre, UCL

17 April 2015

## Summary

This paper will explore the temporal variation of total daily sales for various retail stores in London and establish a temporal profile on weekdays and weekends. Daily sales data is obtained at Shop Keeping Unit (SKU) level from a major UK retailer which has hundreds of retail stores in the UK. The aim of the study is to identify distinguishable temporal patterns in sales behaviour between various types of stores in London and to detect and explain outliers. It would also answer broad questions such as when customer congestion is most likely to occur, how stock can be redistributed during lunch hours and evenings in various retail stores.

**KEYWORDS:** SKU, convenience store, store sales data, temporal profile

## 1. Introduction

In mid-2014, as the UK economy returned to robust growth after six years of economic crisis and austerity, the IGD (Institute of Grocery Distribution) published a 5 year forward look to the likely configuration of the grocery market in 2019. Whilst the overall size of the grocery market was forecast to grow from £175bn to £203bn (16.3%), a major part of the growth is attributed in the estimated sales growth of both online and 'discounter' grocery channels to reach market shares in 2019 of 8.3% and 10.5% respectively and the convenience store grocery sector which is predicted to account for 24.1% market share of total UK grocery sales in 2019 (having already grown to 21.4% in 2014). It is also predicted that the market share of the 'superstores and hypermarkets' would collapse from 42.2% to 34.9% (IGD, 2014).

Over the past decade, important and highly complex shifts in consumer behaviour have been taking place in the UK. Consumers have adjusted to and increasingly embraced online shopping, but additionally have sought convenience at the local/neighbourhood and workplace level. This might have been caused by technological advancement, growth of non-traditional households, the increase in women's participation in the labour force, longer working hours and population ageing (Wrigley and Lambiri, 2014).

The important shifts in consumer behaviour and prevailing cultures of consumption towards the convenience culture is also helped by the intense technological innovations, favouring massive growth of online and 'on the go' retail channels (e-commerce, m-commerce), combined with the adoption of and ownership of technology that has created challenges for even the most sophisticated consumer facing retailers (Wrigley and Lambiri, 2014).

---

<sup>1</sup> syed.uddin.14@ucl.ac.uk

<sup>2</sup> p.longley@ucl.ac.uk

This study seeks to research the customer transactions in a leading retailer's stores (convenience store, superstores, store with click and collect facility etc.) in different locations of London for 24 hour periods over few months and investigate the spatial and temporal variation of sales of various products in these stores. The retailer that forms the focus of this study operates hundreds of supermarkets and convenience stores, employs thousands of staff and deals with millions of customer transactions in each week.

The research will investigate the prospects for improving profitability by changing the composition of products sold at different times of day for certain stores and will investigate the impact of multichannel retailing through the inclusion of 'click and collect' facilities.

## 2. Data:

### 2.1. Store related data:

For this study the retailer has agreed to provide the store attribute data (size, type, product range) and store sales revenue data at an aggregated SKU level in several stores. Each SKU is contained within a logical hierarchy of products sold through the stores. Aggregation of SKUs into key groups will be used in this study, such as Sweet, Fresh Fruit & Veg, Ready Meals, Lunch etc.

### 2.2 Open retail location data:

Open retail location data was collected from the Geolytix website (2015). Following map shows various stores locations in London which include the selected stores for the study as well,



Figure 2.1: Location of various retail stores in London (Geolytix Website, 2015)

For further research, open data sources, UK census data and travel data for retail customers in London could be collected through the Oyster card data from Transport for London (TfL) and could be supplemented by London travel demand survey (LTDS) data and Journey to work Census data from Census 2011.

### 3. Results and Conclusion:

A number of convenience stores are selected mainly at Central London locations and their sales revenue is monitored to establish a temporal profile. Early results has provided a distinguishable temporal patterns in sales behaviour in various stores in London. They are shown in the following figures for total store revenues in weekdays (20days) and weekends (8 days) for about 16 selected retail stores in central London over 4 weeks from mid Jan to mid Feb, 2015.

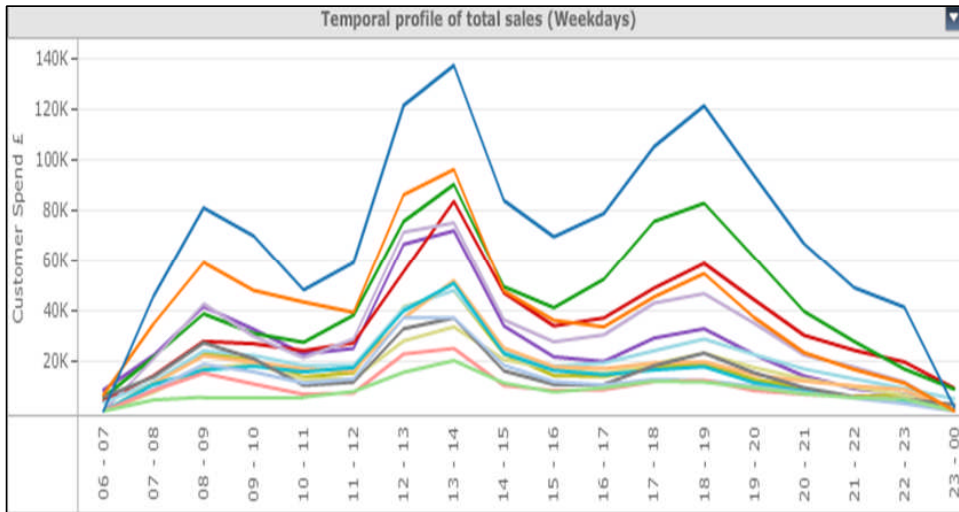


Figure 3.1: Temporal profile of total sales (Weekdays, 20 days)

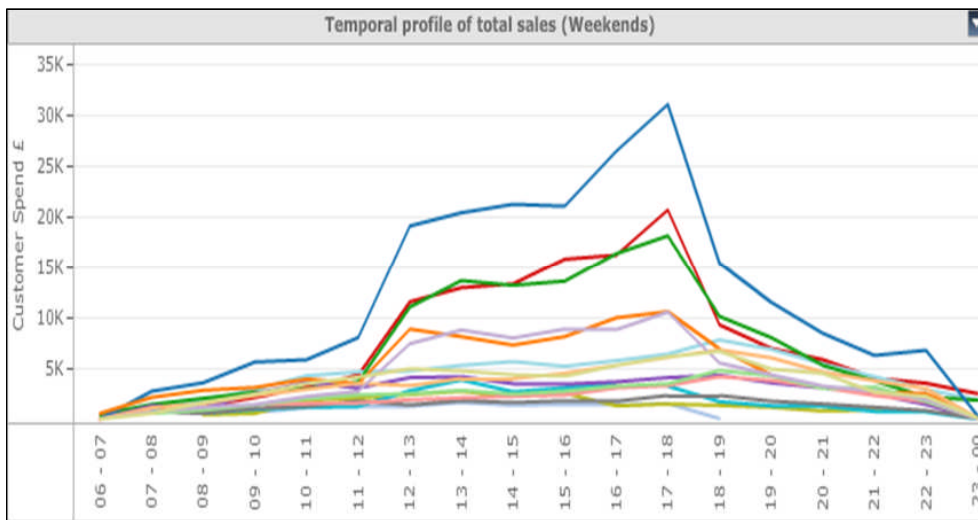


Figure 3.2: Temporal profile of total sales (Weekends, 8 days)

Further research would be carried out in the following areas,

- Look for the temporal profile for various types of retail products (ready food, frozen food etc.) focusing on different group of consumers.
- Analyse the footfall data in Central London area and find relationship of its temporal profile with that of the store revenue profile.
- Research retail consumer's shopping behaviour and link it with their travel pattern in a smart city.



#### **4. Acknowledgements**

This research is funded by an EPSRC and ESRC DTC cohort, in association with the ESRC Consumer Data Research Centre at University College London. Special thanks to Guy Lansley, UCL for helping the collaboration with the retailer.

#### **5. Biography**

Syed Rakib Uddin is a Research Engineer at Consumer Research Centre, UCL pursuing MRes. in Urban Sustainability and Resilience. Before joining UCL, he was a Principal Demand Planning Engineer with Transport for London and has research interests in transport and land use planning & engineering, demand planning, data analysis, analysis of multi-channel retailing and its relationship with urban sustainability and resilience.

Paul Longley is Professor of Geographic Information Science at UCL, where he also directs the ESRC Consumer Data Research Centre.

#### **References**

- IGD (2014), UK Grocery: Market and channel forecasts 2014-2019, IGD Retail Analysis, July 2014.
- Wigley, Neil and Lambiri, Dionysia (2014). Convenience culture and the evolving high street, Chapter 4, ESRC Report, "Evolving High Streets: Resilience and Reinvention, 2014
- Geolytix website (2015) [www.geolytix.co.uk](http://www.geolytix.co.uk) (Accessed 30.03.15)

# Understanding the spatial pattern of urban crime: a developing country's perspective

Faisal Umar<sup>\*1</sup>, James Cheshire<sup>†1</sup> and Shane Johnson<sup>‡2</sup>

<sup>1</sup>UCL Department of Geography, Pearson Building,  
Gower Street, London, WC1E 6BT

<sup>2</sup>UCL Department of Security and Crime Science,  
35 Tavistock Square, London, WC1H 9EZ

January 9, 2015

## Summary

Much of the published spatial analysis research of crime to date has been focused on Euro-American cities. This paper attempts to provide an insight on how we can better understand the spatial pattern of crime in a typical developing country's setting. Data were obtained through extensive field mapping, a block environmental inventory (BEI) and a crime victimization survey to generate a GIS-database of the study area. Grid thematic maps (GTM) for different crime types were produced for visual analysis, which suggests, as observed in many Euro-American studies, crime clusters geographically.

**KEYWORDS:** *GIS, urban crime, hotspots maps, developing country,*

## 1. Introduction

Researching the spatial pattern of urban crime is not a recent development (see Zorbaugh, 1929; Shaw and McKay, 1969; and Burgess and Bogue, 1964) and, for obvious reasons, remains the subject of significant academic attention (Weisburd et al., 2009). Crime clusters spatially (see Shaw and McKay, 1969; Sampson and Groove, 1989; Johnson, 2010) as opportunities for crime are also not evenly distributed across space. The bulk of research to date has pursued a theoretical perspective to better conceptualise and understand crime events alongside the empirical research conducted to test the validity of such theories. However, most research has concerned itself with urban crime in Euro-American cities, largely due to the availability of data in these countries. As data are now gradually becoming available, this paper presents an insight on how we can better understand patterns of crime in the context of developing countries with a particular focus on Kaduna – Nigeria. Research in this context – a typical African setting – is almost non-existent (Igbinovia, 1989; Arthur, 1994; and Mushanga, 2004). To address this we start by asking how does urban crime cluster in a typical developing country's setting? The question is a basic one because it has never been asked of Kaduna and few African cities have been subject to the intensive data collection exercise described here.

## 2. Background

Urban areas in countries like Nigeria often develop with little or no centralized planning and may have features that are far less prevalent or even non-existent in typical Euro-American cities. These characteristics, combined with extremely sparse spatially referenced crime and population datasets, make for a challenging environment in which to undertake the kinds of research taken for granted in more developed countries. The site of this study, Badarawa-Malali district, is an urban district within Kaduna (see Figure 1), the capital city of Kaduna state which also serves as the regional capital of the

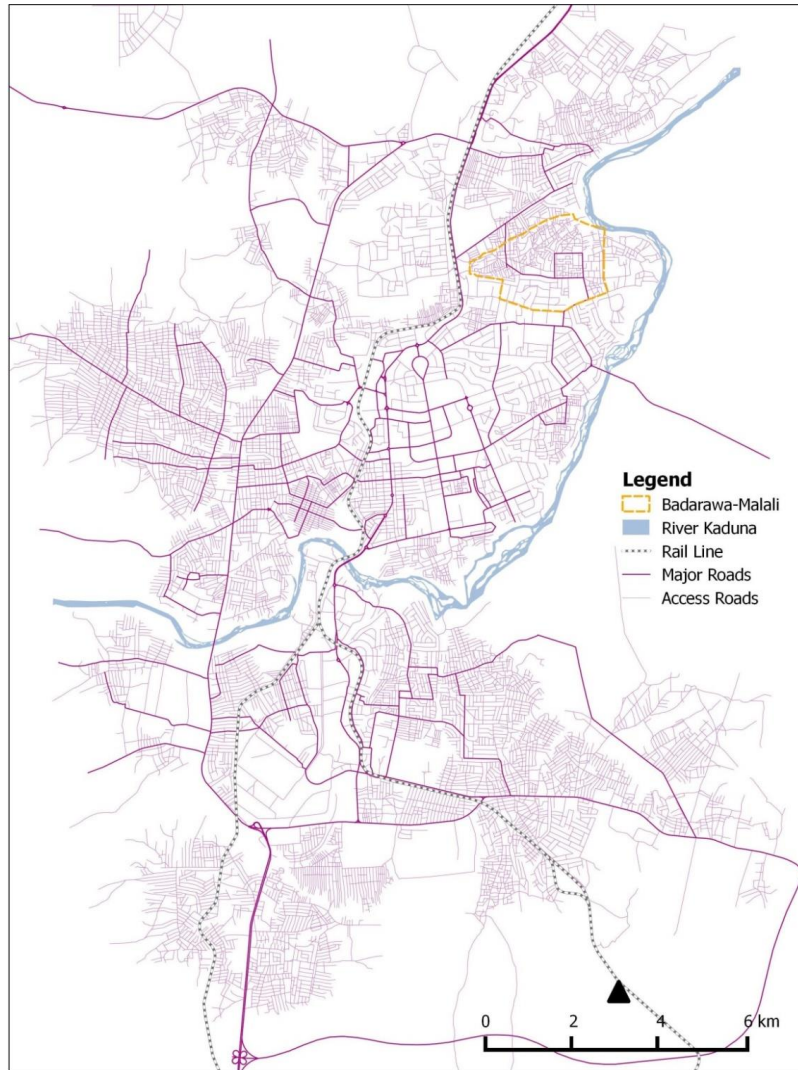
---

\* [f.umar.12@ucl.ac.uk](mailto:f.umar.12@ucl.ac.uk),

† [james.cheshire@ucl.ac.uk](mailto:james.cheshire@ucl.ac.uk)

‡ [shane.johnson@ucl.ac.uk](mailto:shane.johnson@ucl.ac.uk)

Northern Nigeria. The city is an important political, transportation and trade hub and houses a diverse population in terms of socio-economic class and racial make-up. The physical setting of the city is mixed – formal settlements with western style physical planning and informal settlements that usually emerged with little or no centralize planning.



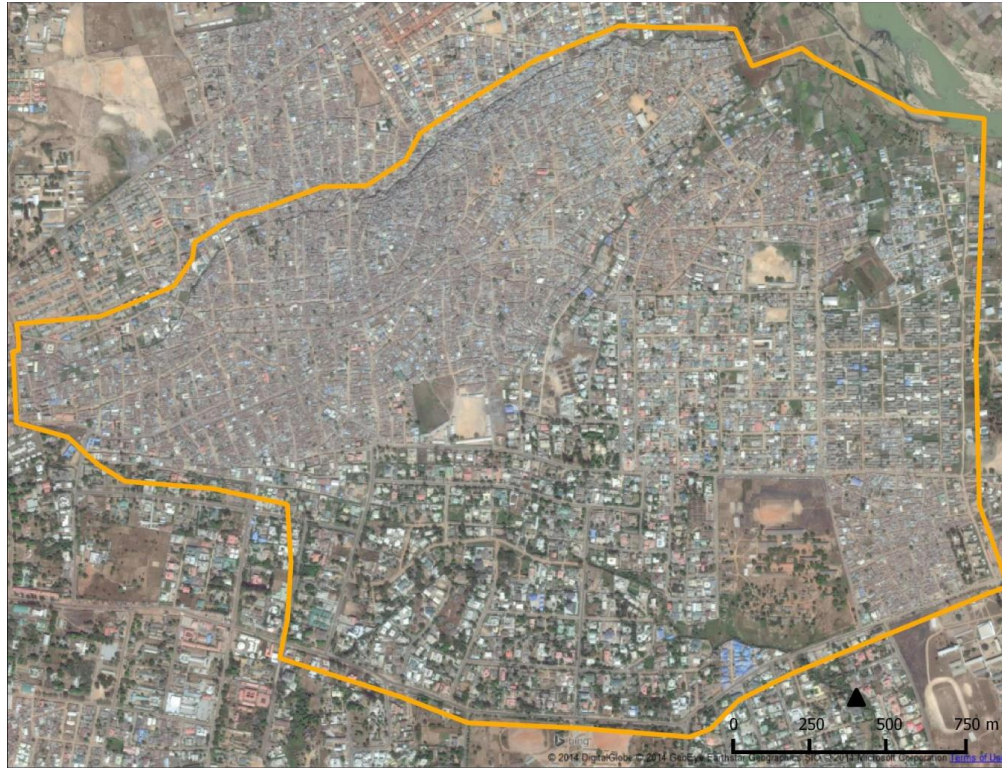
**Figure 1: Kaduna Metropolis**

Kaduna has an estimated population of 1.14million people and covers a total land area of about 250km<sup>2</sup> making the population density to be around 4,560 Person/km<sup>2</sup> (Max-Lock Consultancy Nigeria (MLCN), 2010).

## **2.1 Badarawa-Malali Urban District**

Badarawa-Malala district (see Figure 2) has a population of 96,540 people (MLCN, 2010), which represents 8.4% of the total population of Kaduna. The average household size in this district is 9.3 persons, which is almost the same as the city's average (9.8 persons). As it could be observed from Figure 2, significant parts of the district appear to be densely populated. These are informal settlements, characterised by irregular plot layouts with narrow streets that are mostly unpaved. The other parts of the district, formal settlements, are the low and medium density areas which streets are wide and mostly paved with regular sized plots aligned and well-arranged on large street blocks.

MLCN (2010) suggests that, the traditional community identity varies between the high, medium and low density areas. These variations in both social and physical settings provide for an interesting study of the spatial distribution of urban crime.



**Figure 2:** Satellite image with boundaries of Badarawa-Malali. See Figure 1 for district in context.

### 3. Data and Method

This study obtained primary data using three methods.

(a) Mapping exercise – fieldwork conducted to map the study area using paper maps produced from high-resolution satellite images. During this exercise, the boundaries of every property was marked on the paper maps and a unique reference number (URN) was assigned to each identified property (see Figure 3 for example). The data collected from the fieldwork were digitized using OpenLayers plugin in QGIS 2.0, and all URNs were entered for every property. The URN allowed for the integration of all datasets in a GIS environment at a later stage.



**Figure 3:** A sample of the paper Map (printed map is A3)



(b) Block Environmental Inventory (BEI) (see Perkins et al., 1990) used to obtain data on the condition of the built environment.

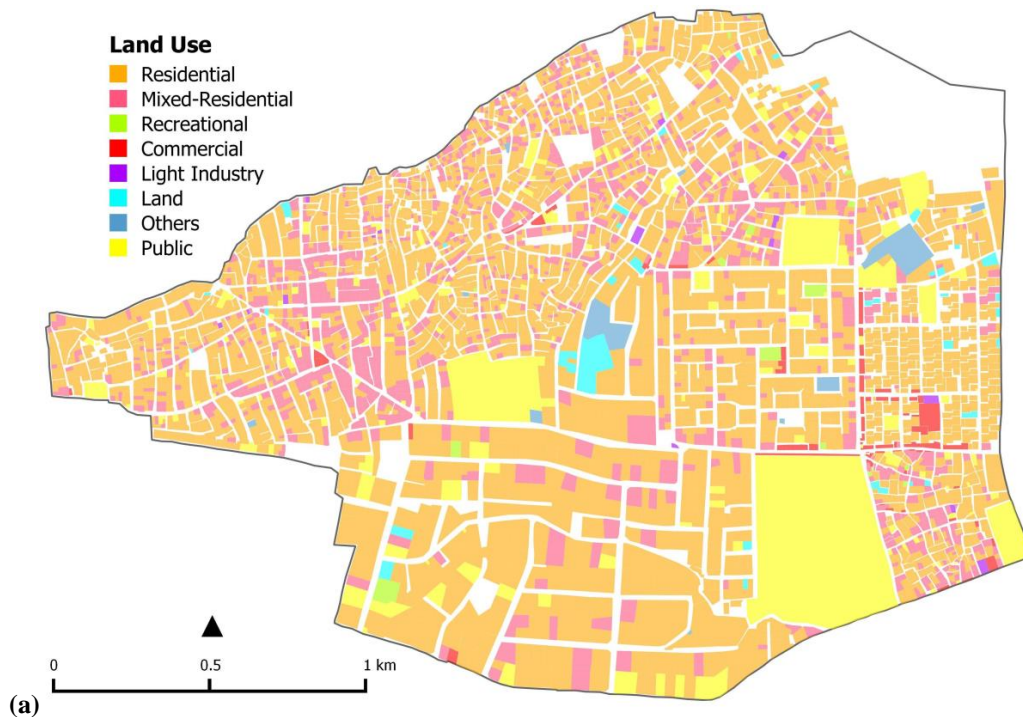
(c) Household and crime victimization survey used to obtain data on crime incidents and a range of other demographic variables. All fieldworks were conducted between March and June 2014.

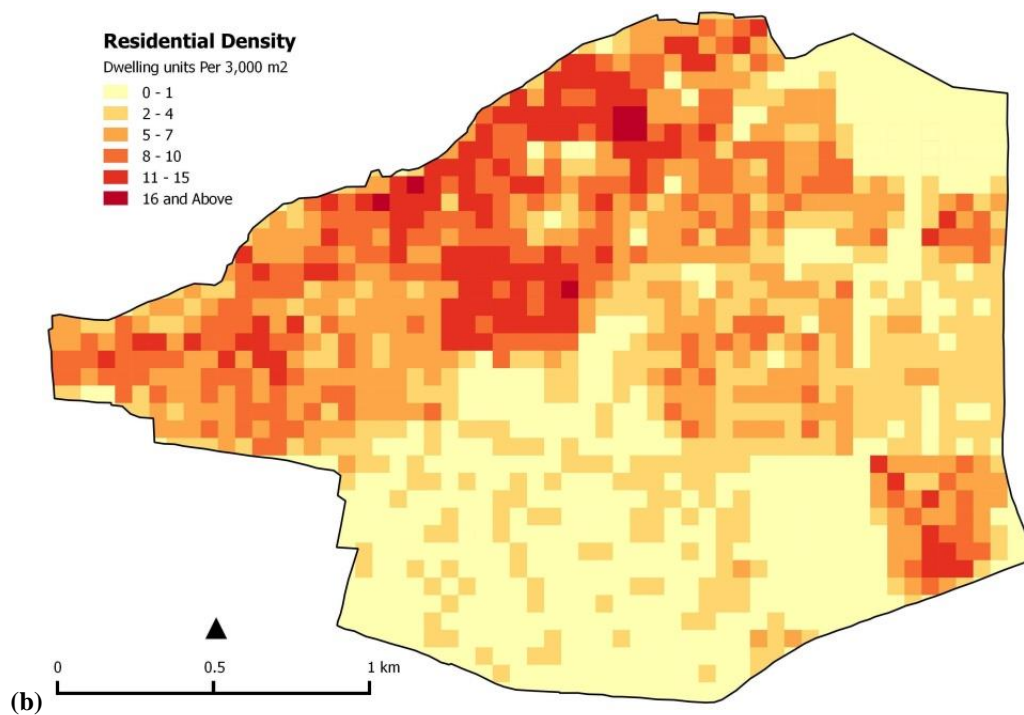
### 3.1 Block Environmental Inventory (BEI)

Trained enumerators collected data on the physical attributes related to properties using a structured BEI form prepared for this exercise (see Figure 4 for the BEI Form). First, the URN assigned to a property from the mapping exercise was entered as the RefNo and all other items on the BEI form were recorded accordingly. The records were entered into a spreadsheet and later joined to the spatial data generated from the mapping exercise.

BLOCK ENVIRONMENTAL INVENTORY FORM															Block Ref: _____								
s/n	Ref No	Land use	Plot				Other uses			Building		Access Control & Target Hindering											
			Occupied	vacant	Abandoned	Under construction	Shops	Kiosk	In-house trading	Out-side trading	Type	Material	High walls	Burglary-proof bars	Gate	Garage	Outdoor sitting	Security Lights	Guards	Open Access	Warning Signs	CCTV Camera	Dogs
1																							
2																							
3																							
4																							
5																							
6																							
7																							
8																							

Figure 4: The Header of the BEI Form





**Figure 5:** Badarawa-Malali urban district (a) Land Use (b) Density of residential units– all datasets used in producing these maps were collected from the BEI exercises.

### 3.2 Household and Crime Victimization Survey

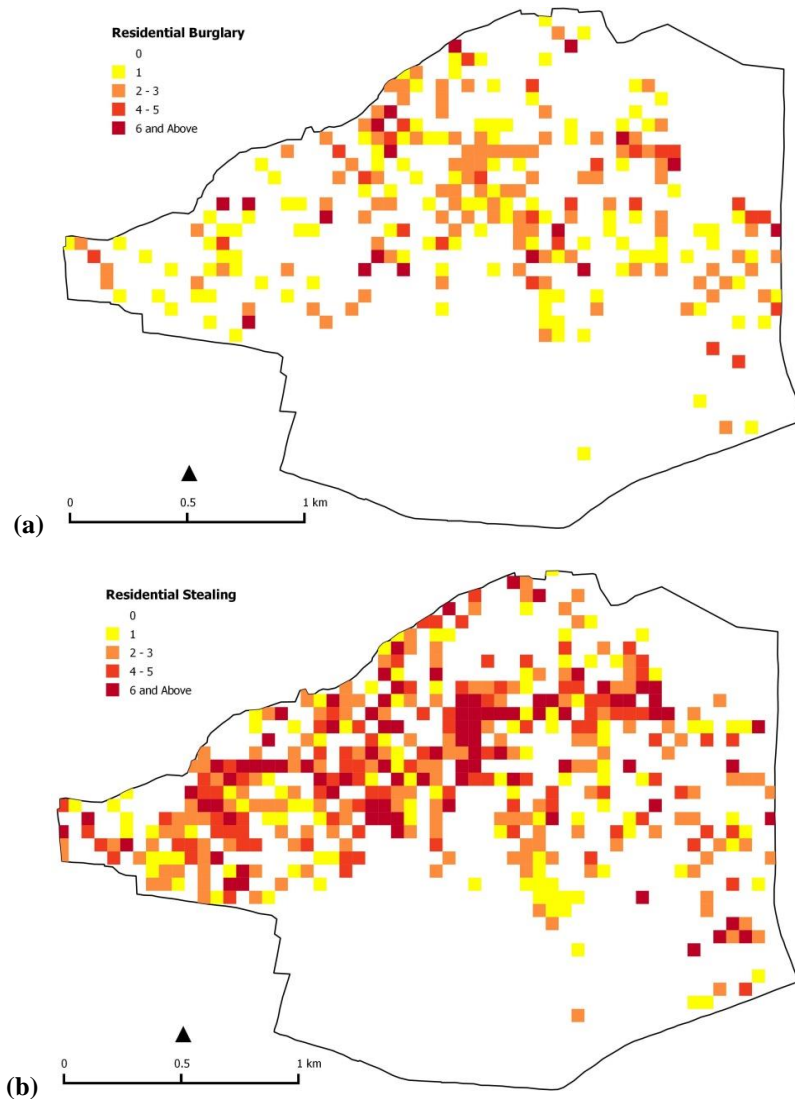
A structured questionnaire interview was designed to collect data on (a) crime incidents, and (b) demographic variables. Each questionnaire has respondent's property URN which corresponds to the one generated during the mapping exercise. This enables the geocoding of the survey data. A total of 2,027 households were interviewed, although, only 1,922 were included in this study. 105 responses were rejected for either lacking or possessing a duplicated URN, or a person at the property declined to respond to most questions. 44 questions were asked – eleven related to demography, such as, ethnicity, household structure and employment status – ten related to collective efficacy – and others related to crime victimisation. Respondents could report: residential burglary; residential stealing; damage to property; theft of automobile; theft from automobile; and damage to automobile.

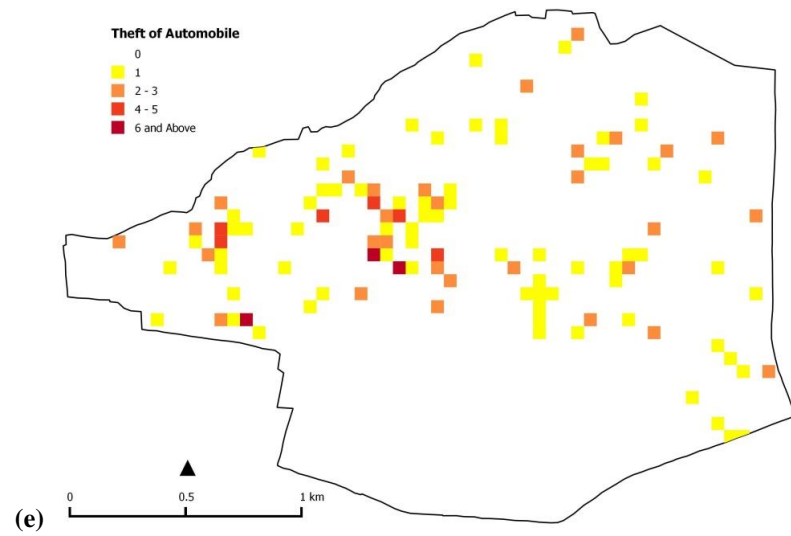
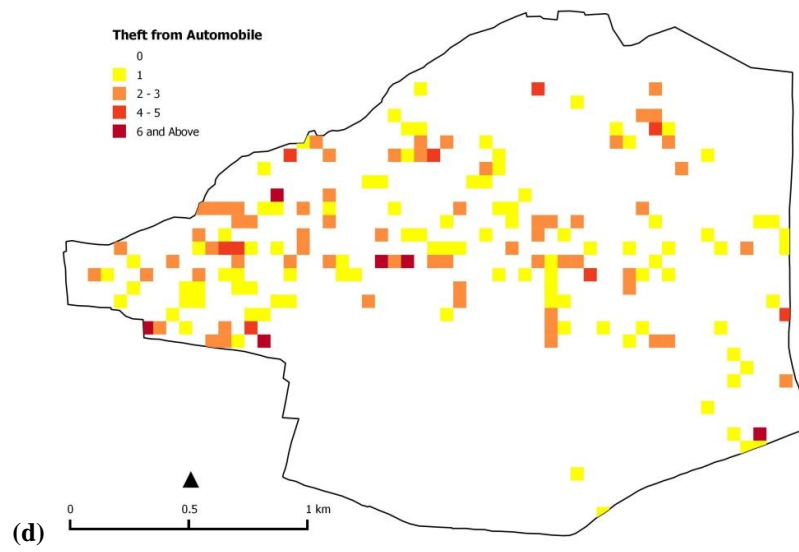
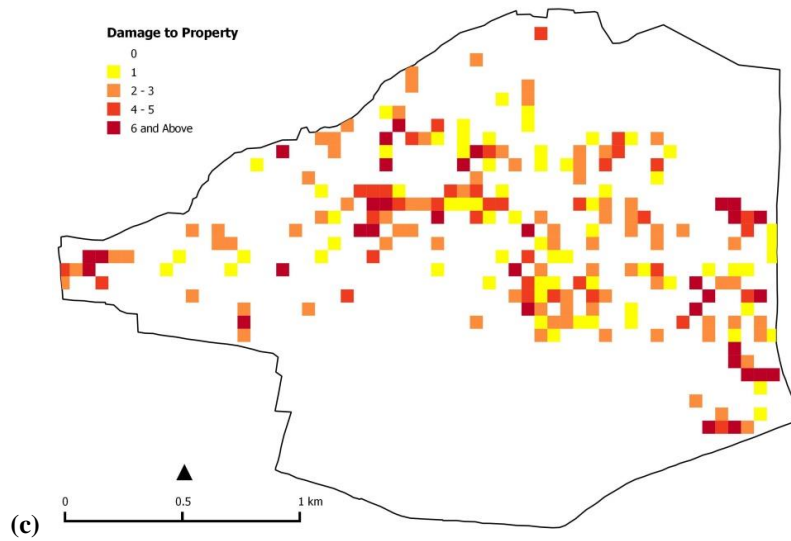
**Table 1:** Number of households reporting crime incidents

Type of Crime	No. of households
Residential Burglary	327
Residential Stealing	664
Damage to Property	278
Theft from Automobile	200
Theft of Automobile	126
Damage to Automobile	290

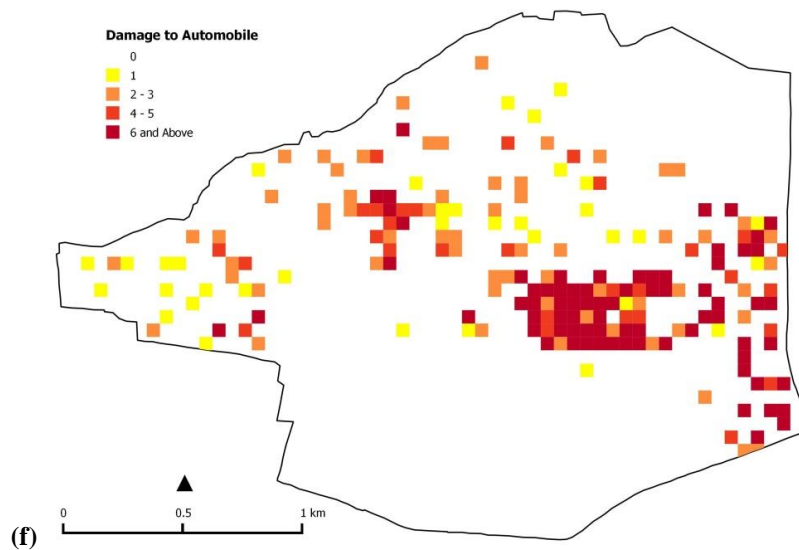
#### 4. Analysis

Although there are arguably more precise approaches (see Chainey et al, 2008), for some obvious reasons one of which is the lack of appropriate smaller geographical unit of analysis in the area under study, Grid Thematic Mapping (GTM) technique was considered at this stage in creating crime hotspot maps using a grid-cell size of 55m square. A count of reported incidents within every grid-cell was taken for each of the six crime types and hotspot maps were produced. A visual analysis of these maps (see Figure 6: a – f) suggests that, crime tends to concentrate in particular grid-cells but not in others and different types of crime show different spatial patterns. This concurred with the general existing knowledge about the spatial patterns of urban crime.









**Figure 6:** Crime hotspot maps of Badarawa-Malali District

## 5. Discussion and Future work

Knowing the precise location of where crimes occur is the starting point to identifying hotspots. This study does not only obtain the locations of crime incidents; it has generated a GIS-database that links it with the associated attributes of such places. The process was time consuming and required a combination of extensive field work, involving several trained enumerators, and some intensive GIS computations. The volume of data is unprecedented and will form the basis for what we hope will be some of the most detailed study of urban crime in a developing country. From this we will seek explanations on the micro-level relationships between urban crime and the features of both social and physical environment in a typical developing country's setting.

Although, the preliminary visual analysis suggests that some places have high crime incidents while others do not, this paper does not purport to provide the conclusion that crime clusters in space in a typical developing country's setting. More work is still needed to provide a robust spatial and statistical analysis of the datasets generated. If the premise that crime clusters geographically stands, as the preliminary analysis revealed, it will also be interesting to address other issues, such as, at what geographical scale does crime clusters? Pursuing this will mean identifying the most appropriate geographical scale of analysis which could involve further aggregation of datasets. It will also be worthwhile to understand the social and environmental correlates of crime in developing countries. It's still a work in progress and the future work will concentrate in addressing these issues.

## Acknowledgements

This research is funded by the Petroleum Technology Development Fund (PTDF) – Nigeria. The authors acknowledged the efforts of all enumerators and the department of Urban & Regional Planning, ABU Zaria and also thank Hafsar for her invaluable contribution.

## Bibliography

*Faisal is a third year PhD student at the UCL department of Geography researching urban crime in developing countries with a particular focus on Kaduna – Nigeria. Research interests are within the broader field of urban planning much of which involves GIS techniques in the analysis and visualization of urban datasets.*

*James Cheshire is a Lecture in Quantitative Human Geography at UCL Department of Geography with interest in spatial analysis and Geo-visualization. Current research focuses on the use of "big" and open datasets for the study of social science.*

*Shane Johnson is a Professor and Deputy Head of Department at the UCL Department for Security and Crime Science. Area of particular research interest, among others, is on exploring how methods from other disciplines can inform understanding of crime and security issues.*

## References

- Arthur J A (1994). Criminology and crime justice research in Africa: problems and prospects. *International Review of Modern Sociology*, 24(1), 75 – 94.
- Burgess E W and Bogue D J (1964). The delinquency research of Clifford R. Shaw and Henry D. McKay and associates. In E W Burgess and D J Bogue (Eds.). *Contributions to urban sociology*. The University of Chicago Press, Chicago.
- Chainey S P, Tompson L and Uhlig S (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 21(1-2), 4-28.
- Igbinovia P E (1989). Criminology in Africa. *International Journal of Offender Therapy and Comparative Criminology*, 33(2), v-x.
- Johnson, S D (2010). A brief history of the analysis of crime concentration. *European Journal of Applied Mathematics*, 21(4-5), 349–370.
- Max-Lock Consultancy Nigeria (2010). *Kaduna Master Plan: Draft Final Report*. Max-Lock Centre, University of Westminster.
- Mushanga T (2004). *Criminology in Africa*. Fountain Publishers, Kampala.
- Perkins D D, Florin P, Rich R C, Wandersman A, and Chavis D M (1990). Participation and the social and physical environment of residential blocks: Crime and community context. *American Journal of Community Psychology*, 18(1), 83–115.
- Sampson R J and Groves W B (1989). Community structure and crime: Testing social disorganization theory. *American Journal of Sociology*, 94 (4), 774-802.
- Shaw C R and McKay H D (1969). *Juvenile delinquency and urban areas*, (Rev. Ed.). The University of Chicago Press, Chicago. (Original work published 1942).
- Weisburd D, Bruinsman G J N and Bernasco W (2009). Units of Analysis in Geographic Criminology: Historical Development, Critical Issues, and Open Questions. D Weisburd, G J N Bruinsman and W Bernasco (Eds). *Putting Crime in its Place*. Springer Science.
- Zorbaugh H W (1929). *The gold coast and the slum: A sociological study of Chicago's near north side*. The University of Chicago Press, Chicago.

# **Reconstructing the Agricultural Landscape of the South Downs, England: an Examination of the 1940 and 1941 World War II Plough-up Campaigns**

**N. S Walford**

School of Geography, Geology and the Environment,  
Centre for Earth and Environmental Sciences Research,  
Kingston University, Penrhyn Road, Kingston upon Thames. KT1 2EE  
Tel: +44 (0)20 8417 2512  
Fax: +44 (0)20 8417 2497  
Email: [N.Walford@kingston.ac.uk](mailto:N.Walford@kingston.ac.uk)

November 7 2014

## **Summary**

World War II and the following decades are often regarded as a pivotal point in the changing fortunes of British agriculture during the 20th Century. Preparations for the wartime emergency included a National Farm Survey that would reveal land capable of more intensive production. This paper outlines the development of an historical GIS of different data sources and focuses on the ensuing 'plough-up' campaigns of 1940 and 1941 on the South Downs, England and their legacy during the post-war decades as well as factors contributing to continuity of occupation by farm families.

## **KEYWORDS:**

Agricultural restructuring, World War II, agricultural landscapes, historical GIS

## **1. Introduction and Background**

World War II is often regarded as a pivotal point in the changing fortunes of British agriculture during the 20th Century. The need for a concerted and coordinated response to maintaining a flow of food for the home population during the war time emergency precipitated an era of state intervention in the agro-food industries unknown in peace time. One manifestation of this effort was the 1941-43 National Farm Survey (NFS) of farmers with 5 or more acres (2.03 ha) of land. One of its objectives was to provide information about land regarded as 'under productive', which could be made improved by the so-called plough-up campaigns and by planting or sowing with crops that were in short supply. Access to the NFS documents became in the mid-1990s and they constitute an unequalled national historical data source of qualitative and quantitative data including maps about individual farms, although care is needed when using them for research purposes (Taylor et al., 2012).

This paper concentrates on the statistical population of over 500 farms recorded in the NFS with addresses and land in a set of 78 contiguous parishes stretching across the South Downs. By the mid-19th century the South Downs had become a well defined sheep and cereals farming region (Short 1999) that started to change early in the 20th century with the arrival of artificial fertilisers and cheap imports of sheep products. During the post-World War II (WWII) era the push towards increased self-sufficiency in temperate crops (e.g. Hodge, 1999) and national and subsequently European political intervention in the agricultural industry incentivised farmers on the South Downs (and elsewhere) to intensify their production systems. Ploughing-up grassland during WWII may be linked with modernisation of farming and continuity of occupation during the post WWII decades.

## **2. Methodology**

The focus on the South Downs in East and West Sussex reflects Short and Watkins's (1994)

assessment of the NFS as being relatively complete and detailed in these counties, although NFS surveyors were more rigorous in recording the area of land ploughed-up on West Sussex farms. Secondly, the South Downs was considered as a candidate area for national park designation shortly after WWII following enactment of the National Parks and Access to the Countryside Act (1949) and achieved this status some 60 years in April 2011. Another reason for focusing on the South Downs is that the research draws on a series of linked projects carried out over a number of years that have included this area as part of a broader exploration of the changing agricultural landscapes and systems of South East England.

The NFS comprises a collection of documents that includes the 4th June 1941 Agricultural Census Return (AC), Supplementary Agricultural Census Returns, the Primary Record (PR) and large scale Ordnance Survey Maps annotated with farm boundaries. Section F of the PR includes data about whole or part fields that the Ministry of Agriculture and Food (MAF) had directed should be ploughed-up and re-sown to one or more crops. These entries include the OS field numbers thus enabling the text to be connected with an individual field, although in cases where ploughing-up was directed on part of a field, the specific area was not identified. Spatial and attribute data were captured from the AC and PR documents and farm boundaries were digitised from georeferenced digital photographic images of the OS maps. A database of 514 farms with complete attribute, address point grid references and area polygons was created. This agricultural geoinformation was combined with three other sources of geospatial data: the fivefold agricultural land classification, the British Geological Survey maps of surface and bedrock geology and the LIDAR 2m 3D digital terrain model.

### **3. Results and Analysis**

The results from the analysis of these data sources are examined in respect of two main research questions:

- what was the extent and impact of the plough-up campaigns on the agricultural landscape of the South Downs?
- were factors such as engagement in plough-up campaign, farm size, elevation and aspect, length of occupancy and presence of motive power at the time of WWII, influential in determining whether a farm family remained in occupation on a farm during the post war decades?

Overall 36 single crop types and 38 combinations of two, three, four or even five crops types in a field were recorded in the combined plough-up data for 1940 and 1941 (there was no plough-up land on the study farms in 1942): most of these occurred with only limited frequency. The most common single crop types respectively in 1940 and 1941 were oats (35% and 34%), wheat (22% and 17%) and barley (5% in both years). In cases where the area of the crop sown was recorded (mainly in West Sussex) this was in the 4.0-19.9 ha range, which accounted for 75% the field area or in many cases the whole field. Some farmers engaged fully with the plough-up campaigns re-sowing in excess of 50 ha in 1940 and 70 ha in 1941, although on the majority of farms less than 25% of their land was entered into the plough-up campaign. Estimating the area of cereals in the pre-war, 1939, harvest by subtracting the areas re-sown to cereals indicates 53% and 47% increases in area in East and West respectively. Examination of the ploughed-up fields in relation to the agricultural land classification reveals that those in the poorer classes (4 and 5) were more likely than those of the better quality to have been direct for ploughing-up. The underlying bedrock of 75% of the parish area comprises various chalk formations, although some parishes extend northwards over Gault Clay and Lower/Upper Greensand. A higher proportion of the area of ploughed-up land was over chalk compared with overall and the percentages on the other types were correspondingly lower, notably over Gault Clay.

### **4. Conclusion**

The results of this analysis question the “locally held view that the Downs were transformed quite radically to an arable monoculture during the Second World War” (Short et al. 2000: 219). The geographical range of plough-up land stretched across the study area from Beachy Head in the east to the border with Hampshire in the west, although it amounted to only some 6% of all farmland within the 78 parishes. The ploughed-up land nearly doubled the amount of cereal cropping in the area compared with the estimated total in 1939 and 43% of the farms sowed plough-up to cereals. The

research contributed to the growing body of research employing GIS as an organising framework for managing and integrating geospatial data from a number of historical data sources and using the analytical tools and techniques available within such software for casting new light on old questions. The conversion and capture of historical data from different formats, including maps, text, photographs and numerical attributes, is a feature of this endeavour, but is not unproblematic. While the ability to envision historical landscapes and environments is undoubtedly a laudable achievement in its own right, potentially greater rewards are to be obtained by seeking to understand the operation of cyclical and linear nature of historical processes. The First and Second World Wars have been interpreted as the ‘forcing house’ of change during the 20th century, but this paper suggests that at least in so far the development of a more intensive agricultural landscape on the South Downs is concerned, the plough-up campaign of WWII had variable impact.

## 5. Acknowledgements

A Small Research Grant from the British Academy helped to initiate research tracing the occupants of farms on the South Downs during the post World War II decades. The work reported here is materially different from the original research, but nevertheless developed from the idea of linking historical data sources to examine agricultural landscape and occupancy changes.

## 6. References

- Hodge, I. (1999). Countryside planning: from urban containment to sustainable development. In Cullingworth, B. (ed.) *Planning: 50 Years of Urban and Regional Policy*. The Athlone Press, London, pp. 91-104.
- Short, B. M. (1999). Agricultural regions and land use c. 1840 in Sussex. In Leslie, K. and Short, B.M. (eds.) *An Historical Atlas of Sussex*. Phillimore and Co., Chichester.
- Short, B. M. and Watkins, C., 1994, The National Farm Survey of England and Wales 1941-43. *Area*, 26, pp. 288-93.
- Short, B. M., Watkins, C., Foot, W. and Kinsman, P. (2000). *The National Farm Survey 1941-43: State Surveillance and the Countryside in England and Wales in the Second World War*. CABI, Wallingford.
- Taylor, K. J., Walford, N. S., Short, B. M. and Armitage, R. (2012). Cautionary notes on using the National Farm Survey records in conjunction with other sources for investigating the agrarian history of Second World War Britain. *Agricultural History Review*, 60(1), pp. 77-96.

## Biography

Nigel Walford, Professor of Applied GIS at Kingston University and Director of the Centre for Earth and Environmental Science Research, focuses on applying GIS to the mapping and analysis of geodemographic and agri-environmental information. Recent journal papers are published in **I.J. of GIS, Env. & Planning A**, and **Population, Space and Place**.

# Visualisation of Spread of Chalara Ash Dieback for Raising Public Awareness and Responsible Woodland Access

Chen Wang<sup>\*1</sup>, David Miller<sup>†1</sup>, Paula Horne<sup>‡1</sup>, Yang Jiang<sup>§2</sup>  
Gillian Donaldson-Selby<sup>\*\*1</sup> and Jane Morrice<sup>††1</sup>

<sup>1</sup>The James Hutton Institute, Aberdeen, UK, AB15 8QH

<sup>2</sup>School of Computing Science and Digital Media, Robert Gordon University, UK

## Summary

A 3D model of ash (*Fraxinus excelsior*) woodland was developed to present information on the symptoms and spread of Chalara ash dieback (*Chalara fraxinea*) as part of a knowledge exchange programme for the Scottish Tree Health Advisory Group. A hypothetical woodland was designed, with characteristics of the vegetation and topography of a site in north-west Scotland. A model of different stages of infection was prepared and represented in a virtual environment. This was presented to audiences in Edinburgh and Aberdeen, and feedback on experiences and understanding of the disease provided to the team monitoring and advising on the disease outbreak.

**KEYWORDS:** Ash Dieback, Woodland, 3D Visualisation, Public Participation, Knowledge Exchange.

## 1. Introduction

*Chalara fraxinea*, a fungal pathogen causing dieback of ash trees (*Fraxinus excelsior*), was first reported in the UK in 2012, having spread across Europe from its identification in the Baltic States in the early 1990s. The disease is particularly destructive on young ash plants, with older trees surviving for many years before succumbing to secondary infections.

Denmark has suffered 60% to 90% loss of ash trees, and similar impact in Scotland would be significant for its 10.7 million ash trees, and associated loss of ecosystem services. The Forestry Commission Scotland Chalara Action Plan (Forestry Commission, 2013) identifies the need to promote and enhance proportionate biosecurity measures amongst the general public as part of a programme of knowledge exchange.

There is increasing use of 3D modelling tools, games engines and virtual reality environments for exploring scenarios of historic change at sites (e.g. Rua and Alvito, 2011; Verhagen, 2008), and scenarios of future change (e.g. Wang et al., 2012; Wang et al., 2013). This paper describes the use of interactive 3D models of an ash wood which is progressively infected by Chalara, as part of the communication programme.

## 2. Methodology

The framework used for developing a 3D model and simulation of disease spreading through a woodland comprised: design of features (e.g. tree species), compilation into a single site model, representation of stages of infection, and implementation in tools for public engagement (Figure 1).

---

\* chen.wang@hutton.ac.uk

† david.miller@hutton.ac.uk

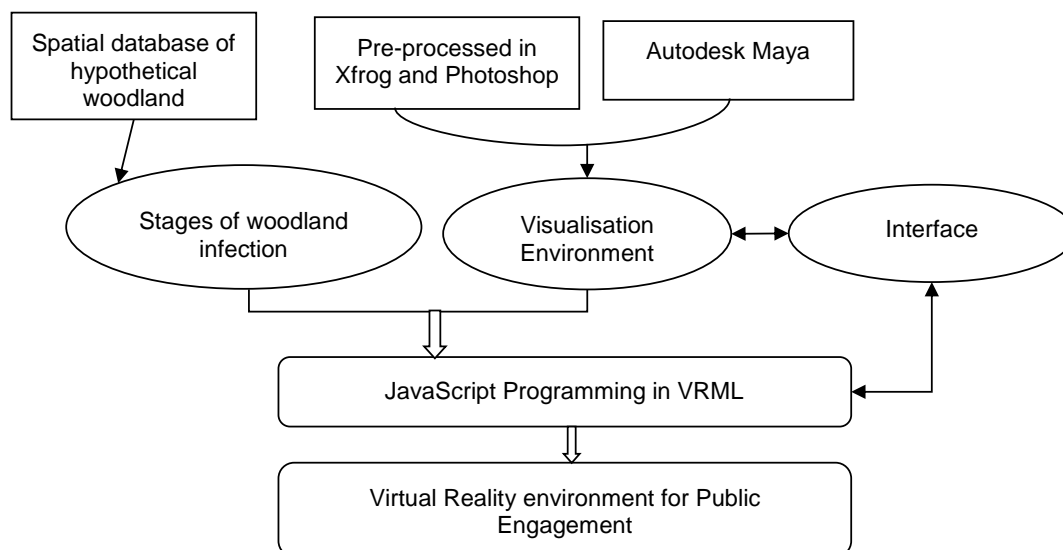
‡ paula.horne@hutton.ac.uk

§ y.jiang2@rgu.ac.uk

\*\* gillian.donaldson-selby@hutton.ac.uk

†† jane.morrice@hutton.ac.uk

Model inputs comprise spatial data and associated imagery, non-geospecific imagery (e.g. photographs of features), which are designed independently and then compiled in a single package (Autodesk Maya). The model is exported into a viewer (Octaga) in which functionality is coded in JavaScript and VRML. The environment within which the model was used in public engagement was the mobile Virtual Landscape Theatre (VLT; Ball et al., 2008; Donaldson-Selby et al., 2012), at events hosted in the John Hope Gateway Centre, Royal Botanic Garden Edinburgh.



**Figure 1.** Framework for development of 3D model and simulation of woodland affected by ash dieback.

## 2.1 Model creation

The content of a hypothetical woodland was prepared based upon the spatial patterns and rocky terrain of Rassal Ashwood, north-west Scotland, using information on the soils, topography, and vegetation. This was developed in ArcGIS to produce a spatial plan which could then be populated with features such as bare rock, ground and tree vegetation. This provided freedom to represent stages of infection and removed the potential for audience misunderstanding of infection at an actual woodland.

## 2.2 3D Model Creation

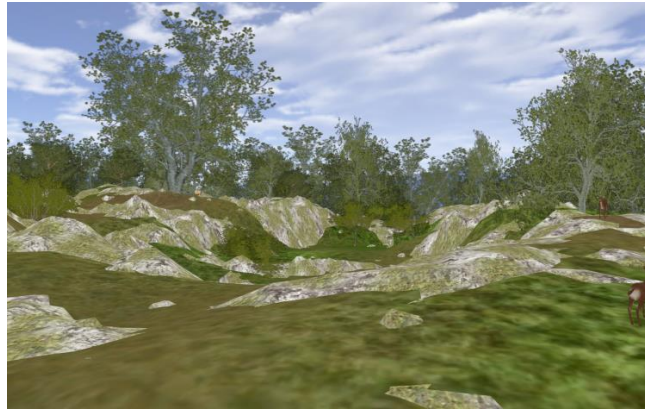
A 3D model was created as follows:

- (i) compilation of GIS database representing characteristics of the current woodland;
- (ii) creation of a terrain model, modified to create physical features to support views of combinations of trees, and enable communication of the narrative of disease spread;
- (iii) preparation of textures of landscape backdrops using high-resolution aerial imagery.

Elements added to the model were:

- (i) 3D models of features associated with woodland environments including trees species such as ash, rowan, yew, willow and birch; and a range of colours of the lichens found locally. The individual features were developed in Xfrog, modified in Photoshop, and compiled in Autodesk Maya;
- (ii) photographs of the symptoms of Chalara on tree branches and leaves;
- (iii) photographs of woodland signs used by the Forestry Commission to inform woodland visitors of the status of infection at a site ([www.forestry.gov.uk/chalara](http://www.forestry.gov.uk/chalara)) and QR codes for the relevant WWW site.

Other features added to the model included wildlife such as red squirrels, sheep and deer. Figure 2 shows an extract from the model with a view across the woodland, including the healthy ash trees, red deer and rocky terrain.



**Figure 2.** A view of a 3D model of a hypothetical ash woodland in Scotland.

### 2.3 Woodland infection sequence

Four stages of infection of woodland were created (Figure 3, a to d), planned using data of the spatial distribution of vegetation for the current state of the hypothetical site. A narrative was developed using expert knowledge of the disease with respect to age of trees and the potential for exposure, which was based on segmenting the woodland into three areas diagonally across the site. The stages of infection were then considered for each area in turn.

Attributes of spatial data features were allocated values for species, age, height, which informed interpretation of timescales for infection and stages of dieback on individual trees. Additional polygons were added to represent where other vegetation species could succeed ash trees. The 3D model of the stages of infection and change was developed with ash trees distributed across the three areas. In summary these are:

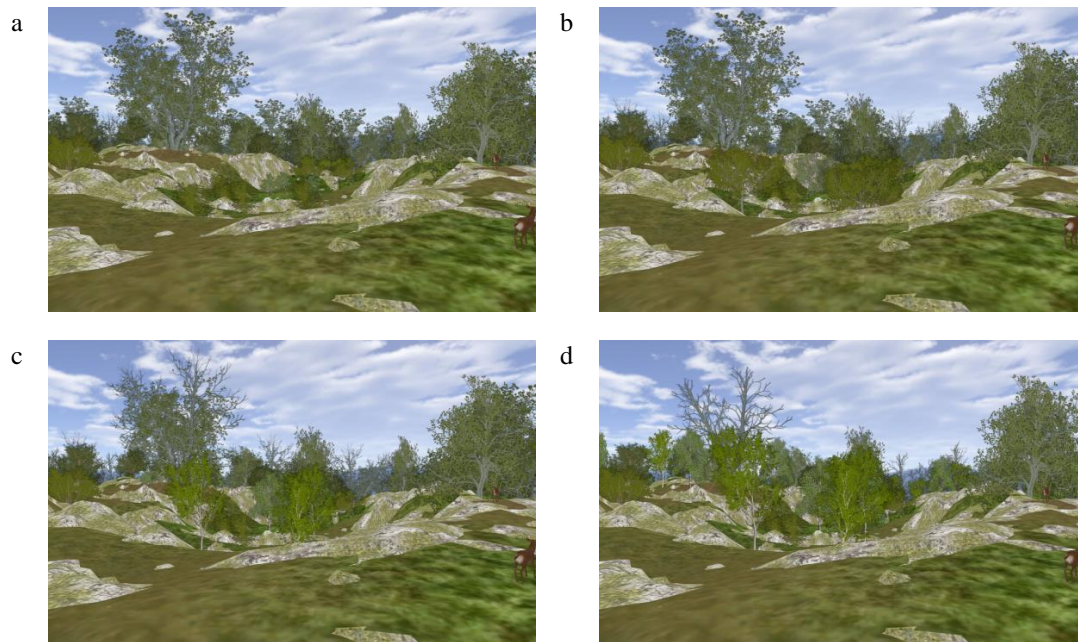
Stage 1: 1/3 ash trees in Area A experience die-back, and ash trees in Area B; Area C remains healthy.

Stage 2: ash trees dieback by 2/3 in Area A; other tree species begin to succeed (e.g. rowan) in Area A; ash trees dieback by 1/3 in Area B; ash trees in Area C remain healthy.

Stage 3: all ash trees in Area A are almost dead; succession trees grow in Area A; all ash trees dieback 2/3 in Area B; other tree species regenerate in Area B; all ash trees dieback by 1/3 in Area C.

Stage 4: worsening of level of dieback across the woodland; all ash trees dead in Area A, and succession trees continuing to grow; all ash trees almost dead in Area B, and succession trees are taller; all ash trees dieback by 2/3 in Area C, and other tree species start growing (rowan and birch).





**Figure 3** (a to d). Visualisation of four stages of woodland infection ((a) none, to (d) trees dead).

## 2.4 Model functionality

Interactive functions were added to the model which allowed the presentation of a narrative about the Chalara Ash Dieback threat to woodlands, including the stages of spread of infection, symptoms of infection, the death of trees, and succession of woodland. The functions support the delivery of the narrative of the spread of infection and dieback, and the evolution of the woodland. These include:

- (i) switching between stages of spread in the infection of a woodland by Chalara;
- (ii) switching on and off visibility of photographs of visual evidence of dieback of leaves of ash trees, and lesions on young branches;
- (iii) switching on and off of icons to represent woodland signs, using representations of the official signs for infection as provided by FC on the Chalara www site;
- (iv) QR code imagery in the 3D model to enable audience members use mobile devices to access relevant www sites;
- (v) switching between presence and absence of some wildlife;
- (vi) preset viewpoints for key parts of the narrative, in particular views to specific trees for viewing embedded images of symptoms of Chalara;
- (vii) a ‘cartoon’ of take-home message (‘Boots, Bikes and Buggies’), visible from specific viewpoints.

## 3. Medium and public engagement

The 3D model and simulation of impacts of ash dieback were used in a public engagement campaign “Moving Forward from Ash Dieback” at Royal Botanic Garden Edinburgh, and in the national tour the wider knowledge exchange programme. The event used the James Hutton Institute’s mobile Virtual Landscape Theatre (VLT) as an interactive immersive environment, for audiences up to 20, for exploring landscapes, changing their content and discussing land management options.

The model was navigable, with interactivity to appeal across the range of prospective audiences. The drop-in sessions of 20 minutes were run throughout each day, with hand-held consoles used for providing feedback on selected options on the topic of ash dieback. Younger participants were able to respond to questions on wildlife and their habitats. Participants included families from across Europe,

United States, Canada, as well as the UK, in all of which there are trees affected by Chalara.



**Figure 4.** A view of a woodland with mature ash trees and a mix of other species such as birch, hazel, yew and willow, presented to the RBGE arboriculture team and apprentice gardeners.



**Figure 5.** A woodland infected by Chalara ash dieback with some dead ash trees and evidence of regeneration of other woodland species.

A high proportion of adults, across nationalities, recognised the term ‘ash dieback’, and the nature of risks posed to ash woodlands. There was awareness of the symptom of ‘die back’ of leaves, but low awareness of the symptom of lesions on branches. The geographic origins of the disease, and its distribution across Europe were not well known, or that north and west Scotland were not yet affected. There was also low awareness of the mechanisms of spread of the disease (i.e. by spores blow between sites, or transferred in soil and tree litter via vehicles or footwear).

Of those who owned or managed woodlands with ash almost all were familiar with the risk. However, for those without direct responsibility for ash trees there was a low appreciation of the potential rate of spread of disease once present in a woodland, and the likely death of young trees within a year compared to several years for more mature trees.

#### 4. Discussion and Conclusions

The knowledge exchange programme appeared to contribute to raising public awareness of Chalara ash dieback, and risks of its spread. The use of spatial models of disease spread and visualisation of stages of infection of woodlands in a virtual reality environment have had several benefits, including:

- (i) accessibility of information on symptoms and impacts of the disease, without taking people to infected sites,
- (ii) explanation, of background to the disease and means of spread,
- (iii) communication, of impacts of disease spread through time, bringing together different stages of infection which may take several years, into a period of a few minutes,
- (iv) representation, of stages of vegetation succession in a woodland, with the loss of one species and the species which can take its place,
- (v) testing, audience understanding of key messages conveyed during the sessions.

A lesson learnt from the Foot and Mouth outbreak was that the ‘countryside remains open for business’, and not closed to access. The knowledge exchange programme forms one part of the strategy to restrict disease spread, and limit damage done. To date, the strategy for restricting spread of the disease appears to be working, with significant areas of north and west Scotland still clear.

#### 5. Acknowledgements

The authors acknowledge the collaboration in this project of staff at the Royal Botanic Garden Edinburgh and Forestry Commission, and financial support from Scottish Government’s Rural and Environment Science and Analytical Services Division (RESAS).

## 6. Biography

Chen Wang is a Landscape and Visualisation Scientist at the James Hutton Institute. He received his BEng at Soochow University, China, and a PhD at the University of Bradford. His research interests include 3D modelling of landscapes; urban environment modelling and reconstruction; character and traffic animation; 3D real time flood simulation.

David Miller is the Knowledge Exchange Coordinator at the James Hutton Institute. His background is in GIS and modelling landscapes and land use. He is Coordinator of the Scottish Government's Strategic Research Programme Theme on Land Use.

## References

- Ball, J., Capanni, N. and Watt, S. (2008) Virtual reality for mutual understanding in landscape planning, *International Journal of Social Sciences*, 27(2): 78-88.
- Bethesda Softworks (2011) *The Elder Scrolls\_ IV: Shivering Isles\_ (Oblivion)*.
- Donaldson-Selby, G., Wang, C., Miller, D.R., Horne, P., Castellazzi, M., Brown, I., Morrice, J., Ode-Sang, A., Testing public preferences for future land uses and landscapes, GIS Research UK Conference 2012, University of Lancaster, April 2012.
- Forestry Commission Scotland (2013) Chalara Action Plan 2013, [www.forestry.gov.uk/pdf/FCSCHALARACTIONPLANSOTLAND.pdf/\\$FILE/FCSCHALARACTIONPLANSOTLAND.pdf](http://www.forestry.gov.uk/pdf/FCSCHALARACTIONPLANSOTLAND.pdf/$FILE/FCSCHALARACTIONPLANSOTLAND.pdf).
- Rua, H. and Alvito, P. (2011) Living the past: 3D models, virtual reality and game engines as tools for supporting archaeology and the reconstruction of cultural heritage - the case-study of the Roman villa of Casal de Freiria, *Journal of Archaeological Science*, 38(12): 3296-3308.
- Verhagen, P. (2008) Dealing with uncertainty in archaeology. In: CAA2008 Session – On the Road to Reconstructing the Past, Programs and Abstracts, Budapest, Hungary, April 2–6, pp. 99. ISBN: 978-963-8046-95-6.
- Wang, C., Miller, D.R., Jiang, Y. and Morrice, J. (2013) Developing a Novel Approach for 3D Visualisation of Tarland. In: Proceedings of 17th IEEE International Conference Information Visualisation, London, 15th to 18th July 2013.
- Wang, C, Wan, T.R. and Palmer, I.J. (2012) Automatic reconstruction of 3D environment using real terrain data and satellite images, *Intelligent Automation and Soft Computing*, TSI, 18(1): 49-63.

# Is the use of 'mobile computer technology' appropriate for locating people with dementia?

Steve Williams<sup>\*1</sup> and J Mark Ware<sup>†1</sup>

<sup>1</sup>Faculty of Computing, Engineering and Science  
University of South Wales, UK

November 06, 2014

## Summary

This paper discusses ethical and viability issues relating to safer walking technology using a mobile phone. This technology is used to locate people with dementia when they get lost (or wander). In particular, the paper highlights problems of accuracy and availability when using GPS based techniques to locate a person, especially when that person is in a built up area or indoors. Experimental results are presented that suggest Wi-Fi based positioning offers a possible solution in such situations. The paper is presented in the context of a larger project that is considering a wider range of ethical and viability concerns.

**KEYWORDS:** safer walking technology, dementia, positioning, GPS, Wi-Fi localisation

## 1. Introduction

Dementia is a term that describes a collection of symptoms that result from damage to the brain. These symptoms can be caused by a number of conditions; the most common of these is Alzheimer's disease (NHS, 2014), which is a progressive brain disorder that damages and eventually destroys brain cells. Common symptoms of dementia include memory loss, difficulty remembering routes and becoming confused in unfamiliar places (NHS, 2014). Dementia affects over 830,000 people in the UK and the cost to the economy is £23BN per year (Alzheimer's Research UK, 2014). This is predicted to grow by 40% over the next 12 years (Alzheimer's Society SW, 2014). Associated risks of the disease include physical harm, emotional distress and premature mortality. It is reported that 40% of those with dementia get lost at some point and about 5% get lost repeatedly. Furthermore, 1% of people with dementia die while lost and half of those who are missing for more than 24 hours die or are seriously injured (McShane, 2013).

This paper details the initial stages of a research project that is concerned with the development of safer walking technologies for use by people with dementia to facilitate greater independence. The idea is to develop a mobile application (to run on a smart phone) that will allow a person who is lost to be located, while preserving as far as possible that person's right to privacy. It is acknowledged that using modern tracking technology to monitor those with dementia is not novel. Systems exist that alert carers when a patient moves outside set boundaries and allow that patient to be located at any time or place. The novelty of this project is that it aims to tackle known viability and ethical issues that current systems or technical solutions presented in the literature do not properly address. This paper deals mainly with viability and in particular the issue of reliable positioning.

---

<sup>\*</sup> steve.williams@designconnectwales.co.uk

<sup>†</sup> mark.ware@southwales.ac.uk

## **2. Ethics**

Tracking is a contentious issue that divides opinion, but current safer-walking technical solutions rarely address the ethical and human rights issues associated with tracking persons with dementia (Zwijnsen et al., 2011; Schaathun et al., 2014). In the article *Geoslavery* (Dobson and Fisher, 2003), tracking is subjected to some scrutiny. The chilling notion that technology will allow a master to control their slave is discussed. The paper, referring to Orwell's 1984, warns that "surveillance can confer control" and, with specific reference to Alzheimer's disease, questions who will decide when the patient is sufficiently impaired "to warrant such control". Mason (1986) discusses invasion of privacy and talks about how degradation of privacy may creep up on us. This theme is continued by Welsh (2003) who states that "Electronic surveillance has insidiously seeped into the fabric of society". It seems that they were right, and this may have already happened to some of us. Large parts of society now voluntarily check in and checkout of places to let everyone know where their location. For example, to date the Foursquare website has had 6 billion check-ins in which a person's location is shared (Foursquare, 2014). Welsh (2003) also discusses the connotations of criminal surveillance, again mentioning the Orwellian prediction of repression and social control.

In relation to dementia, the dilemma is this: Where is the greater breach of rights? Is it a locked door resulting in the loss of liberty or is it monitored autonomous movement using technology that can lead to loss of privacy? Landau et al. (2010) made a preliminary analysis of the thoughts of cognitively intact older people that concluded that they favoured the latter. In another study, little objection was found by the actual users of the technology, but the alternative of moving to a nursing home is used to possibly explain this (Zwijnsen et al., 2011). McKinstry and Sheikh (2013) are more cautious, stating that it is potentially useful, but that further research is required to find the most suitable people to which this technology is best suited.

## **3. Viability and positioning**

Key viability issues, such as battery consumption and GPS accuracy/availability, are also largely ignored in the literature. The project to date has considered the known limitations of GPS and sought to adopt alternative as a complement GPS positioning. Future work will address the energy consumption problem by developing procedures that minimise GPS and network activity, therefore reducing battery load.

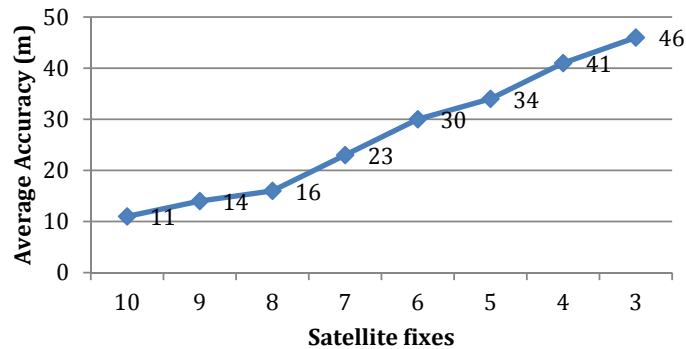
### **3.1. GPS**

GPS has "profoundly changed contemporary navigation, surveying and mapping techniques" (Karimi, Hammad, 2004). However, accuracy largely depends upon the number of satellites that are visible to the receiver. In one study it was found that the phones surveyed were accurate within 10 metres 95% of the time. (Menard et Al., 2011).

In order to test GPS accuracy and availability a working application has been developed. A standard HTC One mobile phone was used for all testing. Results were visualised and analysed via web pages developed using PHP, JavaScript and Google Maps API. When accessing GPS data on the phone accuracy readings are given, and there is a 68% probability that the true location is inside a given distance. (AndroidAccuracy, 2014). Our tests returned a higher probability, with most results falling within less than 10m of the true location (Table 1). It is known that adverse weather conditions have a detrimental effect on GPS signals (Gregorius, 1998). Our tests included taking readings on an overcast day with heavy rain. Results confirmed the relationship between accuracy and satellite availability (Figure 1).

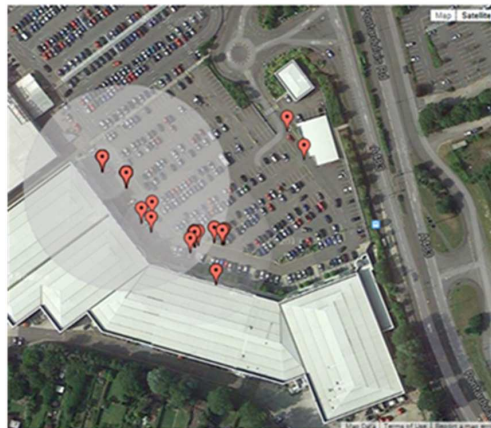
**Table 1** GPS measurements: reported and actual accuracy

Description	n	Accuracy (metres)			
		Reported			Actual
		Min	Ave	Max	
From a window in a built up area	79	17.9	31.6	34.2	≤10
Moving in a car	151	2.8	5.5	24.3	≤5
Rural, in poor weather	182	3.6	29.3	67.9	≤10



**Figure 1** Reported accuracy and the effect of the number of satellite fixes

The biggest problem found with GPS in the scenario of safer walking with dementia is its inability to work indoors. Although our tests confirmed that it does work near windows, they confirmed that indoors GPS reception generally fails. Figure 2 shows locations outside a shopping centre, no fixes were available when walking inside the building.



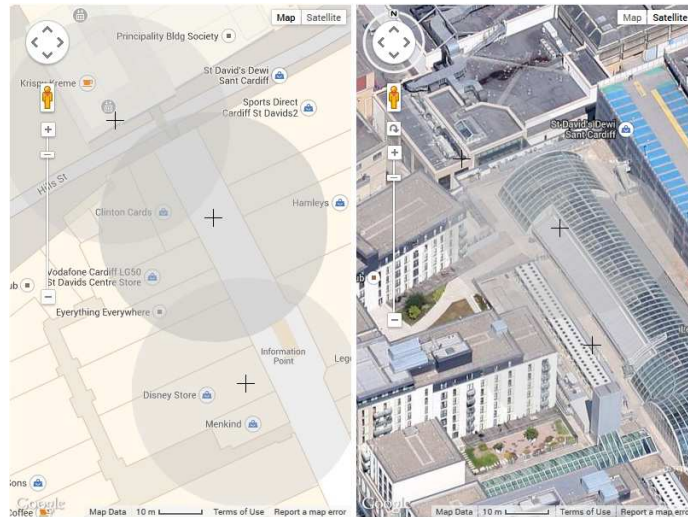
**Figure 2** GPS fails indoors

### 3.2. Wi-Fi

Wi-Fi based positioning makes use of the ever growing number of wireless access points in urban (and, to a lesser extent, rural) areas. If the position of a wireless access point is known (and increasingly this is the case via services offered by the likes of Google) then it is possible to locate a mobile device in



range of that access point. Signal strength may be used to more accurately determine location, and if multiple access points are in range triangulation may be used to pinpoint with some accuracy. We did not develop the triangulation algorithm, but our experiments confirm that this method gives indoor and outdoor location that would be of use in locating a person. Indoor tests were carried out to assess this at a major shopping centre in Cardiff; almost all of positions returned located successfully using this method, with only 5 erroneous results found out of 84 fixes.



**Figure 3** Wi-Fi positioning works indoors

For means of comparison, two points were tested in challenging GPS locations (marked X1 and X2 in Figure 4). As expected GPS location fixes were confused by the tall buildings (circled yellow), whereas a Wi-Fi fix was possible. Using this and other tests it was concluded that Wi-Fi, as a location providing mechanism provides useful results in locating persons indoors, in urban area and in built up rural locations.



**Figure 4** Wi-Fi outperforms GPS in some outdoor scenarios

#### 4. Conclusion

Initial results reported in this paper suggest that Wi-Fi based positioning, used in combination with GPS methods on a mobile phone, can be used to accurately and reliably assist in locating a person

who is lost not only outdoors, but increasingly indoors. The paper has also initiated discussion relating to ethical and rights issues that exist when considering tracking people with dementia.

The project is at an early stage and future work is expected to concentrate on the following areas:

- One of the research questions to be addressed is this: is it possible to effectively locate a person carrying a phone without constantly tracking every movement (and thereby invading privacy)? It is proposed that a mobile smart phone may be used to determine if normally experienced activity is taking place, and that it should be possible to trigger secondary actions if abnormal activity is recognised.
- Another aspect of the work will investigate ways to make the primary carer of a patient the sole recipient and custodian of stored data, thereby reducing propagation of sensitive information.
- A further issue to be investigated is that of the stigma that is attached to tracking technology, due to its association with electronic tagging. The research will consider if devices that may be worn, such as GPS and Wi-Fi enabled watches, are a suitable proposition that may reduce this problem.
- Some key viability issues, such as battery consumption, are also largely ignored in the literature. Even the most modern smart phones have limited battery capacity and as the patient could wander at any time it cannot yet be described how a charging routine may be implemented. Depending on the severity of symptoms it may be that the responsibility of keeping a charged phone with the patient would remain with a spouse or close carer. This may potentially limit usefulness to when the carer lives in very close proximity with the patient and is able to supervise. As this contrary to the ideal requirement of providing autonomous movement research is necessary to look at ways of addressing this such as in the adoption of wearable technology or by mitigating the problem, by for example by developing energy management alerts or procedures that minimise GPS and network activity, therefore reducing battery load.

## 5. Biography

Steve Williams is a student just having completed his M.Sc. in Mobile computing. His research interests are in the area of the Internet of Things, Digital Identity, GPS location, Wi-Fi location, assistive technology and in interdisciplinary collaboration.

Mark Ware is a Reader in GIS. His research interests include automated cartography, GIS and disaster management, Open Source GIS and mobile GIS.

## References

- Alzheimer's Research UK (2014). *Dementia Statistics*. Retrieved August 14<sup>th</sup> 2014 from <http://www.alzheimersresearchuk.org/dementia-statistics/>
- Alzheimer's Society SW. (2014). *Safer walking technology*. Retrieved February 6th, 2014, from Alzheimer's Society: [http://www.alzheimers.org.uk/site/scripts/documents\\_info.php?documentID=579](http://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=579)
- AndroidAccuracy (2014). Android Developers Page. Retrieved November 6<sup>th</sup> 2014 <http://developer.android.com/reference/android/location/Location.html#getAccuracy%28%29>



- Dobson and Fisher. (2003, Spring). Geoslavery. *IEEE Technology and Society Magazine*, Spring, 47-52.
- Foursquare (2014). *About Foursquare*. Retrieved November 6<sup>th</sup> 2014 from <https://foursquare.com/about>
- Gregorius, T. and Blewitt, G. (1998). The effect of weather fronts on GPS measurements. *GPS World*, May, 52-60.
- Karimi, Hammad. (2004). *Telegeoinformatics - Location-Based Computing and Services*. Florida 33431, USA: CRC Press LLC.
- Koester, R.I. (1992). Lost subject profile of Alzheimer's. *Journal of Search, Rescue, and Emergency Response*, 11(4), 20-26.
- Landau, R., Werner, S., Auslander, G.K., Shoval, N. and Heinik, J. (2010). What do cognitively intact older people think about the use of electronic tracking devices for people with dementia? A preliminary analysis. *International Psychogeriatrics*, 22(8), 1301-9.
- Mason, R.O. (1986). Four ethical issues of the information age. *MIS Quarterly*, 10(1), 5-12.
- McKinstry, B. and Sheikh, A. (2004). The use of global positioning systems in promoting safer walking for people with dementia. *Journal of Telemed and Telecare*, 19(5), 288-292.
- McShane. (2013). Should patients with dementia who wander be electronically tagged? Yes. *BMJ* 2013;346:f3603.
- Menard, T., Miller, J., Nowak, M. and Norris, D. (2011). Comparing the GPS Capabilities of the Samsung Galaxy S, Motorola Droid X, and the Apple iPhone for Vehicle Tracking Using FreeSim\_Mobi. 14<sup>th</sup> International IEEE Conference on Intelligent Transport Systems (5 pages).
- NHS (2014). *Dementia guide: symptoms of dementia*. Retrieved November 6<sup>th</sup> 2014 from <http://www.nhs.uk/conditions/dementia-guide/pages/symptoms-of-dementia.aspx>
- Rowe, M.A. and Bennett, V. (2003). A look at deaths occurring in persons with dementia lost in the community. *American Journal of Alzheimer's Disease and Other Dementias*, 18(6), 343-348.
- Schaathun, H.G., Molnes, S.I., Berg, H. and Einang R. (2014). Electronic Tracking of Users with Cognitive Impairment. *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics*, 73-90.
- Welsh et al. (2003). Big brother is watching you--the ethical implications of electronic surveillance measures in the elderly with dementia and in adults with learning difficulties. *Aging and Mental Health*, 7 (5), 372-375.
- Zwijssen, S.A., Niemeijer, A.R. and Hertogh, C.M. (2011). Ethics of using assistive technology in the care for community-dwelling elderly people: an overview of the literature. *Aging and Mental Health*, 15(4), 419-27.

# Exploratory spatiotemporal data analysis and modelling of public confidence in the police in central London

Williams D<sup>\*1</sup>, Haworth J<sup>†1</sup> and Cheng T<sup>‡1</sup>

<sup>1</sup>Department of Civil, Environmental and Geomatic Engineering, University College London

January 9, 2014

## Summary

Improving public confidence in the police is one of the most important issues for the London Metropolitan Police Service (Met). Public confidence varies over geographic space and changes over time. Spatiotemporal analysis and modelling becomes more manageable with a thorough understanding of the underlying spatiotemporal autocorrelation structure of the phenomena under scrutiny. In this study, exploratory spatiotemporal analysis is conducted on repeated cross-sectional survey data from the Metropolitan Police Public Attitude Survey. This confirmed the presence of second order nonstationarity in public perceptions of the Met police.

**KEYWORDS:** spatiotemporal autocorrelation, public confidence, police, kriging

## 1. Public Confidence and the Metropolitan Police

Improving public confidence in the police is one of the most important issues for the Metropolitan Police Service (MPS/Met). In fact, the Mayor's Office for Policing and Crime (MOPAC) and the Met have agreed to the goal of improving public confidence in the Met by 20% up to 75% by 2016 (Mayor's Office for Policing and Crime, 2013). However, MOPAC/ the Met are not on track to achieving this target (Mayor's Office for Policing and Crime, 2014a). Figure 1 is a time series graph of the percentage of Public Attitude Survey (PAS) respondents rating the police as "good" or "excellent" for the period October 2006 to September 2013. Superimposed (in red) is the trend line of the percentages required to achieve a 20% confidence increase by 2016.

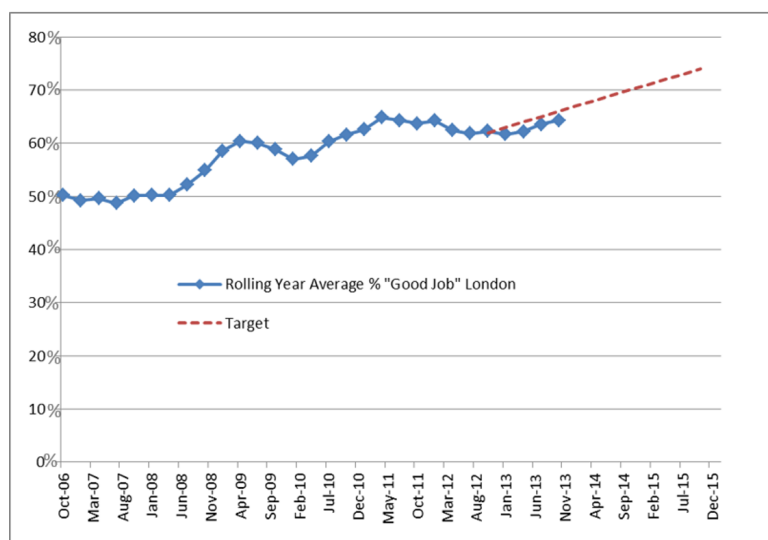
---

\* dawn.williams.10@ucl.ac.uk

† j.haworth@ucl.ac.uk

‡ tao.cheng@ucl.ac.uk

**Time series plot of confidence in the police in London**



**Figure 1** Trend line of confidence values per rolling quarter for London from October 2006 to September 2013 (blue) contrasted with the trend line required to meet the MOPAC target (red).

Current policy targets the improvement of public confidence at the aggregate level. However, to achieve this goal it is necessary to understand how confidence varies at the local level. Levels of confidence may vary throughout geographical space and change with time, and understanding these patterns is crucial to targeting improvement strategies. A local model based on the underlying spatiotemporal autocorrelation structure of public confidence may better model the heterogeneity in the phenomenon, and may also enable prediction of public confidence into the future. Examining the underlying spatiotemporal autocorrelation structure of public confidence is an important step toward this goal.

## 2. Spatiotemporal variability in public confidence

### 2.1. Description of the study area

London is the capital city of the United Kingdom. Located in the South-East region of England it consists of all the area within the M25 motorway over an area of approximately six hundred and seven (607) square miles. For administrative purposes London is divided into thirty-two boroughs. These boroughs are shown in Figure 2. Approximately 8.4 million people live in London.

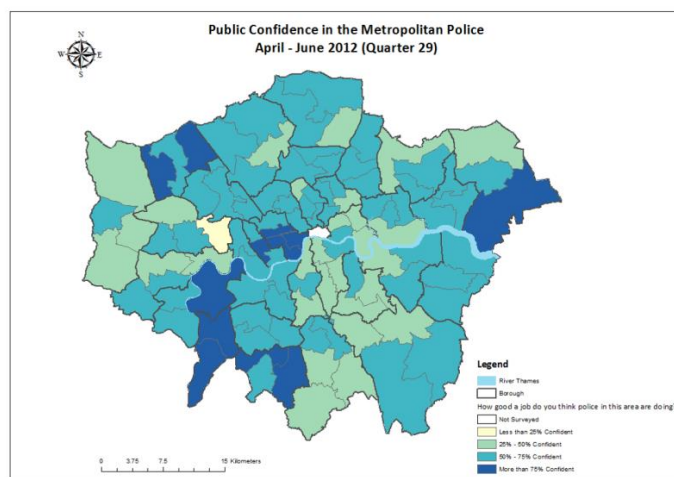


**Figure 2** Map of the thirty-two boroughs of London and the City of London.

## 2.2. Description of the dataset

The Metropolitan Police Public Attitudes Survey (PAS) has collected data on the experiences and perceptions of Londoners with respect to crime and anti-social behaviour since 1983 (Mayor's Office for Policing and Crime, 2014b). The survey is conducted on a rolling basis whereby face-to-face interviews of approximately 100 persons per quarter are conducted.

At the borough neighbourhood level (two to three wards) confidence appears to be higher in West London, particularly South West London, although this trend becomes less apparent at lower levels of spatial aggregation. Figure 3 is a choropleth map of the percentage of Public Attitude survey (PAS) respondents rating the police as “good” or “excellent” for the period for the period April to June 2012. Darker color indicates greater confidence. This snapshot of confidence for the period April to June 2012 is an example of this tendency for more confidence in the west.



**Figure 3** Choropleth map of confidence in the police April – June 2012.

### 2.3. Approach

Exploratory spatiotemporal analysis was conducted to understand the variations in confidence over the study period. The temporal, spatial and spatiotemporal autocorrelation structure of public confidence was examined using tools such as the autocorrelation functions (ACF), spatial, and spatiotemporal variograms. A progressive deepening approach was taken, such that the data was examined at different spatial and temporal resolutions. Table 1 provides some detail on the levels of aggregation used for each tool used.

**Table 1** Spatial and temporal resolutions used for analysis

Tool	Temporal	Spatial
Time series	rolling quarter	London wide
ACF	rolling quarter	London wide
Choropleth Map	quarterly snapshots	Met borough neighbourhood
Variogram	quarterly snapshots	Met borough neighbourhood (centroids)
Spatiotemporal variogram	quarter	Met borough neighbourhood (centroids)

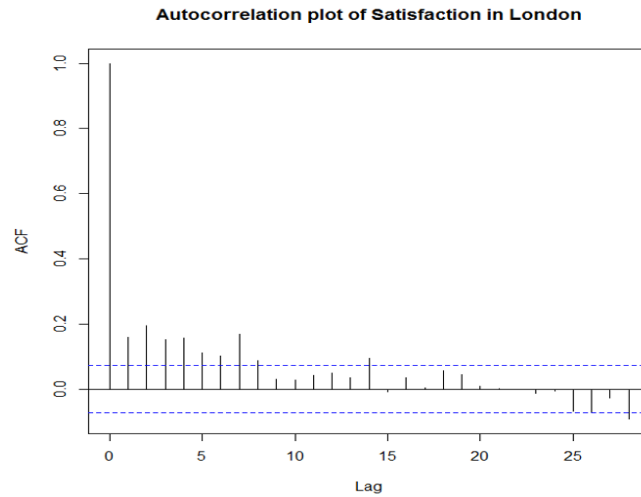
The temporal autocorrelation analysis was carried out using the ACF which measures the cross-correlation between observations of a series separated by temporal lags. The spatial autocorrelation analysis was measured using Local Moran's I and spatial variograms. The local Moran's I statistic is a local indicator of spatial association which allows significant clustering around individual locations to be discovered (Anselin, 1995). Variograms are powerful tools used to describe the autocorrelation structure of a dataset (Haining et al., 2010). Spatiotemporal autocorrelation analysis was conducted using the spatiotemporal variogram. This technique is less widely used and therefore warrants a more detailed explanation. The semivariance between all pairs of measurements is computed, the pairs grouped into bins according to the distance of separation and the average within each distance band is computed. For spatiotemporal variograms the semivariance is computed, as shown in Equation 1, for pairs of points at all spatiotemporal distances, where  $h$  is spatial lag,  $v$  is temporal lag,  $s$  is a point in space and  $t$  is point in time. As with the spatial variograms, these semivariances are grouped by spatiotemporal distance into bins and the average value per bin computed.

$$\gamma(h;v)=\frac{1}{2}E(Z(s,t)-Z(s+h,t+v))^2 \quad (1)$$

### 2.4. Findings

#### 2.4.1. Temporal

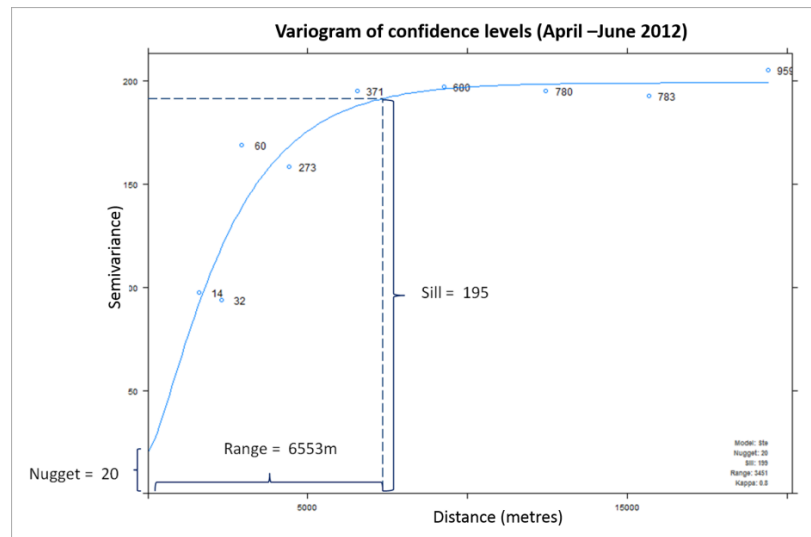
Public confidence in the police exhibits weak temporal autocorrelation. The autocorrelation function plot below, figure 4, shows the amount of "relatedness" between confidence levels in London as a whole from 2006 to 2013. Figure 4 suggests that temporal autocorrelation exists for eight quarters with temporal autocorrelation strongest on the second lag with a value of +0.2.



**Figure 4** Temporal autocorrelation plot for London for the study period 2006 – 2013.

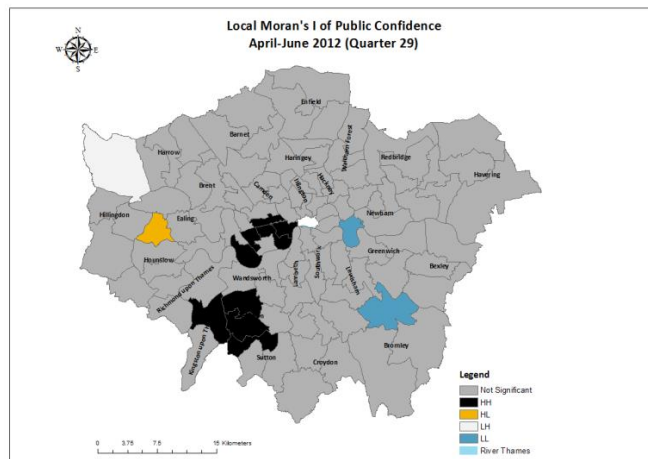
#### 2.4.2. Spatial

Spatial autocorrelation levels varied considerably from quarter to quarter, with some quarters exhibiting no spatial autocorrelation at all. Figure 5, a variogram, describes the semivariance present for the period April to June 2012. The variogram is well structured with a clear nugget, sill and range. As expected, the variance between values increases as the distance of separation increases. Figure 5, suggests that the spatial autocorrelation tails off after approximately 7km.



**Figure 5** Variogram of confidence in the police April – June 2012

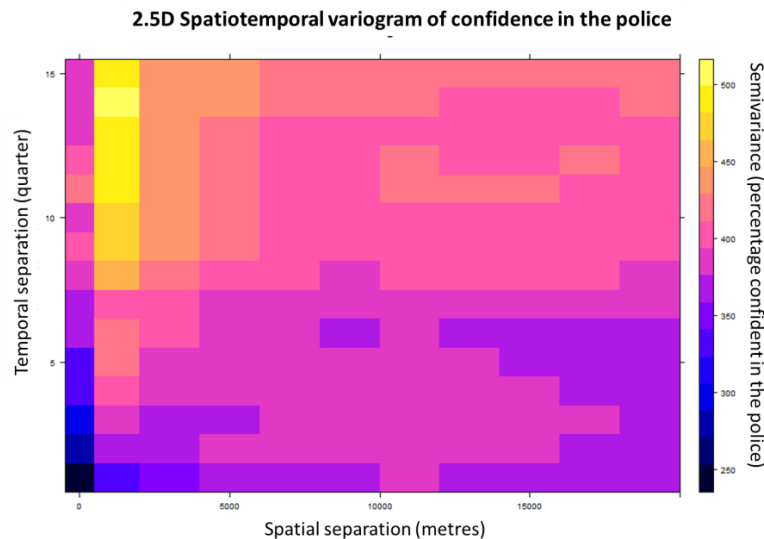
A local Moran's I test, Figure 6, confirmed the present of high-high clustering to the south-west in parts of Kingston, Merton and Sutton, as well as in some parts of west central London.



**Figure 6** Local Moran's I map of confidence in the police April – June 2012

### 2.4.3. Spatiotemporal

As expected, spatiotemporal semivariance is smallest at lag zero and with variation increasing as the number of spatiotemporal lags increase. The strongest spatiotemporal autocorrelation occurs with the first and second spatiotemporal lags. In Figure 7, a 2.5 D representation of the spatiotemporal variogram, there is a clearer decomposition of autocorrelation in time (along the y axis) as opposed to space (along the x axis), especially apart from lag zero. This reflects the earlier finding of very weak temporal autocorrelation at the London-wide level.

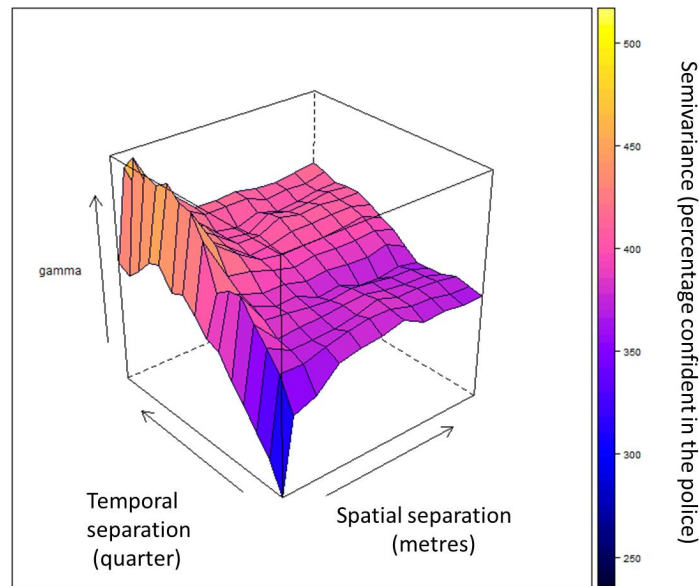


**Figure 7** 2.5D representations of a spatiotemporal variogram of confidence in the police

Spatiotemporal variability in the autocorrelation structure is visible in the 3D representation, Figure 8. The clear structure suggests that spatiotemporal autocorrelation is present. Particularly low semivariances can be seen at zero metres. This suggests that confidence of a borough neighbourhood can be confidently predicted for the next quarter, up to four quarters into the future. The presence of fairly low semivariance values up to 10,000 metres suggests that confidence values taken in one

borough neighbourhood can be used to improve predictions of confidence levels in areas up to 10,000 kilometres away.

**3D Spatiotemporal variogram of confidence in the police**



**Figure 8** 3D representations of a spatiotemporal variograms of confidence in the police

However, it should be noted that the sparsity of the survey data collected requires caution in interpreting these trends.

### 3. Conclusions

Exploratory spatiotemporal data analysis confirmed that public confidence in the police exhibits spatiotemporal heterogeneity and nonstationarity. Local models are preferable to global models particularly when modelling phenomena which are not stationary. The presence of spatiotemporal autocorrelation suggests that prediction of public confidence may be achieved. A model such as a local, dynamic spatiotemporal indicator kriging model may be able to achieve this. This model must be adapted to overcome the challenge of data sparsity. Various tools were used to examine the spatiotemporal autocorrelation structure of public confidence. Of the tools used the spatiotemporal variogram provided the most succinct picture of the autocorrelation structure and would be useful for parameterizing the model chosen.

### 4. Acknowledgements

This work is part of the project - Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks ([www.ucl.ac.uk/cpc](http://www.ucl.ac.uk/cpc)), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) is highly appreciated. Dawn is a Commonwealth Scholar, funded by the UK Government.

### 5. Biography

Dawn Williams is a member of the SpaceTimeLab for Big Data Analytics



(<http://www.ucl.ac.uk/spacetimelab>), at University College London. She is a second year PhD student attached to the Crime, Policing and Citizenship (CPC) Project. Her research interests include spatiotemporal data mining, visualization, big data analysis and sustainable development.

James Haworth is a lecturer in spatio-temporal analytics at the SpaceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimelab>), at University College London. His main interests lie in the analysis, modelling and forecasting of spatio-temporal data using machine learning methods.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimelab>), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

## 6. References

- Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geogr. Anal.* 27, 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x
- Fotheringham, A.S., 2009. “The Problem of Spatial Autocorrelation” and Local Spatial Statistics. *Geogr. Anal.* 41, 398–403. doi:10.1111/j.1538-4632.2009.00767.x
- Haining, R.P., Kerry, R., Oliver, M.A., 2010. Geography, Spatial Data Analysis, and Geostatistics: An Overview. 地理学、空间数据分析及地统计学：综述. *Geogr. Anal.* 42, 7–31. doi:10.1111/j.1538-4632.2009.00780.x
- Mayor’s Office for Policing and Crime, 2013. MOPAC Challenge Quarterly Performance Paper. Mayor’s Office for Policing and Crime.
- Mayor’s Office for Policing and Crime, 2014a. MOPAC Challenge Board: Confidence Transcript (Transcript).
- Mayor’s Office for Policing and Crime, 2014b. Public Confidence in Policing London: Public Attitude Survey (PAS) Frequently asked questions.
- Myhill, A., Bradford, B., 2012. Can Police Enhance Public Confidence by Improving Quality of Service? Results from Two Surveys in England and Wales. *Polic. Soc.* 22, 397–425. doi:10.1080/10439463.2011.641551

# Can the sentiment expressed in trail users' tweets help to assess the effectiveness of Environmental Stewardship Agreements? An exploratory analysis of the Pennine Way National Trail, England.

Tom Wilson<sup>\*1</sup> and Robin Lovelace<sup>†1</sup>

<sup>1</sup>School of Geography, University of Leeds

January 9, 2015

## Summary

This paper presents an exploratory analysis into the feasibility of using the sentiment expressed within trail users' Twitter messages to assess the effectiveness of Environmental Stewardship Scheme agreements in place along the Pennine Way National Trail, England.

**KEYWORDS:** GIS, big data, sentiment analysis, environmental stewardship, volunteered geographic information

## 1. Introduction

The Environmental Stewardship Scheme (ESS), an agri-environmental scheme, represents the most widespread approach to environmental management in England. ESS provides financial rewards to farmers and land managers in return for reductions in farming intensity and the adoption of measures to protect the surrounding environment. The success of early agri-environmental schemes was measured by levels of participation but more recently (and for the lifetime of ESS which was introduced in 2005-2006) the focus has shifted to analyse the environmental benefits provided under the scheme (Franks & Emery, 2013), for example with regard to landscape character (Natural England, 2014a), the enhancement of grassland, moorland and heath (Natural England, 2014b), bird populations (Davey et al., 2010), and the provision of ecosystem services (Defra, 2009).

England has 15 designated National Trails which pass through diverse landscapes and expanses of agricultural land, a substantial amount of which is managed under ESS. In 2012 approximately 12 million visits were made to England's National Trails (Ramblers, 2012). Despite the interactions that trail users unwittingly have with land managed under ESS there is currently no method to specifically obtain their opinions concerning the effectiveness of ESS in preserving and protecting the environment. The opinions of trail users are in fact limited to broad, large-scale qualitative surveys of visitors to the countryside in general, such as the Monitor of Engagement with the Natural Environment (MENE) which examines the adult population's engagement with the natural environment (Natural England, 2013). The National Trail User Surveys (The Countryside Agency, 2005; Natural England/Countryside Council for Wales, 2007) were discontinued in 2007.

This research aims to address this potential discontinuity of knowledge by exploring the feasibility of utilising the sentiment conveyed in trail users' Twitter messages (tweets) to assess the effectiveness of ESS specifically from the trail user perspective. This exploratory analysis will focus on the Pennine Way National Trail. We propose a methodology to geographically and lexically filter relevant tweets that originate from the proximity of the trail, and then perform sentiment analysis to extract the

---

<sup>\*</sup> gy10tlw@leeds.ac.uk

<sup>†</sup> r.lovelace@leeds.ac.uk

sentiment expressed. Finally we will determine whether this information can be used to assess the effectiveness of ESS.

## 2. Related research

Currently the effectiveness of ESS is measured with regard to the delivery of ES objectives and environmental benefits, for example in the provision of ecosystem services (Defra, 2009), the effects on landscape character (Natural England, 2014b), the ecological status of grassland, moor and heath (Natural England, 2014a), and the effects on bird populations (Davey et al., 2010). The Monitor of engagement with the Natural Environment (MENE) is a comprehensive nationwide survey which aims to measure the population's engagement with the natural environment (Natural England, 2013). The MENE survey is a representative but broad approach, and is not specific to trail users.

Sentiment Analysis (SA) is the process of extracting positive or negative opinions from text with the aim of analysing people's sentiment; this is seen as an important task, particularly since our opinions influence our behaviours, and this has led to a particular commercial interest (Liu, 2012). More recently however, the growth of the social web has led to an increased interest in SA of short informal text for the purposes of social research, to gain insights into specific events, and to study the affective dimension of the social web (Thelwall et al., 2012).

This research aims to address the potential shortfall of information regarding trails users' opinions by conducting SA of short informal text sent via the micro-blogging service Twitter. Specifically we explore whether the sentiment expressed within the tweets of trail users can be used to assess the effectiveness of ES agreements in delivering environmental benefits.

## 3. Study Area

The focus of this research is the 431 km (268 mile) long Pennine Way National Trail (PWNT), the oldest of England's 15 National Trails which officially opened on 24th April 1965 (Figure 1). Along its route the PWNT crosses agricultural land managed under ESS. For this analysis a 5km spatial buffer was drawn around the PWNT which will hereafter referred to as the PWNT corridor.

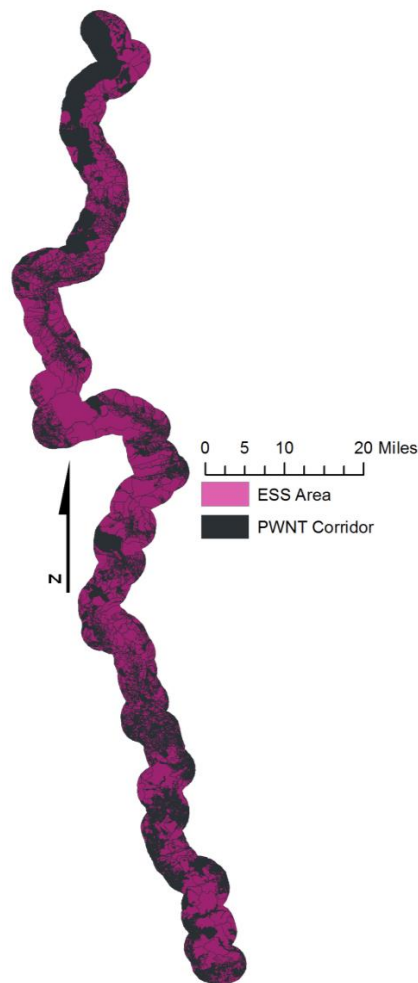


**Figure 1:** Location of the Pennine Way National Trail, England.

#### 4. Data

Geocoded Twitter data forms the basis of this research. The Twitter data were collected between 2014-06-03 and 2014-07-25 via the Twitter Streaming API, a service which provides limited access to the Twitter 'stream' (Lovelace et al., 2014). The full dataset consisted of 60,434 geocoded tweets which originated (i.e. were sent) from the region of the PWNT. Each row in the dataset represented a tweet and included the longitude and latitude of origin and the tweet's text. Additional metadata included a date and time stamp, and the sender's (tweeter's) number of followers, following, and tweets sent. The dataset did not include a twitter username.

A GPX file of the PWNT was obtained from the National Trails website (Walk Unlimited, 2014). A 5km spatial buffer was drawn in ArcMap (ESRI, 2011) to create the aforementioned trail corridor. A shapefile of ESS agreement boundaries for England were obtained from Natural England (Natural England, 2014c). These boundaries were spatially clipped to the PWNT corridor resulting in 1717 individual ESS agreements within the trail corridor, covering 74.09% of the land (Figure 2).



**Figure 2:** The PWNT Corridor and ESS agreement boundaries. © Natural England copyright. Contains Ordnance Survey data © Crown copyright and database right [2014]

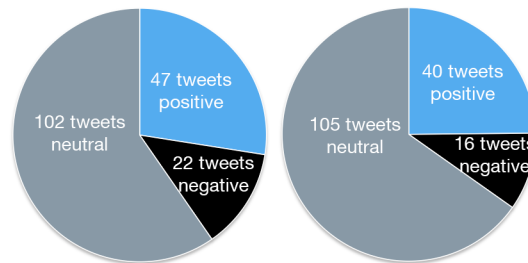
## 5. Methodology

Processing of the Twitter dataset was completed using the statistical package R version 3.1.2 (R Development Core Team, 2008). The Twitter data were spatially clipped to the PWNT trail corridor so that only tweets that originated from within the 5km PWNT corridor were included in further analysis. Regular expression terms were used to filter the tweets relevant to trail use, e.g. "Pennine Way" and "hike", plus some significant locations along the trail for example "Kinder Scout" and "Cheviots". In total 16 search terms were used. Further data 'cleaning' was performed to remove tweets such as broadcast traffic alerts and direct replies between Twitter users. This process resulted in 161 tweets for sentiment analysis.

Sentiment Analysis was conducted using SentiStrength, a lexicon-based sentiment analysis tool developed specifically for short informal text (Thelwall et al., 2010; Thelwall et al., 2012) such as that found on Twitter. SentiStrength returns an integer for both the positive (+1 to +5) and negative (-1 to -5) sentiment expressed within text. A value of 1 or -1 denotes an absence of positive or negative sentiment respectively. Overall sentiment is obtained by combining the two integers to determine sentiment polarity of the text: A combined score of 0 denotes neutral text.

## 6. Selected findings

Figure 3 provides a breakdown of the sentiment expressed in the twitter messages along the PWNT trail corridor. 47 tweets expressed positive sentiment, 22 negative sentiment and 102 expressed no sentiment (i.e. neutral). 10 tweets expressed both positive and negative sentiment. Overall tweet sentiment consisted 40 positive, 16 negative and 105 neutral.



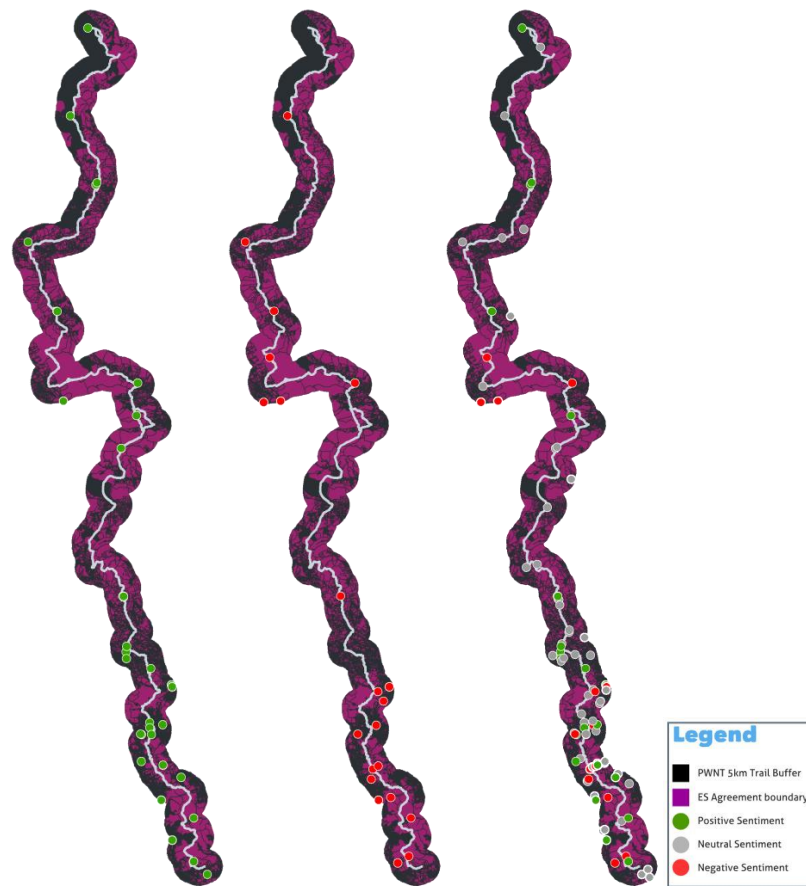
**Figure 3:** The sentiment expressed in trail users' tweets. Left: Expressions of sentiment in each tweet. Right: The overall sentiment of the tweets

Figure 4 provides a spatial overview of the sentiment of tweets along the PWNT corridor. 7 overall positive and 3 overall negative tweets originated from within an ESS boundary.

The majority of the 161 tweets are neutral and did not express sentiment. Further investigation revealed that of the 105 overall neutral tweets 94 (90%) contain a URL to an external source such as an image or website.

## 7. Conclusions

This research has demonstrated a methodology to filter relevant tweets and perform a sentiment analysis on short informal texts sent by trail users along the Pennine Way National Trail. With regard to the feasibility of assessing the effectiveness of ESS agreements the initial findings are limited, both in terms of the number of relevant tweets and the sentiment conveyed. Additional work is needed to collect additional data, perhaps over a longer period of time. A key finding is that a high percentage of neutral tweets contained a URL to an image and it could be that these images are intended to convey sentiment. Further research should therefore focus on developing methods to extract the sentiment, if any, conveyed in these images.



**Figure 4:** Sentiment of tweets originating from within the PWNT corridor. Left: tweets expressing positive sentiment Middle: tweets expressing negative sentiment. Right: the overall sentiment of tweets. © Natural England copyright. Contains Ordnance Survey data © Crown copyright and database right [2014]

## 8. Acknowledgements

Special thanks to Dr. Andrew Evans and Dr. Paul Norman for their continued support and counsel; and to Phillapa Swanton, Hazel Thomas, and Steven Westwood of Natural England.

## 9. Biography

Tom Wilson is a GIS Online Distance Learning master's degree student at the University of Leeds, and an early-career researcher who also works as a program director of a non-profit conservation group in Arizona. His interests are in the applications of GIS, big data analysis, and geovisualisation.

Dr. Robin Lovelace is a geographer and environmental scientist, and currently a TALISMAN researcher at the University of Leeds specialising in methods of spatial microsimulation.

## 10. References

- Davey, C. M., Vickery, J. A., Boatman, N. D., Chamberlain, D. E., Parry, H. R., & Siriwardena, G. M. (2010). Assessing the impact of Entry Level Stewardship on lowland farmland birds in England. *Ibis*, 152(3), 459-474.
- Department of Food and Rural Affairs. (2009). Provision of ecosystem services through the Environmental Stewardship Scheme. Defra project code NR0121
- ESRI. (2011). ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute
- Franks, J. R., & Emery, S. B. (2013). Incentivising collaborative conservation: Lessons from existing environmental Stewardship Scheme options. *Land Use Policy*, 30(1), 847-862.
- Lawton, J.H., Brotherton, P.N.M., Brown, V.K., Elphick, C., Fitter, A.H., Forshaw, J., Haddow, R.W., Hilborne, S., Leafe, R.N., Mace, G.M., Southgate, M.P., Sutherland, W.A., Tew, T.E., Varley, J., & Wynne, G.R. (2010) Making Space for Nature: a review of England's wildlife sites and ecological network. Report to Defra.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), pp1-167.
- Lovelace, R., Malleson, N., Harland, K., & Birkin, M. (2014). Geotagged tweets to inform a spatial interaction model: a case study of museums. [online] Available at: <http://arxiv.org/abs/1403.5118>
- Natural England. (2013). Monitor of Engagement with the Natural Environment: The national survey of people and the natural environment. Annual Report from the 2011-12 survey. Natural England Commissioned Reports, Number 122
- Natural England. (2014a). Assessment of the effects of Environmental Stewardship on landscape character. Correlative analysis of datasets to assess the degree of success in delivery of Environmental Stewardship objectives. Natural England Commissioned Reports, Number 158
- Natural England. (2014b). Assessment of the effect of Environmental Stewardship on improving the ecological status of grassland, moorland and heath. Natural England Commissioned Reports, Number 156
- Natural England. (2014c). Environment Stewardship Scheme Agreements. [shapefile] Created by Natural England. August 2014. [online] Available at: <http://www.geostore.com/environment-agency/WebStore?xml=environment-agency/xml/ogcDataDownload.xml>
- Natural England/Countryside Council for Wales. (2007). Results of the national trail user survey 2007.
- Ramblers. (2012). National Trails: A fantastic future. [online] Available at: <http://www.ramblers.org.uk/what-we-do/news/2012/september/securing-a-fantastic-future-for-national-trails.aspx> [Accessed 7 January 2015].
- R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <http://www.R-project.org>
- The Countryside Agency. (2005). Results of the national trail user survey 2005.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.* (63)1, pp163-173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), pp2544-2558.

Walk Unlimited (2014). The Pennine Way. [online] Available at: <http://nationaltrail.co.uk/pennine-way> [Accessed 7 January 2015].



# A Model Officer: An Agent-based Model of Policing

Sarah Wise\*<sup>1</sup> and Tao Cheng<sup>1</sup>

Department of Civil, Environmental, and Geomatic Engineering  
University College London

November 7, 2014

## Summary

The way police officers create guardianship is poorly understood, in part because of the complexities of policing. However, in order to understand how to advise the police, researchers must have an understanding of how the current system works. The work presents an agent-based model that simulates the movement of police vehicles, using a record of calls for service to emulate the demands on the police force. The GPS traces of the simulated officers are compared with real officer movement GPS data in order to assess the quality of the generated movement patterns.

**KEYWORDS:** agent-based modelling, crime, policing, GIS

## 1. Introduction

The term *guardianship* is a criminological concept that refers to the way guardians, such as property owners and the police, prevent potential offenders from committing crimes (Cohen and Felson, 1979). When potential offenders are choosing whether to commit a crime, they consider how likely they are to be apprehended or stopped by any fellow citizens or police officers in their immediate area (Kleck and Barnes, 2008). The offender's choice to offend is therefore based in part on his or her interactions with other people, and the higher-level crime patterns that result from the choices of all of the individual offenders are influenced by the physical presence (or absence) of police and citizens. Thus, guardianship depends on the spatial, temporal, and behavioural interactions of offenders, citizens, and police officers.

However, the way police create guardianship is not obvious. Policing is a complex, culturally specific process, and officers have many intersecting responsibilities (Policy Studies Institute, 1996). While Robert Peel identified the prevention of crime and disorder as the first goal of policing (Home Office, 2012), police forces are also asked to help with finding missing persons, providing security at public festivals, and handling traffic accidents (Metropolitan Police, 2014). For researchers attempting to influence crime rates by suggesting police policy, ignorance of the way these other commitments constrain officer presence and movement will result in policy suggestions divorced from reality.

Researchers have historically failed to consider these complications in their models of policing and guardianship. Throughout the literature, very few models of guardianship creation exist, and those that do explore it in trivial ways. In general, the issue these models have faced is two-fold: firstly, that many methodologies cannot aggregate lower-level behaviours in order to understand the purposive behaviour of a system comprised of many individuals, and secondly that there is an absence of data that could support such a model. A

---

<sup>1</sup> s.wise@ucl.ac.uk

number of simulations emulate the process of officers carrying out responsibilities as part of a larger group, but ignore the many complicating factors, such as the purposeful movement of officers and the fact that they have other responsibilities (e.g. Birks et al., 2012; Groff, 2007). Further, simulated officers are unimpeded by the time-consuming process of actually dealing with offenders (e.g. Melo et al., 2006; Dray et al., 2008). These simplifications significantly bias officer movement patterns, generating patterns of guardianship that do not match the guardianship created by real officers.

In all of these cases, researchers have been hampered by a lack of access to information about officer duties, incapable of incorporating the complexities of policing into their simulations for want of data. As a result of our working relationship with the London Metropolitan Police, we have access to this kind of information. This work presents a simulation which seeks to capture the complex realities of policing, using a combination of data and behavioural research to create a realistic model of police activity. Given that policing is complex, spatial, temporal, and profoundly influenced by human decision-making, we utilise an agent-based model (ABM).

## **2. The Model**

The model presented here utilises an agent-based framework to explore how officers translate their assignments into movement and actions in the context of the environment in which they find themselves. As a methodology, ABM has been particularly successful in incorporating criminological concepts such as routine activity theory (Cohen and Felson, 1979), rational choice theory (Cornish and Clarke, 1987), and crime pattern theory (Brantingham and Brantingham, 1984) into simulations (e.g. Groff, 2006; Birks et al., 2012; Malleson et al., 2012).

In this work, the ABM attempts to capture the behaviours of Metropolitan Police constables, simulating them at the level of the vehicles to which they are assigned. The simulation models the vehicles moving over a road network, specified with 1m<sup>2</sup> resolution, and are updated on a temporal scale of one minute per simulation step. The model framework is built in Java, using the MASON simulation toolkit, an open-source multiagent simulation library. The following sections will describe the environment in which the vehicles exist, the way vehicles are represented in the simulation, and the way vehicle behaviours are translated into actions.

### **2.1 Environment**

In order to represent the environment in which police vehicles exist, the model combines information about the real-world road network with records of calls for service from the community. The model was tested with road network data derived from the Ordnance Survey MasterMap Integrated Transport Network Road (ITN) dataset. The locations of police stations, which factor into the activities of the police vehicles, are taken from the data provided to us by the Met Police. The locations of traffic lights, which impose a time cost on the movements of officers throughout the environment, is taken from data provided by Transport for London (TfL).

In addition to the physical constraints of the environment, vehicles are influenced in how they move by the calls for service that they receive from the general public. The timing and location of these calls for service are drawn from the records of the Call Aided Despatch (CAD) system of the Met Police, and these incidents are used to direct officers to the real-

world locations of incidents at the appropriate times. Thus, the pressures and constraints upon the officers are rendered in a simulated setting.

## 2.2 Agents

The model represents the actions of police officers in terms of the movement and interactions of police vehicles. Vehicle agents have attributes that inform their actions. Table 1 provides an overview of the attributes that characterise the Vehicle agent at any given point in time, specifying the range of values these attributes may take on and providing examples of such values. In particular, Vehicles have a current location in space, a home station, a unique call sign, a current activity, and a current status, as well as a Tasking object.

**Table 1.** Vehicle attributes

Attribute	Possible Values	Example Value
Current Location	Point in space	(3487, 2387)
Home Station	Point in space	(3487, 2387)
Call Sign	String	EK8N
Tasking	Tasking Object	Response Tasking
Current Activity	Patrolling, Occupied, On Way to Tasking, On Way to Station, Waiting	Patrolling
Current Status	Available, Off Duty, Occupied, Meal Break	Available

All officers obey a daily schedule of returning to their home station every eight hours, to simulate the change in officers. In addition to this shared structure, individual vehicles are assigned to “taskings”, or assignments of duty. The vehicle’s assignment corresponds to the real-world process of assigning officers to carry out different tasks during the day, as is arranged during the briefing before a shift begins. These assignments dictate who will respond to calls, who will focus on patrolling, who will be responsible for coordinating with other officers in order to pick up offenders, and so forth. They structure the officer’s day, and dictate how he progresses from one activity to another. The existing assignments are:

- **Reporting:** the vehicle is primarily responsible for responding to non-urgent calls for service. It will move around the environment in response to these calls, spending time dealing with the caller when it reaches the site of the incident. The vehicle will spend any unoccupied time patrolling, which is modelled here as moving randomly about the environment.
- **Transporting:** the vehicle is responsible for coordinating with other vehicles who have detained suspects. When the Transporting vehicle receives a request of transport, it will move to the suspect and then transport him back to the police station.
- **Responding:** the vehicle is responsible for responding to urgent calls for service. When it receives a call, the responding vehicle begins to move at a faster speed and ignores streetlights, both in its calculation of the shortest path and in its movements. When it reaches the site of the incident, it spends time assessing the incident, potentially apprehending an offender. If an offender is apprehended, the vehicle will call for a Transport vehicle and wait until it arrives. It will spend any unoccupied time patrolling.



**Figure 1.** Normalised heatmaps showing the road usage associated with the real data (A), the random-patrolling model (B), and the tasking-service model (C).

The vehicles communicate with an object called the Despatcher, which informs them of incidents based on the data fed into the simulation and coordinates among the agents, taking a request for transport from a Responding vehicle and transmitting the request to a Transport vehicle. Vehicles plan the shortest path to a point in terms of time, and obey traffic lights except in the case of responding vehicles moving to an urgent request.

### **3. Results**

In order to investigate how the existence of assignments influences the behaviour of simulated agents, we compare two models: one in which all the vehicles spend their time patrolling and receive no calls for service, and another in which vehicles are assigned taskings and receive calls for service as drawn from the real CAD data. The first model emulates the existing ABMs of police movement, while the second represents our contribution.

Figure 1 shows the real road usage data compared with the random-patrolling model and the tasking-service model. Briefly, the real data shows officers making greater use of major roads. The random-patrolling model demonstrates a much less concentrated focus on these roads, while the tasking-service model is more concentrated. However, the tasking-service model is more focused in the south and centre of the region than is the real data. The simulated models generate many more records than exist in the real data, which poses interesting questions about how to compare synthetic versus real data.

### **4. Conclusions**

The simulation generates interesting results, with the behavioural model generating more realistic patterns of road usage than the commonly utilised random movement model. More questions exist with regard to comparing the real data with the generated data, and suggest further investigations into the growing field of ABM validation efforts. The work presented here both addresses the lack of nuanced simulations of policing and pushes forward the practice of inserting real-world data into simulations in order to emulate rich environments for behaviourally complex agents. In the future, this ABM will allow us to explore counterfactual situations, comparing the projected effectiveness of different policing strategies.

### **5. Biography**

Sarah Wise is a postdoctoral researcher in the Department of Civil, Environmental, and Geomatic Engineering at University College London. Her research interests include agent-based modelling, social network analysis, data mining, and geographical information systems, and her past research has dealt with crisis situations, health, crime, and social media.

Tao Cheng is a Professor in GeoInformatics, and Director of SpceTimeLab for Big Data Analytics (<http://www.ucl.ac.uk/spacetimelab>), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, visualisation and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

## References

- Birks, D., Townsley, M., & Stewart, A. (2012). *Generative Explanations of Crime: Using Simulation To Test Criminological Theory*. *Criminology*, 50(1), 221–254.
- Brantingham PJ and Brantingham PL, 1984, *Patterns in crime*. Macmillan, New York, NY.
- Cornish D and Clarke R, 1987, *Understanding crime displacement: An application of rational choice theory*. *Criminology*, 25(4): 933–947.
- Cohen L and Felson M, 1979, *Social Change and Crime Rate Trends: A Routine Activity Approach*. *American Sociological Review*, 44(4): 588–608.
- Dray, A., Mazerolle, L., Perez, P., & Ritter, A. (2008). *Policing Australia's "heroin drought": using an agent-based model to simulate alternative outcomes*. *Journal of Experimental Criminology*, 4(3), 267–287.
- Groff, E. R. (2007). "Situating" *Simulation to Model Human Spatio-Temporal Interactions: An Example Using Crime Events*. *Transactions in GIS*, 11(4), 507–530.
- Home Officer. (2012). *Policing by Consent*.  
<https://www.gov.uk/government/publications/policing-by-consent>
- Kleck, G., & Barnes, J. C. (2008). *Deterrence and Macro-Level Perceptions of Punishment Risks: Is There a "Collective Wisdom"?* *Crime & Delinquency*, 59(7), 1006–1035.
- Melo, A., Belchior, M., & Furtado, V. (2006). *Analyzing Police Patrol Routes by Simulating the Physical Reorganization of Agents*. In J. S. Sichman & L. Antunes (Eds.), *Multi-Agent-Based Simulation VI* (pp. 99–114). Springer Berlin Heidelberg.
- Metropolitan Police. (2014) *Specialist Crimes and Operations*,  
<http://content.met.police.uk/Site/specialistcrimeoperations>

# Estimates of ethnic mortality in the UK revisited

Wohland P<sup>\*1</sup> and Rees P<sup>†2</sup>

<sup>1</sup>Institute for Health and Society, Newcastle University, Biogerontology Research Building, Campus for Ageing and Vitality, Newcastle, NE4 5PL United Kingdom

<sup>2</sup>School of Geography, University of Leeds, Leeds, LS2 9JT United Kingdom

March 20, 2015

## Summary

The ethnic diversity of the UK population is increasing rapidly. Between 1991 and 2011 the share of the population defining themselves as not White increased from 7% to 14%. Still, information on mortality for ethnic groups, an important population health indicator, is not routinely collected everywhere the country.

Previously, we developed the first estimates of ethnic mortality for local areas in the UK for 2001 using information on the geographical distribution of the ethnic populations as well as health and vital statistics information and found profound variations across groups. Because these previous estimates have been challenged by subsequent work, in this paper we review methods and results of a literature starting in 1984 before embarking on new estimation.

**KEYWORDS:** Ethnic mortality, Mortality by country of birth, England and Wales, United Kingdom, Local Areas

## 1. Introduction

The ethnic composition of the UK population is changing. Diversity has steadily increased since 1991 and the proportion of people defining themselves as not White has doubled between 1991 and 2011 and increased from 7% to 14% of the total population. The part of the population who define themselves as not White British was almost 20% in 2011 (Jivraj, 2012; Rees et al., 2009). However, unlike other immigration countries such as the USA, Australia or New Zealand, information on mortality for ethnic groups, an important population health indicator, is still not routinely collected everywhere the country. Only Scotland, which has a lower ethnic density compared to other parts of the country, had the far-sightedness and have started collecting ethnic information on the death certificate in 2012 (Christie, 2012).

## 2. Ethnic groups life expectancy estimates 2001

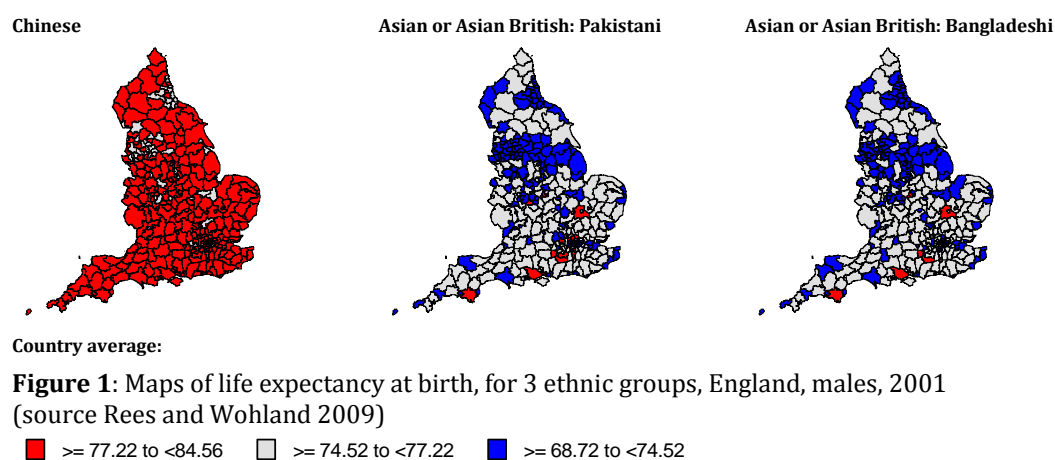
Previously, in the course of projecting the future ethnic population for UK local areas (Rees et al., 2009) we developed the first ethnic mortality estimates for the UK and English and Welsh local areas for 2001 using information on the geographical distribution of ethnic populations as well as health

---

\* Pia.Wohland@ncl.ac.uk

† P.H.Rees@leeds.ac.uk

and vital statistics information. Profound differences between groups and across space were found. Figure 1 shows maps presenting life expectancy for three ethnic groups at the opposite spectrum of expected length of life. Chinese men born in 2001 could expect to live between 77 and 85 years in most parts of the UK. On the other hand, Pakistani and Bangladeshi men, born in 2001 could only expect to live between 75 to 77 years in about half of local areas and just between 68 and 75 years in the other half. Only in a few places could these groups expect to reach the same age as the Chinese group. We found that life expectancies (male and female) for many ethnic minorities were below those of the White British. Life expectancies for ethnic minorities were related to the degree of deprivation groups experienced, counter-acted in part by recent arrival.



### 3. Discussion of mortality estimates for migrants and ethnic groups in the UK

Before our estimation of ethnic group mortalities, similar previous research concentrated on immigrant mortality. Immigrant mortality in England and Wales has been monitored since the 1971 Census when country of birth was recorded for the first time. This in combination with the country of birth information on the death certificate made it possible to study the mortality experience of people by nativity, that is, of mainly first generation immigrants. Various studies (see Table 1) have used this information to compare mortality experience between different immigrant groups.

The different studies of immigrant or ethnic group mortality used diverse mortality measures, ranging from standardised mortality ratios to relative risks of dying, hazard ratios of mortality and life expectancy. In addition, different reference populations -White British or total population- were used as well as different ethnic group categories examined. Most studies also varied in terms of the age range they looked at. For these reasons a comprehensive comparison of the results is difficult.



When reviewing literature we therefore examined only whether a study found better or worse or similar health outcomes compared to the reference group. Results are graphically represented in Figure 2. We refer to the different mortality measures as health experience. Two studies did not supply confidence intervals for their results Marmot et al. (1984) and Rees et al. (2009), for this reason, only show better or worse health compared to the reference population. Other studies often did not find significant differences, which in part might be the result of using small sample populations.

In this paper we will discuss common findings and recommendations resulting from the literature review. For example, the studies in general agreed that Irish and Scottish people had poorer health compared to the reference population and that other Europeans/White Other had better health. On the other hand, there are groups where findings varied. One example here is the Caribbean group. Whereas Marmot et al. (1984) found better health for Caribbean migrants if they were men and worse if they were women, Scott and Timaeus (2013), who did not distinguish by gender found better health for the foreign born Caribbean group and worse for UK born Caribbean. Rees et al. (2009) who distinguished by gender, but looked at ethnic groups, that is first and later generations combined, found worse health for both gender. Observations like these suggest future research into ethnic group health should include foreign and UK born categories.

Table 1 Various studies since the 1970s investigating the mortality experience of migrants to the UK or UK ethnic groups

Study (in order of year of publication)	Measure	Groups		Time frame
Marmot, M. G., Adelstein, A. M., & Bulusu, L. (1984). Lessons from the Study of Immigrant Mortality. <i>Lancet</i> , 1(8392), 1455-1457.	PMRs SMRs	Ireland Poland Italy Indian-subcontinent Caribbean	Indian-Indian Indian-British Africa Europe	Country of birth 1970-72
Wild, S., & McKeigue, P. (1997). Cross sectional analysis of mortality by country of birth in England and Wales, 1970-92. <i>British Medical Journal</i> , 314(7082), 705-710.	SMRs	Scotland Ireland East Africa (68% South Asian) West Africa (73% Black African) Caribbean South Asia		Country of birth 1970-1992 Two time points (1970-1972) (1990-1992)
Wild, S. H., C. Fischbacher, A. Brock, C. Griffiths, and R. Bhopal. 2007. "Mortality from All Causes and Circulatory Disease by Country of Birth in England and Wales 2001-2003." <i>Journal of Public Health</i> 29 (2): 191-198. doi:10.1093/pubmed/fdm010	Indirect SMRs	E&W Scotland Ireland Eastern Europe East Africa West Africa	West Indies Middle East Bangladesh India Pakistan China and Hong Kong	Country of birth 2001-2003
Rees, P. H., Wohland, P. N., & Norman, P. D. (2009). The estimation of mortality for ethnic groups at local scale within the United Kingdom. <i>Social Science &amp; Medicine</i> , 69(11), 1592-1607. doi: 10.1016/j.socscimed.2009.08.015	LE and SMR	16 Ethnic groups (Census definition England and Wales, 2001)		Ethnic group 2001
Scott, A. P., and I. M. Timaeus. 2013. "Mortality Differentials 1991-2005 by Self-reported Ethnicity: Findings from the ONS Longitudinal Study." <i>Journal of Epidemiology and Community Health</i> 67 (9): 743-750. doi:10.1136/jech-2012-202265	Relative risk of dying Also compares foreign born to UK born	White Black Caribbean Black African Other Black Indian	Pakistani Bangladeshi Chinese Other Asian Other	Ethnic group 1991-2005
Wallace, M., & Kulu, H. (2014). Low immigrant mortality in England and Wales: A data artefact? <i>Social Science &amp; Medicine</i> , 120, 100-109. doi: DOI 10.1016/j.socscimed.2014.08.032	Hazard ratios of mortality	E&W Scotland NI Ireland India Pakistan Bangladesh Jamaica Other Caribbean	E&S Africa W&C Africa W Europe E Europe China Other Asia Rest of World Unresolvable	Country of birth 1971-2001
Morris, M., Woods, L. M., & Rachet, B. (2015). A novel ecological methodology for constructing ethnic-majority life tables in the absence of individual ethnicity information. <i>Journal of Epidemiology and Community Health</i> . doi: 10.1136/jech-2014-204210	Majority population life table	White Black Asian		Ethnic group 2001

First Author	Marmot	Wild	Wild	Scott	Wallace		Morris
Ages	20-69 20-69 15-64	20-69 20-69	20+ 20+	1-79 1-79 1-79	20+	0-100	1-80
Gender	Men Women Men	Men Women	Men Women	All All All	ALL	Men Men Women Women SIR GWM SIR GWM	Men Women
Additional Info				Born inUK Born abroad			
Measure	SMRs SMRs SMRs	SMRs SMRs	SMRs SMRs	Relative risk of dying	Hazard Ratio	LE LE LE LE	LE LE
White	Ireland Poland Italy Europe Indian-British	Scotland Ireland	E&W Scotland Ireland Eastern Europe	White	E&W Scotland NI Ireland W Europe E Europe	ALL WBR WIR OWH	White
Asian	Indian-subcontinent	South Asia East Africa (68%)	East Africa Bangladesh India Pakistan China and Hong Kong	Indian Pakistani Bangladeshi Chinese Other Asian	India Pakistan Bangladesh China Other Asia	IND PAK BAN OAS CHI	Asian
Black	Caribbean Africa	Caribbean West Africa (73%)	West Indies West Africa	Black Caribbean Black African Other Black	Jamaica Other Caribbean E&S Africa W&C Africa	BCA BAF OBL	Black
Other	Better= Same= Worse= Compared to standard		Middle East North Africa	Other	Best of World Unresolvable	OET WBC WBA WAS OMI	

Figure 2 Graphical representations how ethnic/migrant group health compared to reference categories (White British or total population)

Notes: SIR = method uses standardised illness ratios, GWM = geographically weighted method

Wild et al. (2007) and Rees et al. (2009) also found significant health differences within broad groups, especially the Asian group. This confirms the importance of using finer groupings, wherever possible.

In the 1970s, when research into the health of migrant minorities began, it was effective to research the mortality of first generation immigrants to establish health inequalities between the minority populations compared to the majority population. Nowadays, many second, third and fourth generation descendants of post WW2 immigrants live in the UK. These people's country of birth is the UK, but their ethnicity is "not White British". Various small cohort studies show that specific health outcomes vary by ethnic group (in which first and later generations are combined). This suggests that ethnic groups overall have different mortality experiences. Alas, the data to prove this are not available. In the ESRC funded NewETHPOP project we will provide further evidence on mortality differentials of UK ethnic groups and also for the first time show how ethnic groups' mortality changed over time, using the newest census information from 2011, analysing the mortality of ethnic-nativity (born in the UK/born outside the UK) groups. New methods will be needed to generate life tables for these groups from available data.

## Acknowledgements

The ethnic mortalities presented were results from the ESRC Funded project, *Ethnic group population trends and projections for UK local areas: dissemination of innovative data inputs, model outputs, documentation and skills*, 1 October 2010 to 30 September 2011, ESRC Research Award RES-165-25-0162.

## Biography

Dr Pia Wohland is a Senior Research Associate at the Institute for Health and Society (IHS) and Newcastle University Institute for Ageing (NUIA). She has researched ethnic mortality and health differences for local areas in the UK and developed software for ethnic population projections.

Philip Rees is Emeritus Professor of Population Geography at the University of Leeds, with interests in ethnic population projections, health outcomes and ageing of the population.

## References

- Christie, B. 2012. *Scotland introduces record of ethnicity on death certificates*.
- Jivraj, S. 2012. *How has ethnic diversity grown 1991-2001-2011?* Manchester: Centre on Dynamics of Ethnicity, The University of Manchester.
- Marmot, M.G. et al. 1984. Lessons from the Study of Immigrant Mortality. *Lancet*. **1**(8392), pp.1455-1457.
- Morris, M. et al. 2015. A novel ecological methodology for constructing ethnic-majority life tables in the absence of individual ethnicity information. *Journal of Epidemiology and Community Health*.
- Rees, P. and Wohland, P. (2009) Estimates of ethnic mortality in the UK. *Working Paper 08/04*, School of Geography, University of Leeds, Leeds, UK. Online at: <http://www.geog.leeds.ac.uk/fileadmin/documents/research/csap/08-04.pdf>
- Rees, P.H. et al. 2009. The estimation of mortality for ethnic groups at local scale within the United Kingdom. *Social Science & Medicine*. **69**(11), pp.1592-1607.
- Scott, A.P. and Timaeus, I.M. 2013. Mortality differentials 1991-2005 by self-reported ethnicity: findings from the ONS Longitudinal Study. *Journal of Epidemiology and Community Health*. **67**(9), pp.743-750.
- Wallace, M. and Kulu, H. 2014. Low immigrant mortality in England and Wales: A data artefact? *Social Science & Medicine*. **120**, pp.100-109.
- Wild, S. and McKeigue, P. 1997. Cross sectional analysis of mortality by country of birth in England and Wales, 1970-92. *British Medical Journal*. **314**(7082), pp.705-710.
- Wild, S.H. et al. 2007. Mortality from all causes and circulatory disease by country of birth in England and Wales 2001-2003. *Journal of Public Health*. **29**(2), pp.191-198.

# Designing 3D Geographic Information for Navigation Using Google Glass

Kelvin Wong<sup>\*1</sup> and Claire Ellul<sup>†2</sup>

<sup>1</sup>Department of Computer Science, University College London, London, UK, WC1E 6BT

<sup>2</sup>Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK, WC1E 6BT

January 05, 2015

## Summary

No longer bound by traditional 2D physical representations, there is a steady shift towards three-dimensional (3D) data. Existing research recognises landmarks to be important navigational but specific geometric and semantic attributes in 3D have not been identified. This study offers a user-centred investigation into assessing of the saliency of environmental objects which facilitate pedestrian navigation. A novel real-world navigation experiment using Google Glass is carried out with fourteen participants. Results show geometric and semantic detail for navigation are most pertinent between 1.65 – 7.5m for buildings. Visual characteristics such as colour, shape and texture are more relevant than function and use.

**KEYWORDS:** 3D GIS, Navigation, Google Glass, Landmarks, User-Centred Design

## 1. Introduction

Navigation is an implicit requirement of our daily lives. In recent years, the way we traverse the world has been strongly impacted by the emergence of mobile navigational technologies, altering our perception of space and specifically, our navigational strategies. A complex and ever-evolving phenomena, the study of navigation has a long history in a diverse number of fields ranging from psychology to geography to computer science. Navigation experiments, however, remain predominantly in virtual and simulated environments. There is a lack of real world experiments examining human navigation behaviour, perhaps due to the lack of control and higher costs incurred. Where existing studies have found landmarks to be the most salient feature in an urban landscape for navigation (May et al., 2003), saliency of the 3D geometric and semantic attributes of these landmarks have not been identified. This study aims to address the above issues by exploring landmark saliency at a finer, intrinsic level through a real-world navigation experiment. The paper will outline the design of the user-centric experiment and the subsequent results from its first iteration. It will conclude with possible directions for future work.

## 2. Methodology

In order to capture the true dynamism of real navigation, fourteen subjects were asked to navigate in the real-world rather than in a computer-simulated environment. A novel approach using a pair of Google Glass was implemented to record the gaze and movement of the participants. The device is minimalist and light optical head-mounted display, allowing for unobtrusive and natural tracking. The experiment was carried out around UCL and the Bloomsbury, an area of predominantly residential and office spaces.

Participants were instructed to follow a specified route on the map provided. The selected route (1.518km) was designed to test the navigation strategies of the participants and passed through an area

---

<sup>\*</sup> kelvin.wong.11@ucl.ac.uk

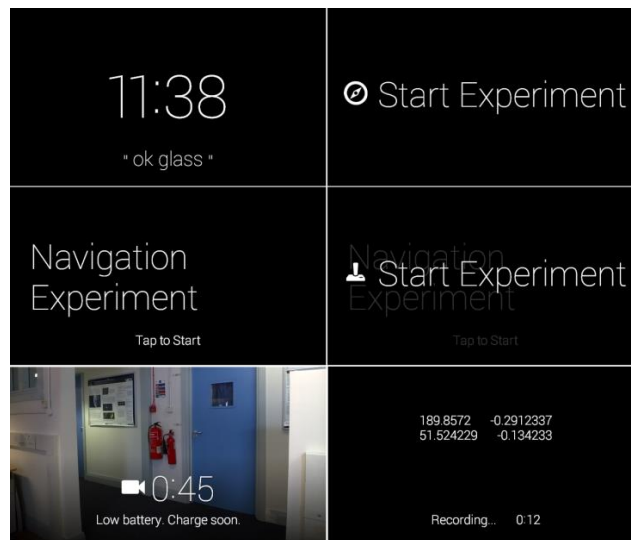
<sup>†</sup> c.ellul@ucl.ac.uk

without with major branding or obvious iconic buildings. By choosing a complex and diverse route, a significant navigation task demand was imposed on the participants and promoted the description of geometric and semantic features of navigational landmarks. Participants were told that the task was not to navigate the route most successfully or the fastest but were asked to produce a set of written instructions of the route for an unfamiliar traveller who had not been to the area before. This was to promote active thinking and to engage with the process of navigation. The participants were then fitted with Google Glass, presented with the route map and were allowed to navigate freely (Figure 1).



**Figure 1** Participants wearing Google Glass and navigating

A Google Glass application was specifically designed and written for this experiment. Developed in Android Glass Development Kit Preview 4.4.2, the application runs on Google Glass XE 19.1 and was paired with a smartphone (iPhone 5S) in order to obtain locational information via GPS. The user does not directly interact with the Google Glass, but rather it passively records a first person video as well as tracking their gaze vector (orientation and pitch), location (latitude and longitude) and elapsed time. The video was recorded concurrently while the gaze vectors and location were logged every 0.125 seconds. Figure 2 shows the various screens of the application during the experiment.



**Figure 2** Screenshots of Google Glass application flow

The experiment collected two sets of data: 1) qualitative written instructions and; 2) quantitative gaze vector tracking.

[illegible]

The results showed that road names and landmarks feature heavily. This is consistent with existing literature on terms used in navigation instructions (May et al., 2003). It is key to note here the use of OS StreetView as a base for the route map may have led to a distorted high usage of road names and road signs. It can be concluded, however, where buildings are visually homogenous and there is a lack of obvious landmarks, street names become more important, especially for egocentric navigation strategies using personal directional instructions.

694

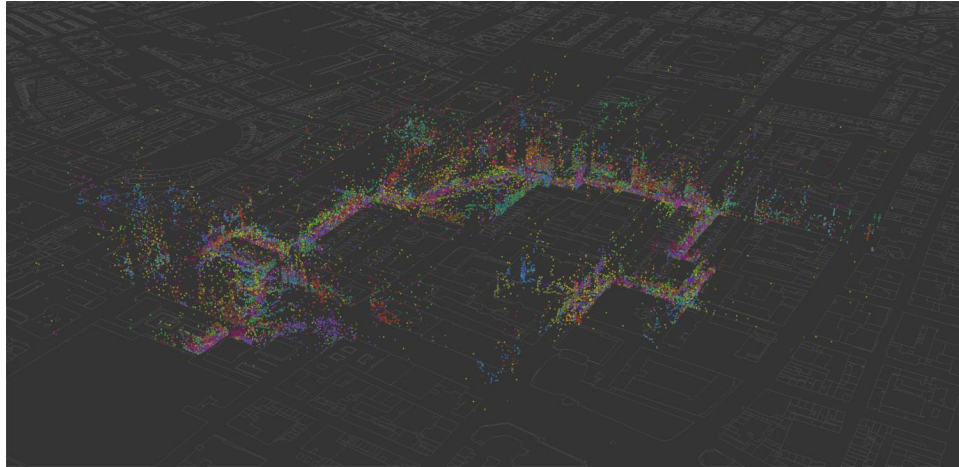
**Table 1** Summary table of words preceding and following “building”

Description	Count	Description Type	Description	Count	Description Type
red	7	Visual	UCL	1	Semantic
glass	5	Visual	cube	1	Visual
columned	3	Visual	cream	1	Visual
large	3	Visual	Rubin	1	Semantic
brick	3	Visual	University of London	1	Semantic
big	3	Visual	unusual	1	Visual
classical style	2	Visual	slatted	1	Visual
grey	2	Visual	UCL Women's Health	1	Semantic
stone	2	Visual	green	1	Visual
huge	2	Visual	fancy	1	Visual
UCL Hospital	2	Semantic	old	1	Visual
pointy	2	Visual	giant	1	Visual
black	1	Visual	grand	1	Visual
quaker's	1	Semantic	UCL Engineering	1	Semantic
impressive	1	Visual	small	1	Visual
church-like	1	Visual	white	1	Visual
dark	1	Visual	Grant Museum of Zoology	1	Semantic
single-storied	1	Visual			

In this study, a gaze vector is defined as a geodesic line which most accurately represents the shortest distance between the participant location and the gaze position. The data collected from the participants were corrected for GPS error using individually digitised routes from the video. The median adjustment of 5.773m is consistent with the average median error of 8m for iPhone’s integrated positioning technologies (Zandbergen, 2009). Each data point was also corrected to have a viewing height of 1.65m and the gaze vectors were then mapped using the orientation, pitch and locational data collected from the Google Glass. An assumption was made that the gaze position is calculated to be the first building or built structure from OS MasterMap<sup>‡</sup> that the gaze vector intersects within 200m (Figure 4). This was necessary as the device lacked true eye-tracking.

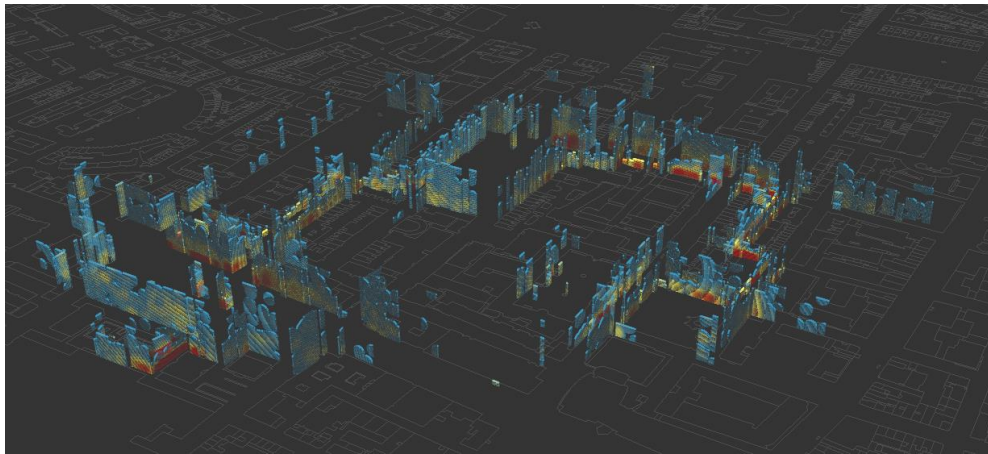
<sup>‡</sup> OS MasterMap Topography Layer Building Height Attribute was used to supplement this process.





**Figure 4** Gaze Positions of All Participants mapped in 3D

Initial exploratory statistics showed that gaze height was predominantly between 1.65 – 7.5m, with navigators focusing around eye-level to the first two stories. Roof characteristics, though useful for completeness and shape, are largely irrelevant pedestrian navigation as they could not be seen. The ideal subsequent analysis for this study would be to intersect gaze position points with a true 3D model, thereby identifying the exact features viewed while navigating every 0.125s i.e. the window on the ground floor or a street sign at an intersection. This, however, was not possible as a true 3D building city model with a high enough geometric and semantic detail was not available. This limitation stems not from the proposed methodology or data but rather the inherent deficit in suitable 3D GIS tools and datasets. An alternative exploratory analysis was carried out using 3D heat maps.

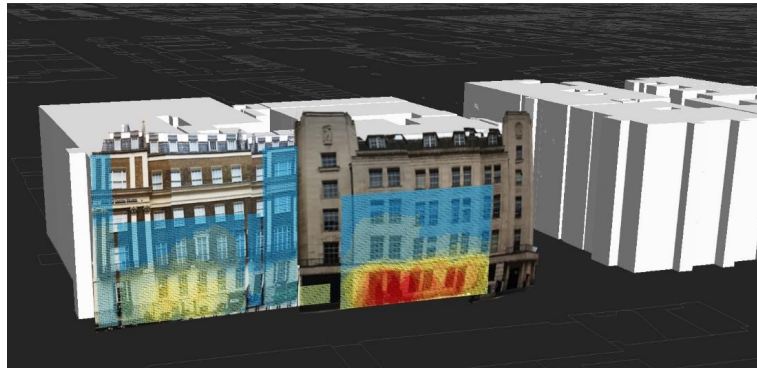


**Figure 5** Gaze Positions Point Density Function of All Participants mapped in 3D

Every gaze positions was passed through a point density function to create a heat map for each building façade using the Gram-Schmidt orthogonalization process [see Wong (2012) for more detail] (Figure 5). A similar clustering approach was adopted in a web-based experiment identifying the features in a number of urban scenes which could be used in forming navigational instructions (Bartie et al., 2014). The 3D heat maps were overlaid against buildings in Google Earth as well as over textured buildings (Figures 6&7). The results showed that ground floor front-facing building facades were most examined followed by road signs. Buildings at decision points such as junctions were also used as landmarks.



**Figure 6** Gaze Position Heat Map for Endsleigh Gardens in Google Earth



**Figure 7** Gaze Position Heat Map for Endsleigh Gardens with building textures

The study and its subsequent analysis clearly demonstrates a severe shortcoming within the current state of GIS – where 2D is insufficient in analysing the data collected yet, true 3D data of adequate quality geometrically and semantically is unavailable. Further, there is a lack of GI systems able to perform geospatial queries in 3D which are readily available in 2D such as buffering, intersection and topological relations.

#### 4. Conclusion

This study has provided a novel methodology in gathering requirements for a 3D navigation dataset. It outlines the first steps in an iterative study whereby the above recommendations would be implemented in developing a true 3D GIS dataset for navigation. Further work with participants from different age ranges and cultural backgrounds would be desirable to capture a representative sample of navigation strategies. In addition, true eye tracking and the availability of a 3D city model with full geometric and semantic attributes would enable the realisation of the full potential of the experiment. While the study demonstrates the possible value of 3D, it also shows the inherent deficiencies in the wider 3D field. In all, 3D is a multifaceted and ill-defined problem and it is unclear whether the benefits of the extra dimension outweighs its complexity. This study shows 3D is beneficial in the application of pedestrian navigation but argues existing technologies are incapable of delivering the envisaged true 3D navigation system. Where 2D maps works well on existing smartphone technologies, 3D navigation may require other enabling technologies such as a form of heads-up display or an ambient device. Regardless the direction 3D takes, what is key is that a user-centric design approach will ensure resulting outcome is effective, efficient, and enjoyable to use.

## 5. Acknowledgements

This project was funded and supported by the Ordnance Survey.

## 6. Biography

Kelvin Wong is an EngD research engineer at the UCL Centre for Virtual Environments, Interaction and Visualisation, Department of Computer Science, University College London. His research interests focuses on the challenges of deploying 3D geographic datasets at a national level with particular interests in usability, applications and data quality of 3D geographic information. Additional research relates to 3D visualisations and 3D requirements gathering.

Claire Ellul is a Lecturer in Geographical Information Science at University College London. Prior to starting her PhD, she spent 10 years as a GIS consultant in the UK and overseas, and now carried out research into the usability of 3D GIS and 3D GIS/BIM integration. She is the founder and current chair of the Association of Geographical Information's 3D Specialist Interest Group.

## References

- BARTIE, P., MACKANESS, W., PETRENZ, P. & DICKINSON, A. Clustering landmark image annotations based on tag location and content. Proceedings of RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production, 31st August 2014, 2014.
- MAY, A. J., ROSS, T., BAYER, S. H. & TARKIAINEN, M. J. 2003. Pedestrian Navigation Aids: Information Requirements and Design Implications. *Personal Ubiquitous Comput.*, 7, 331-338.
- MILLER, J. & CARLSON, L. 2011. Selecting landmarks in novel environments. *Psychonomic Bulletin & Review*, 18, 184-191.
- PARTALA, T., NURMINEN, A., VAINIO, T., LAAKSONEN, J., LAINE, M. & VÄÄNÄNEN, JUKKA. Salience of visual cues in 3D city maps. Proceedings of the 24th BCS Interaction Specialist Group Conference, 2010. 428-432.
- WONG, K. 2012. *3D Geographic Information and Solar Panel Positioning*. University College London.
- ZANDBERGEN, P. A. 2009. Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS*, 13, 5-25.

# The role of the SQL date functions in a GIS for ecological conservation

Greg Wood<sup>\*1</sup>, Duncan Whyatt<sup>†1</sup> and Carly Stevens<sup>‡1</sup>

<sup>1</sup> Lancaster Environment Centre, Lancaster University, Lancaster, UK, LA1 4YQ  
October 7<sup>th</sup>, 2014

## Summary

This paper demonstrates the use of temporally dynamic GIS attributes in devising and communicating ecological conservation constraints to construction projects. We embed SQL date functions into table view queries, using an open source database management system, to automatically calculate the ages of ecological observations in real time. In turn, the age attributes are then used in ecological analysis and monitoring applications, such as prioritising ecological survey efforts and estimating resilience to disturbance events. We conclude that the incorporation of temporally dynamic attributes in ecological datasets can help to reduce unnecessary development constraints, delays and expense.

**KEYWORDS:** temporal GIS; conservation; wildlife management; development; dynamic attributes

## 1. Introduction

European and UK laws expressly prohibit the anthropogenic disturbance of various animal species including bats, badgers, otters, great crested newts and a selection of birds (see Chartered Institute of Ecology and Environmental Management, 2014). This has a direct impact upon the construction industry, as measures must be taken to ensure that new development activities do not cause significant noise, vibrations or human presence in close proximity to the places that these animals occupy as shelters. New developments therefore often necessitate substantial conservation effort, involving the mapping of all protected animals inhabiting the site, continuous monitoring of how their distribution changes over time and the calculation of the spatial constraints these distributions place on construction activities (English Nature, 2002, Natural England, 2013, Scottish Natural Heritage, 2013a, b). As an example of the latter, works are not permitted within 30m of an otter holt (Chartered Institute of Ecology and Environmental Management, 2014), meaning that the locations of all otter holts must be found, mapped, and buffered accordingly.

In themselves, ecological practices of this manner can prove to be formidable tasks. The challenge however is intensified when developments have a multi-year duration, forcing ecologists to account not only for spatial constraints, but also their change through time due to temporal dynamics in species populations, migrations and habitat usage. Whilst guidance offered by ecological conservation bodies, such as Natural England and Scottish Natural Heritage, and the British Biodiversity Standard (British Standards Institute, 2013) detail the survey methodologies and buffer distances required for such a conservation style, they fail to address many of these associated temporal complexities. For example, Wood et al. (submitted-a) found that no guidance is given regarding the length of time an otter's resting place must remain unused before its protection can be dropped. If protection is enforced unnecessarily, then the financial cost associated with the redesign of the site and/or ecological mitigation (such as creating an artificial resting place) is equally unnecessary. Similarly, dropping the protection prematurely would facilitate the removal of a resting place which may have

---

<sup>\*</sup> g.wood1@lancaster.ac.uk

<sup>†</sup> d.whyatt@lancaster.ac.uk

<sup>‡</sup> Stevens.c@lancaster.ac.uk

been used in the future.

Although GIS solutions have long proved to be a popular choice for ecologists to capture, record and present ecological data relating to development works, and to calculate associated buffer distances, ‘out of the box’ GIS software fails to adequately represent these temporally dynamic data. This difficulty is exacerbated by the fact that the skill set of many practicing ecologists and decision makers (in both the spheres of regulation and development planning) encompass only basic GIS techniques (Wood et al., submitted-b).

This paper presents initial research into a GIS architecture which enables the automatic calculation of temporally dynamic attributes, and illustrates its application to protected species conservation. We aim to address a number of issues identified during previous work (Wood et al., submitted-b). First, we discuss a system architecture to facilitate a centrally stored pool of ecological data, which can be accessed by a variety of users, each using different software and possessing different GIS capabilities. Second, we outline a methodology to automate the processing of spatio-temporal dynamics to increase usability among novice users. Third, simple examples are illustrated using a case study site located in Scotland. Finally, explore the potential for more complex spatio-temporal modelling such as accounting for seasonal variations in resilience to anthropogenic disturbance such as that documented by Ruddock and Whitfield (2007).

## 2. Methodology

PostgreSQL with the PostGIS spatial extension served as the database management system (DBMS) platform for the solution. Tables were created for each set of protected features exhibited on the development site, including badger sett entrances, otter holts and rests, bat roosts and nests for various species of bird. Geoserver was then utilised to generate web feature services (WFS) and GeoJSON data structures to feed into various client GIS interfaces including ArcGIS, QGIS, uDig and will later feed an android application currently in development (Figure 1). By relying on interfaces that were familiar to the user, the learning curve in data exploration and utilisation was significantly reduced when compared to implementing a custom browser based interface for example. Feeding ecological data to clients in this manner also meant that it could be used to spatially query other spatial data held on the client, a feature deemed important by potential stakeholders as part of earlier research (Wood et al., submitted).

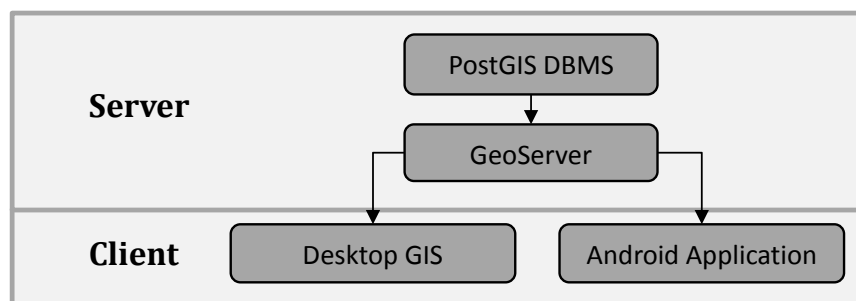


Figure 1 System architecture

Each record in a PostGIS table represents an individual protected ecological feature, each with three date fields: start\_date, last\_observed\_date and end\_date. An additional activity attribute was added allowing ecologists to fill in ‘High’ ‘Medium’ ‘Low’ and ‘Zero’. When a field ecologist observes the feature, as part of ongoing monitoring efforts, they take one of two courses of action. If no changes in activity are noticed, the ecologist simply updates the last\_observed\_date attribute. If a change is noticed, then the current date is entered into the end\_date attribute and a new record is made for the feature containing the current date in start\_date and last observed date attributes (see Tables 1a, b and c). Views were then created in PostGIS to dynamically populate time dependent attributes each time an update is requested from the client.

<b>Name</b>	<b>start_date</b>	<b>last_observed_date</b>	<b>end_date</b>	<b>Activity</b>
Sett 1	01/01/2014	01/01/2014	NULL	High

Table 1a **Example record of a badger sett observation**

<b>Name</b>	<b>start_date</b>	<b>last_observed_date</b>	<b>end_date</b>	<b>Activity</b>
Sett 1	01/01/2014	01/02/2014	NULL	High

Table 2b **Example update to Table 1a where activity remains the same**

<b>Name</b>	<b>start_date</b>	<b>last_observed_date</b>	<b>end_date</b>	<b>Activity</b>
Sett 1	01/01/2014	01/01/2014	01/02/2014	High
Sett 1	01/02/2014	01/02/2014	NULL	Low

Table 3c **Example update to Table 1a where activity has changed**

### 3. Implementations and discussion

In the simplest form of date dependent attributes illustrated here, an age attribute was created using the function `age(last_observed_date)`. This gives the number of days since the ecologist working on the case study site last observed the feature, which proved extremely useful in planning daily work schedules. This attribute was also of interest to the environmental regulator, who ensures that all development works comply with national and international laws. The addition of an age attribute allowed the regulators to see that ecological monitoring activities, prescribed as part of the planning conditions, were being adhered to.

Since `start_date` is never altered, running the function `age(start_date)` on the most recent instance of the feature, yielded the length of time activity remained constant for. For otter holts for example, where ‘zero’ activity was consecutively recorded for over one year, the status of the feature could be dropped from ‘active’ to ‘inactive’, allowing decision makers to apply less weight to these records in further spatial analysis. After two years of no activity the feature status could be dropped further to ‘historic’ where protection zones derived under CIEEM (2014), Natural England (2013) and Scottish Natural Heritage (2013a, b) guidelines, were removed altogether (Figure 2). The timespans were agreed through face to face discussion between ecological experts and regulators undertaken throughout the summer of 2014. When compared to more traditional approaches to ecological constraints mapping (see British Standards Institute, 2013), which typically use a single temporal snapshot to represent data for the entire time span of the project, this new methodology facilitated notable improvement. By removing unnecessary ecological constraints, developers are saved both time and money that would have either been spent on lengthy licence applications to disturb the feature or to undertake alternative work procedures. Equally, if the methodology becomes more widely adopted, the number of unnecessary licence applications submitted to regulatory authorities would be cut and would free up their time to pay closer attention to those applications that warrant it.





**Figure 2** Demonstrating the use of temporally dynamic attributes to categorise the otter holts. The top map shows results from the initial otter survey in 2011 and the bottom shows the results of the spatiotemporal modelling in 2014 outlined in the main text

#### 4. Conclusion and future research

This paper presented two simple examples where use of the `age()` function (part of many DBMS' including Postgres) coupled with a table view, have added significantly richer information to ecological GIS data. Though the `age()` function is relatively easy to use in this manner, its use in creating dynamic attributes is seldom implemented.

The next step in our research is to model spatio-temporal variation in disturbance susceptibility. This will consist of a lookup table with different buffer distances for different times of year and for different ages of the feature. Different types of badger sett experience different usage frequencies on a seasonal basis for example (Roper et al., 2001), and could be reflected with larger buffer distances in periods of heightened activity. Additionally, Ferguson-Lees et al. (2011) note that different bird species vary in time taken to build nests, egg incubation periods and juvenile rearing, each of which can be associated with different susceptibilities to disturbance. An android application will then be developed to use these spatio-temporally variable buffers as geofences (see Namiot and Sneps-

Sneppe, 2013) to alert site engineers if they accidentally move into a protective buffer zone. Ultimately, it is hoped that the examples of temporally dynamic attributes presented here, will serve in creating more robust ecological constraints mapping. Additionally, it is hoped that our use of the `age()` function and table views, will spark ideas for its application in other areas of research.

## 5. Acknowledgements

The authors would like to thank the Centre for Global Eco-Innovation, partly funded by the European Regional Development fund, for their financial support during this study. Thanks is also given to Richard Castell and David Hackett of Solum Environmental Ltd., for their helpful discussions during the development of this research.

## 6. Biographies

*Greg Wood* is a third year PhD student at Lancaster University with research experience in modelling spatio-temporal interactions of ecology and habitat connectivity in academia and for governmental environment regulators.

*Duncan Whyatt* is a Senior Lecturer in GIS at Lancaster University with research interests in air pollution at a variety of spatial and temporal scales.

*Carly Stevens* is a Lecturer in ecology at Lancaster University whose main research interests are related to the impacts of global change on plants and soils.

## 7. References

- BRITISH STANDARDS INSTITUTE 2013. Bs 42020:2013: biodiversity – Code of Practice for Planning and Development. London: British Standards Institute.
- CHARTERED INSTITUTE OF ECOLOGY AND ENVIRONMENTAL MANAGEMENT. 2014. *Technical Guidance Series* [Online]. Available: <http://www.cieem.net/technical-guidance-series-tgs-> [Accessed 16th June].
- FERGUSON-LEES, J., CASTELL, R. & LEECH, D. 2011. *A Field Guide to Monitoring Nests*, UK: British Trust for Ornithology.
- NAMIOT, D. & SNEPS-SNEPPE, M. 2013. Geofence and Network Proximity. In: BALANDIN, S., ANDREEV, S. & KOUCHERYAVY, Y. (eds.) *Internet of Things, Smart Spaces, and Next Generation Networking*. Springer Berlin Heidelberg.
- NATURAL ENGLAND. 2013. *European Protected Species: Mitigation Licensing - How to Get a Licence* [Online]. Available: [http://www.naturalengland.org.uk/Images/wml-g12\\_tcm6-4116.pdf](http://www.naturalengland.org.uk/Images/wml-g12_tcm6-4116.pdf) [Accessed 22nd May].
- ROPER, T. J., OSTLER, J. R., SCHMID, T. K. & CHRISTIAN, S. F. 2001. Sett Use in European Badgers *Meles Meles*. *Behaviour*, 138(2), 173-187.
- RUDDOCK, M. & WHITFIELD, D. P. 2007. A Review of Disturbance Distances in Selected Bird Species. *A report from Natural Research (Projects) Ltd to Scottish Natural Heritage*.
- SCOTTISH NATURAL HERITAGE. 2013a. *Badger Licences - Development* [Online]. Available: <http://www.snh.gov.uk/protecting-scotlands-nature/species-licensing/mammal-licensing/badgers-and-licensing/dev/> [Accessed 12th October].
- SCOTTISH NATURAL HERITAGE. 2013b. *Otters and Development* [Online]. Available: <http://www.snh.org.uk/publications/on-line/wildlife/otters/effects.asp> [Accessed 18th March].
- WOOD, G., WHYATT, D., HACKETT, D. & STEVENS, C. submitted-a. Spatio-Temporal Challenges in Representing Wildlife Disturbance within a GIS.
- WOOD, G., WHYATT, D. & STEVENS, C. submitted-b. Towards Integrating Planning for the Built Environment and Biodiversity, through Collaboration.



# Understanding car ownership elasticities in England and Wales: Advancing the evidence base with new data sources

Godwin Yeboah<sup>\*1</sup>, Jillian Anable<sup>+1</sup>, Tim Chatterton<sup>λ2</sup>, Jo Barnes<sup>γ2</sup>, R. Eddie Wilson<sup>Ω3</sup>, Oliver Turnbull<sup>θ3</sup>, Sally Cairns<sup>φ4</sup>

<sup>1</sup>The Centre for Transport Research, School of Geosciences, University of Aberdeen, Scotland.

<sup>2</sup>Air Quality Management Resource Centre, University of the West of England, Bristol, UK.

<sup>3</sup>Intelligent Transport Systems, University of Bristol, Bristol, UK.

<sup>4</sup>Transport Research Laboratory and University College London, London, UK.

March 12, 2015

## Summary

This study presents global and local models explaining household car ownership elasticity in England and Wales based on new datasets from Experian household *median income* and 2011 Census released by UK Government agencies. Latest empirical evidence on car ownership elasticity across the area is based on 2001 Ward level household *average income* estimates and 2001 Census. In using different income estimates and new datasets, new evidence is compared with what we already know about car ownership elasticity. Geographically weighted regression is utilized to estimate and forecast car ownership elasticities at both Ward and Lower layer Super Output Areas. With our initial modelling in this paper, we suggest that future work should incorporate road worthiness tests data, at lower geographies when released, from Ministry of Transport (MOT) as a proxy for car ownership (and use) to undertake a comparative analysis towards deepening our understanding of car ownership (and use) trends to inform transport policy in England and Wales.

**KEYWORDS:** car ownership, median income, geographically weighted regression, elasticity, MOT test

## 1. Introduction

Existing 21<sup>st</sup> Century approaches used to estimate car ownership models may be grouped into two categories: a-spatial and spatial approaches. Most of the studies found fall in the a-spatial category (Clark, 2009; Dargay and Hanly, 2007; Dargay, 2002, 2001; Leibling, 2008; Litman, 2013; Whelan, 2007); with limited studies using a spatial approach (Clark and Finley, 2010; Clark, 2007). Figure 1 indicates a range of a-spatial and spatial approaches employed in earlier studies whilst Figure 2 shows an overview of the variety of datasets used since 2000 to study car ownership in the UK.

---

\* Godwin.Yeboah@abdn.ac.uk

+ J.Anable@abdn.ac.uk

λ Tim.Chatterton@uwe.ac.uk

γ Jo.Barnes@uwe.ac.uk

Ω Re.Wilson@bristol.ac.uk

θ Oliver.Turnbull@bristol.ac.uk

φ SCairns@trl.co.uk

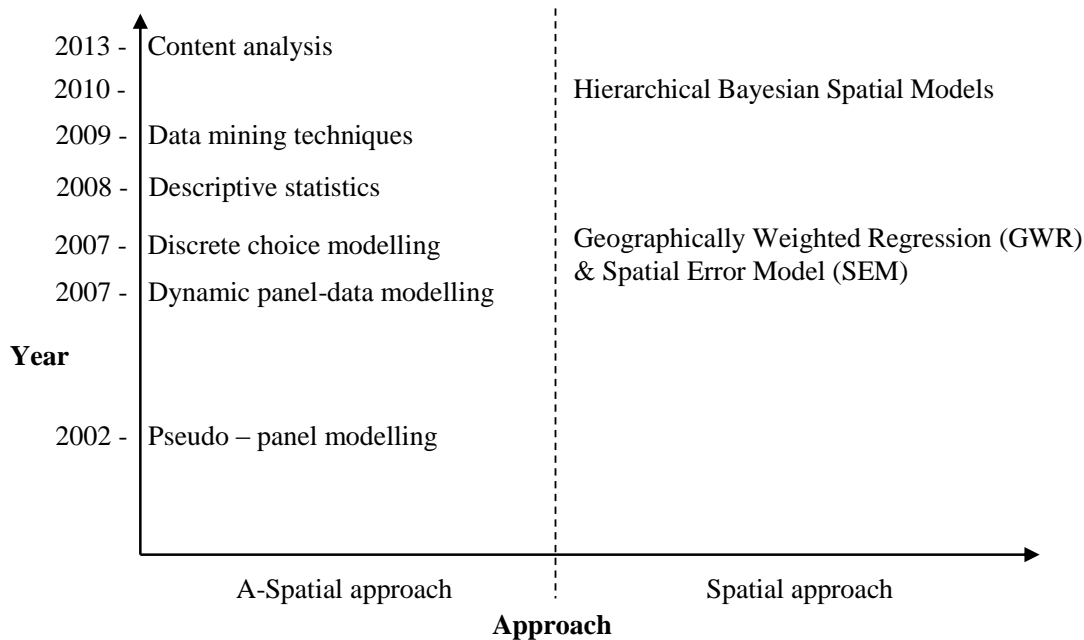


Figure 1: Overview of car ownership study approaches in the 21st Century

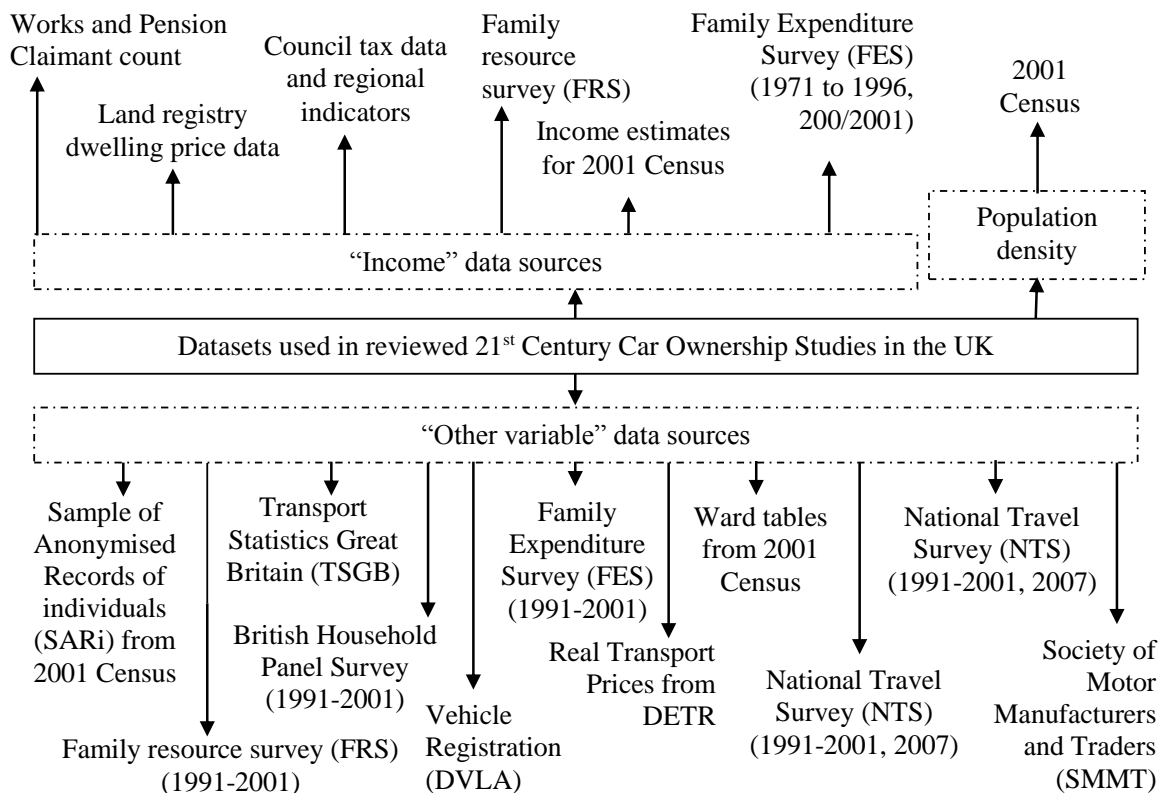


Figure 2: Sketch of datasets used in reviewed 21st Century Car Ownership Studies in the UK

ONS means Office for National Statistics; BHPS means British Household Panel Survey; DVLA means Driver and Vehicle Licensing Agency; DETR means Department for Environment Transport and the Regions.

The overarching research questions in this study comprise: 1) *What is the relationship between car ownership and household median income at both LSOA and Ward geographies?*; 2) *How does the strength of the relationship influence future elasticity scenarios of the number of cars in England and Wales?*; and, 3) *Does the modifiable area unit problem (MAUP) have an effect on the outcome of derived elasticity scenarios?* Elasticity is the measurement of how responsive an economic variable is to a change in another and could be considered a tool for measuring the responsiveness of one variable to changes in another. Elasticity has the advantage of being a unitless ratio, independent of the type of quantities being varied. In addressing these questions, a robust spatial analytical technique, geographically weighted regression (GWR), is used to estimate individual strength of the relationship for each LSOA and Ward along with future forecasts of the number of the size of vehicles in the area. The use of two geographies (i.e. LSOA and Ward) provides the empirical bases in addressing whether modified area unit problem (MAUP) has effect on the outcome of the future forecasts in this study; using at least two geographies is suggested (Flowerdew et al., 2008, p. 1254). The paper is divided into four main sections with this section inclusive: introduction, study area and data specification, analysis and results, and discussion and conclusion.

## **2. Study area and data specification**

This study examines car ownership elasticities in England and Wales by using available datasets described in this section. We found available M.O.T tests datasets at the time of analysis too coarse (postcode district level) to use but we intend to include dataset (when released to the authors) in further analysis. Conducted literature review in this paper justifies the use of three key variables in our analysis; car ownership, income and population density.

### **2.1. Car ownership dataset**

The car ownership data is taken from the 2011 Census (KS404EW) in the UK. The UK Census is taken every decade and the most recent is the 2011 Census. Car ownership at the household level is the number of cars or vans that are owned, or available for use, by one or more members of a household. This includes company cars and vans that are available for private use. It does not include motorbikes or scooters, or any cars or vans belonging to visitors. The count of cars or vans in an area relates only to households. Cars or vans used by residents of communal establishments are not counted. The Census gives an excellent snapshot of the country at a time. The 2001 Census version was only used to test results from previous study.

### **2.2. Household Median Income dataset**

The availability of Household median income estimate, from Experian Limited, provided at LSOA geographies make analysis at LSOA level possible (Experian, 2011). Additionally, we argue that since one of the recent key findings from ONS suggest that growth in UK median household disposable income mirror closely growth in GDP from 1997 to 2012, using it to understand car ownership trends is more appropriate than household average income estimates (ONS, 2013). Further evidence is provided by Tim Harford, a UK Economist, that the median earnings in the UK, unlike US, have increased by 1.25% annually since 1968 (Harford, 2011). Average income estimates might not always reflect true income variations in economies where income disparities are wider. In the UK, although income inequality fell within 2011-2012 at its lowest since 25 years (Stewart and Osborne, 2013), the trend of income inequality is at alarming rate (EqualityTrust, 2013); and growing faster than any other rich country according to OECD with the top 10% having incomes 12 times greater than the bottom 10% (Ramesh, 2011). Given that income inequality is prevalent in the UK, we argue that median income estimates suggest a better reflection of “wealth” than average income estimates. Moreover, there seem to be no other recent relevant income data to use as the 2011 Census did not collect income information. The 2011 and 2004 median income estimates were used.

### **2.3. Geographic dataset and other explanatory variables**

Two main geographies used are 2001 and 2011 Lower layer Super Output Areas (LSOAs) together with 2003 and 2011 Wards. Population density information from 2011 and 2001 Censuses were used in this study. Household structure and size could as well be used but this is not included due to possible co-linearity since it depends on number of people and number of household which are present in already chosen variables (Clark, 2007); see Figure 2 for other data sources.

## **3. Analysis and results**

### **3.1. Mapping significant clusters of car ownership, income and population density**

As a first step, we explored areas in England and Wales where some Wards and LSOA with high car ownership can be determined to be statistically significantly ( $p < 0.01$ ) different to the national household car ownership average. Applying Repley's K algorithm (a multi-distance spatial cluster analysis) gave a fixed distance of 17.5km exhibiting maximum spatial clustering effect which is used for the mapping in Figures 3 and 4. As shown in Figure 3 and Figure 4, there are differences in significant clusters at Ward and LSOA levels of analyses suggesting that further analysis at LSOA could provide an enhanced evidence to what we know at Ward level.

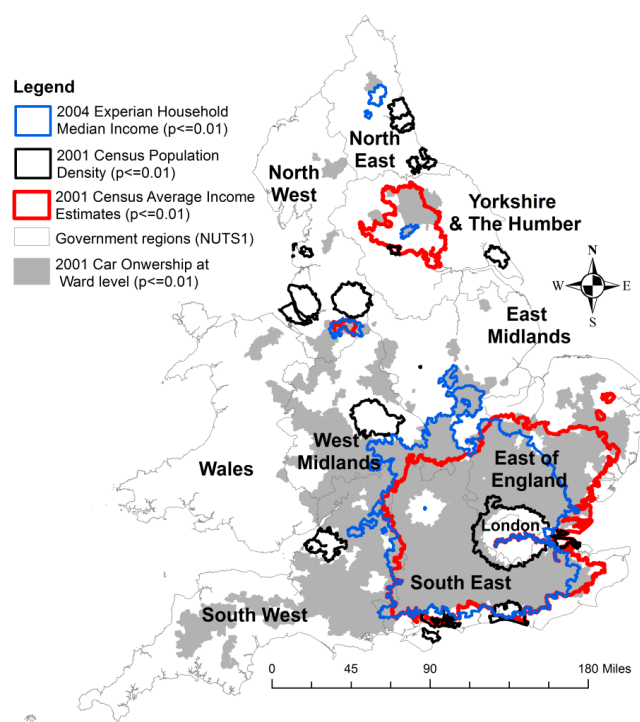


Figure 3: Statistically significant ( $p \leq 0.01$ ) clusters of car ownership, population density, average and median income in England and Wales (Ward)

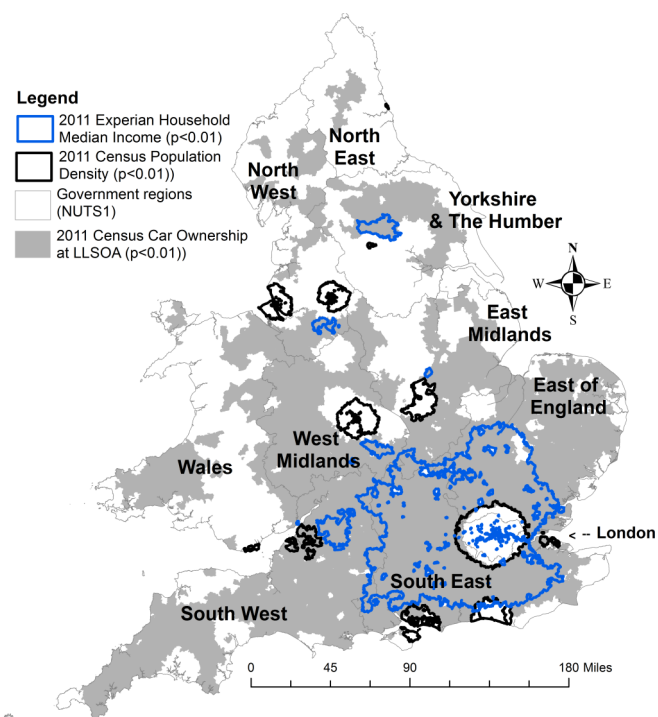


Figure 4: Statistically significant ( $p \leq 0.01$ ) clusters of car ownership, population density, and median income in England and Wales (LSOA)

### 3.2. Car Ownership Models estimation with income and population density

#### 3.2.1. Global models estimation at LSOA and Ward geographies

Logarithmic regression technique is used here to model car ownership with Experian's household median income and Census population density by comparing Clark's study which used 2001/2002 average income estimates data at Ward level along with other parameters. The comparison is done loosely with awareness that the income datasets are different in both studies but, the expectation is that these parameters should be reasonably close despite approximately two year's difference in income datasets and inherent estimation process. The average car ownership per household computed for the 8,805 Wards was still the same (1.223) as in Clark's study (which used 8,837 Wards). Results of global logarithmic regression model of car ownership at 2003 version of Ward geography (with 2004 Experian Median Income) show an adjusted R of .762:

$$\begin{aligned} \text{cars per household} = & \\ -5.051 + 0.638 \text{Ln}(2004 \text{ Median Income}) - 0.104 \text{Ln}(2001 \text{ Population Density}) & \quad \text{Eq.1} \\ (0.063) \quad (0.006) & \quad (0.001) \end{aligned}$$

The value of 0.52 (i.e. 0.638/1.223) shows that 10% rise in household median income gives rise to a 5.2% increase in car ownership at ward level across England and Wales. This is about 0.9% difference in prediction when the estimated household average income is used at ward level across England and Wales as shown in Clark's study. The average car ownership per household across England and Wales at LSOA is 1.19. Results of logarithm regression model of car ownership at LSOA geography (2011 Experian Income) show an adjusted R of .637:

$$\begin{aligned} \text{cars per household} = & \\ -3.585 + 0.503 \text{Ln}(2011 \text{ Median Income}) - 0.144 \text{Ln}(2011 \text{ Population Density}) & \quad \text{Eq.2} \\ (0.038) \quad (0.004) & \quad (0.001) \end{aligned}$$

The value of 0.42 (i.e. 0.503/1.19) shows that 10% rise in household median income gives rise to a 4.2% increase in car ownership at LSOA level across England and Wales. The coefficient of the median income parameter is reduced which in-turn affects the elasticity parameter for the prediction. Results of logarithm regression model of car ownership at 2011 Ward geography (2011 Household Median Income) show an adjusted R of .539:

$$\begin{aligned} \text{cars per household} = & \\ 1.156 + 0.050 \text{Ln}(2011 \text{ Median Income}) - 0.147 \text{Ln}(2011 \text{ Population Density}) & \quad \text{Eq.3} \\ (0.024) \quad (0.002) & \quad (0.002) \end{aligned}$$

The average car ownership per household at 2011 Ward geography is 1.3. The value of 0.04 (i.e. 0.05/1.3) suggests that 10% rise in household median income gives rise to a 0.4% increase in car ownership at LSOA level across England and Wales.

#### 3.2.2. Local model estimates at LSOA and Ward geographies

Here, we are interested in finding out if any systematic spatial pattern exists in error estimates in the global regression models (Equation 2 & 3); our solution for eliminating any systematic pattern is the utility of Geographically Weighted Regression (GWR) technique in modelling these variables. GWR generates spatially calibrated regression models by generating separate regression equation for features

in a sample data being investigated to explain the extent of spatial variation. Fotheringham et al. (2002) provide detail discussion of GWR. Even when spatial autocorrelation is identified, or not, in a global model, literature suggest that global measures tend to be misleading and that the use of localised version of spatial autocorrelation technique be used in the examining spatial arrangements of data (Fotheringham et al., 2002, pp. 14–15). Spatial autocorrelation still existed in the error estimates after the global logarithmic regression. Clark argued that goodness of fit for his case (i.e. at 2003 Ward level using 2001 Census data) improved from about 75.3% to 90.1% at 5% of neighbours (i.e. number of Wards) and it was due to considering the spatial component of the regression model. Thus, the use of spatially calibrated regression models provides better explanation than a-spatial regression regimes. The result from our calculation at 2011 LSOA suggests an improvement of goodness of fit from about 63.7% to 83.8% at 5% of neighbours. In using 2011 Ward level values, the goodness of fit in our case improved from about 53.9% to 74.9% at 5% of neighbours.

### **3.2.3. Estimating number of cars and vans under six elasticity scenarios**

We estimated the change in local and national car ownership of a 10% rise using household median income, and population density as co-variable, to show the effect of using different elasticity estimates. Table 1 shows results from six scenarios of predicted size of the national and London cars and fleets. Our results from the computed mean of car ownership elasticity ( $\eta_{co} = 0.42$ ), from equation 2, which is a constant global elasticity assuming car ownership behaviour across England and Wales, suggest that relative increases of cars and van fleets, using 2011 LSOA level median income data, reflect similar differences (i.e. about 1.9% = 5.9-4.0) when 2001 Ward level specific elasticity ( $\eta_{co} = 0.744/y_i$ ) is used as reported (i.e. about 1.9% = 8.6-6.7) by Clark (2007, p. 195). However, our computation using 2011 Ward specific elasticities from ( $\eta_{co} = 0.05/y_i$ ), gave relatively bigger difference (i.e. about 3.4% = 4.1- 0.7) for England and Wales and London. This might mean that average income affects predictions in London more than in England and Wales as Clark (2007, p. 194) argues that “estimated national [i.e. England and Wales] parameter value of 0.744 is too high for the London situation”.

Scenario	England and Wales	London (excluding the City of London)
Number of cars and vans	27,294,656	2,662,722
Car and van fleet size after a 10% increase in median income		
Scenario 1: Using mean car ownership elasticity ( $\eta_{co} = 0.423$ ) with respect to 2011 LSOA median income for prediction	28,386,442 (+4.0)	2,819,823 (+5.9)
Scenario 2: Using mean car ownership elasticity ( $\eta_{co} = 0.039$ ) with respect to 2011 Ward median income for prediction	27,376,540 (+0.3%)	2,676,036 (+0.5%)
Scenario 3: The third scenario is to assume LSOA specific elasticities from ( $\eta_{co} = 0.503/y_i$ ), which takes account of local circumstances but still assumes a uniformly estimated value for the strength of the relationship between car ownership and median income	28,604,800 (+4.8%)	2,849,113 (+7.0%)
Scenario 4: The fourth scenario is to assume 2011 Ward specific elasticities from ( $\eta_{co} = 0.05/y_i$ ), which takes account of local circumstances but still assumes a uniformly estimated value for the strength of the relationship between car ownership and median income	28,413,737 (+4.1%)	2,681,361 (+0.7%)
Scenario 5: The fifth scenario is the most flexible at LSOA level, it uses the locally (LSOA) estimated GWR median income parameters when estimating the elasticity $\beta_{LSOAmedianincome(ui,vi)}/y_i$	29,068,809 (+6.5%)	2,795,858 (+5.0%)
Scenario 6: The sixth scenario is the most flexible at Ward level, it uses the locally (Ward) estimated GWR median income parameters when estimating the elasticity, $\beta_{Wardmedianincome(ui,vi)}/y_i$	27,594,897 (+1.1%)	2,713,314 (+1.9%)

Table 1: Estimated number of cars and vans under six elasticity scenarios (3 scenarios x 2 spatial resolutions)

#### 4. Discussion and conclusion

This study is in line with previous car ownership studies establishing income as a significant predictor, but different results were achieved using two levels of geographies (LSOA and Ward) when household median income was considered. GWR based approach was used to examine the relationship between household median income and car ownership along with population density at both Lower Layer Super Output and Ward geographies. We found similar values when Clark's results were re-computed for the three elasticity scenarios (only Ward levels) except small changes in the estimates for London scenarios;  $0.744/y_i$  and  $\beta_{income}/y_i$  were 9.4% and 7% respectively (Clark study estimated 8.6% and 6.5% respectively). Despite the increase in parameters in GWR compared to global regression models, GWR model allows for a true estimation of the local parameter spatially (Clark, 2007). Although Clark (2007) suggested that his methodological approach could be used to understand other variables in 2001 Census and their relationship to income and other explanatory variables to map the outcome across England and Wales, we used the approach along with new datasets (particularly 2011 Census and 2011 Experian median income) to explain how car ownership relates to household median income using 2011 population density as a covariate data. MAUP was found to have varying effect on the derived elasticity scenarios as shown when the scenarios are compared (i.e. 1 & 2, 3 & 4, and 5 & 6); the differences in the case of London are bigger compare to the rest of England and Wales. It is suggested that future work should incorporate road worthiness tests data, at LLSOA and Ward geographies, from MOT as a proxy for car ownership and use to deepen our understanding of car ownership (and use) trends to inform transport policy in England and Wales. This is because the current MOT data released only contain postcode area information for Vehicle Test Stations (VTS) which is considered too coarse



(Chatterton et al., 2014); and, therefore inappropriate for a comparative analysis in this paper. Having registered keepers information together with pass/fail MOT tests at LSOA geography, for example, will provide alternative to derived car ownership measure from the Census for understanding car ownership from a different perspective. It might also be worth estimating elasticities using National Trip End Model (NTEM) Zones while taking into consideration of one of the NTEM sub-model, the National Car Ownership Model (NATCOP), and associated scenarios (DfT, 2014).

## **5. Acknowledgement**

The authors acknowledge the support of an EPSRC Advanced Research Fellowship (grant ref. EP/E055567/1). Data used contains National Statistics data © Crown copyright and database right 2014; Contains Ordnance Survey data © Crown copyright and database right 2014; Contains National Statistics data © Crown copyright and database right 2014. Authors also acknowledge Experian data from UKDataService - SN: 5738 Experian Demographic Data, 2004-2005 and 2008-2011. Finally, we thank an anonymous reviewer and Paul Emmerson at Transport Research Laboratory.

## **6. Biography**

Dr. Godwin Yeboah is a Research Fellow in Transport and based at the Centre for Transport Research at the University of Aberdeen in Scotland. His broad research interests: Geomatics; Intelligent Mobility solutions and Transport Policy; Big Data and Visual Analytics to knowledge; Energy Demand; Modelling and Simulation; Social Media Research; Machine Learning techniques.

Prof. Jillian Anable is a Senior Lecturer and a Personal Chair in Transport and Energy Demand. She is based at the Centre for Transport Research at the University of Aberdeen. Research interests: Transport, energy demand, climate change, travel/behaviour change, consumer demand for low-carbon vehicles, public acceptability/political deliverability, statistical segmentation.

Dr. Tim Chatterton is a Senior Research Fellow at the Air Quality Management Resource Centre at the University of the West of England in England. His research interests: Energy, air quality, climate change, behaviour change.

Dr. Jo Barnes is a Research Fellow at the Air Quality Management Resource Centre at the University of the West of England in England. Her research interests: air pollution and GIS, air quality, climate change, behaviour change.

Prof. Eddie Wilson is the Chair in Intelligent Transport Systems at the University of Bristol in England.

Dr. Oliver Turnbull is a Teaching and Research Fellow based at the Department of Aerospace Engineering at the University of Bristol in England.

Dr. Sally Cairns is a Senior Research Fellow and works jointly at Transport Research Laboratory and University College London. Her primary research interests relate to transport policy, traffic reduction and travel behaviour change, with a focus on analysis of complex empirical evidence from real-world experience.

## **7. Reference**

Chatterton, T., Barnes, J., Wilson, E.R., Anable, J., Cairns, S., 2014. Variations in car type, size, usage and emissions across Great Britain and relationships with socio-demographic characteristics. In: Proceedings of the 46th University Transport Studies Group Annual Conference, Newcastle, January 2014. Newcastle.

Clark, S., Finley, A., 2010. Spatial Modelling of Car Ownership Data: A Case Study from the United Kingdom. *Appl. Spat. Anal. Policy* 3(1), 45–65.

- Clark, S.D., 2007. Estimating local car ownership models. *J. Transp. Geogr.* 15(3), 184–197.
- Clark, S.D., 2009. The determinants of car ownership in England and Wales from anonymous 2001 census data. *Transp. Res. Part C Emerg. Technol.* 17(5), 526–540.
- Dargay, J., Hanly, M., 2007. Volatility of car ownership, commuting mode and time in the UK. *Transp. Res. Part A Policy Pract.* 41(10), 934–948.
- Dargay, J.M., 2001. The effect of income on car ownership: evidence of asymmetry. *Transp. Res. Part A Policy Pract.* 35(9), 807–821.
- Dargay, J.M., 2002. Determinants of car ownership in rural and urban areas: a pseudo-panel analysis. *Transp. Res. Part E Logist. Transp. Rev.* 38(5), 351–366.
- DfT, 2014. SUPPLEMENTARY GUIDANCE NTEM Sub-Models. Accessed at [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/370919/TAG\\_Supplementary\\_-\\_NTEM\\_Sub-Models-January2014.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/370919/TAG_Supplementary_-_NTEM_Sub-Models-January2014.pdf)
- Experian, 2011. Experian Demographic Data, 2004-2005 and 2008-2011 [WWW Document]. Accessed at <http://discover.ukdataservice.ac.uk/catalogue/?sn=5738&type=Data catalogue>
- Flowerdew, R., Manley, D.J., Sabel, C.E., 2008. Neighbourhood effects on health: Does it matter where you draw the boundaries? *Soc. Sci. Med.* 66(6), 1241–1255.
- Fotheringham, A.S., Brunson, C., Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.
- Harford, T., 2011. Bye, bye easy money. *Financial Times Magazine*. Accessed at <http://www.ft.com/cms/s/2/f7bfa5be-34b4-11e0-9ebc-00144feabdc0.html#axzz2v1afCLJo>
- Leibling, D., 2008. Car ownership in Great Britain. The Royal Automobile Club (RAC) Foundation, London. Accessed at [http://www.racfoundation.org/assets/rac\\_foundation/content/downloadables/car\\_ownership\\_in\\_great\\_britain\\_-\\_leibling\\_-\\_171008\\_-\\_report.pdf](http://www.racfoundation.org/assets/rac_foundation/content/downloadables/car_ownership_in_great_britain_-_leibling_-_171008_-_report.pdf)
- Litman, A.T., 2013. *Understanding Transport Demands and Elasticities - How Prices and Other Factors Affect Travel Behavior*. Victorian Transport Policy Institute. Accessed at <http://www.vtpi.org/elasticities.pdf#page=1&zoom=auto,0,716>
- ONS, 2013. Middle Income Households, 1977-2011/12 (The original date of publication was 2 December 2013 and was corrected on 17 February 2014). Accessed at [http://www.ons.gov.uk/ons/dcp171776\\_341133.pdf](http://www.ons.gov.uk/ons/dcp171776_341133.pdf)
- Ramesh, R., 2011. Income inequality growing faster in UK than any other rich country, says OECD. *Guard.* Accessed at <http://www.theguardian.com/society/2011/dec/05/income-inequality-growing-faster-uk>
- Stewart, H., Osborne, H., 2013. UK income gap shrinks to narrowest margin for 25 years. *Guardian*. Accessed at <http://www.theguardian.com/uk-news/2013/jul/10/income-gap-narrowest-margin-25-years>
- Whelan, G., 2007. Modelling car ownership in Great Britain. *Transp. Res. Part A Policy Pract.* 41(3), 205–219.

# Modelling the long-term economic and demographic impacts of major infrastructure provision: a simultaneous model approach

Chengchao Zuo<sup>\*1</sup>, Mark Birkin<sup>†1</sup>

<sup>1</sup>School of Geography, University of Leeds

Apr. 01, 2015

## Summary

This paper reports investigations into the feedback and linkages between demographic change and infrastructure provision. In this paper, we seek to explore the coupled dynamic of demographics, the economy and infrastructure simultaneously as a series of subsystems. The modelling results will explore different policy scenarios for regional infrastructure investment to offer an initial proof of concept of the feasibility of implementing a coupled model of demographic and economic growth over a medium to long time horizon, and promises a distinct and exciting perspective on the co-dynamic interplay of social and economic policies, regional development, infrastructure provision and prosperity.

**KEYWORDS:** infrastructure, demographic change, economic development, scenario modelling, policy analysis.

## 1. Introduction

This paper reports investigations into the feedback and linkages between demographic change, economic development and infrastructure provision, which are being undertaken by the Infrastructure Transitions Research Consortium (ITRC). National infrastructure systems (NIS) provide a foundation for economic productivity and human wellbeing. They shape many of the interactions between human civilisation and the natural environment [1]. However, the NIS for Great Britain faces considerable challenges in the future to serve a globalised economy and to meet the government's commitment on reduction in greenhouse gas emission [2]. Infrastructure UK (IUK), with support from organisations such as the ITRC, are amongst many groups on the international stage who are tasked with addressing such problems.

In the ITRC programme to date the reverse coupling of demographics to infrastructure has been articulated less explicitly. Interregional migration flows are typically viewed as business as usual, in common with core national projections. However the dependence of future demographic change on infrastructure is obvious – thus a new high-speed link between London and Birmingham would change relative accessibilities, which are the key driver of migration and commuting flows these regions. Infrastructure can also influence population change indirectly through economic growth – for example, the construction of a new desalination plant in East Anglia would create new jobs, tending to encourage the inflow of migrant workers. In short, “population growth leads to increased demand for infrastructure services, but better infrastructure services also attract population to a

---

<sup>\*</sup> geocz@leeds.ac.uk

<sup>†</sup> m.h.birkin@leeds.ac.uk

region” (Beaven, et al, 2014).

The population of the UK is currently growing rapidly under the influence of both international migration and natural change. This growth has been spatially uneven, which has important implications for infrastructure provision. ITRC has therefore laid down a series of spatially explicit demographic scenarios as a driver of future infrastructure requirements (Zuo et al, 2014).

Since 2010, the UK government starts to publish its National Infrastructure Plan (NIP) in an annually basis in order to setting out *a broad vision of the infrastructure investment required to underpin the UK’s growth*. In the NIP2014, an ambitious infrastructure plan was set out for the till 2020 and beyond by underpinning a pipeline of over £460 billion of planned public and private investment. The diagram below (Figure 1) shows the investment plan specified by the NIP2014. According to the plan, majority of the investment is going to transport and energy sector and mainly located in London and south of England.

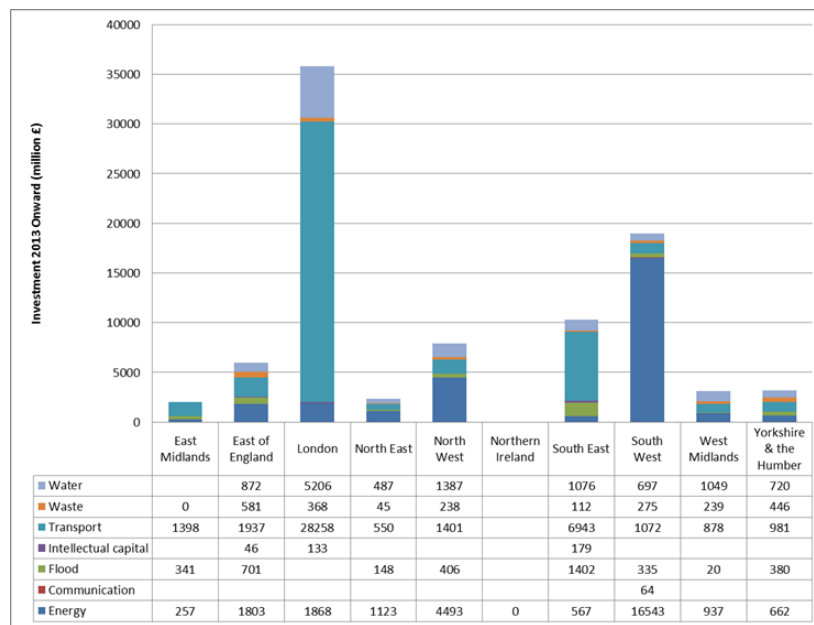


Figure 1. The 2013 onward infrastructure investment

This paper therefore seeks assess the socio-economic impact of these infrastructure developments by deploying a dynamic co-dependence of demographics, the economy and infrastructure, as series of coupled subsystems.

Models of this type have been suggested in the past for abstract multi-agent systems, and co-evolutionary models have been explored to some extent in the context of both ecological systems and economic markets. None of these models includes either infrastructure or a spatially explicit representation of a real demographic system. Of course linkages between population (or at least ‘demand’) and economic sectors, including infrastructure, are a feature of well-established input-output models, but although substantial work has been invested in the regional disaggregation of such models these approaches in turn lack an co-dependence perspective. The approach to be adopted here is therefore unique in exploring the co-dependence of infrastructure economy and demographics within a spatially explicit modelling framework.

## 2. Modelling Framework

The structure of the model in its current form is illustrated in Figure 1. The link from population to the economy is indicated through the flow of labour as a factor of production, while the reverse link is effected through a combination of spatial processes which underpin population movements. The role of infrastructure is articulated as of particular importance in view of the substantive focus of this work.

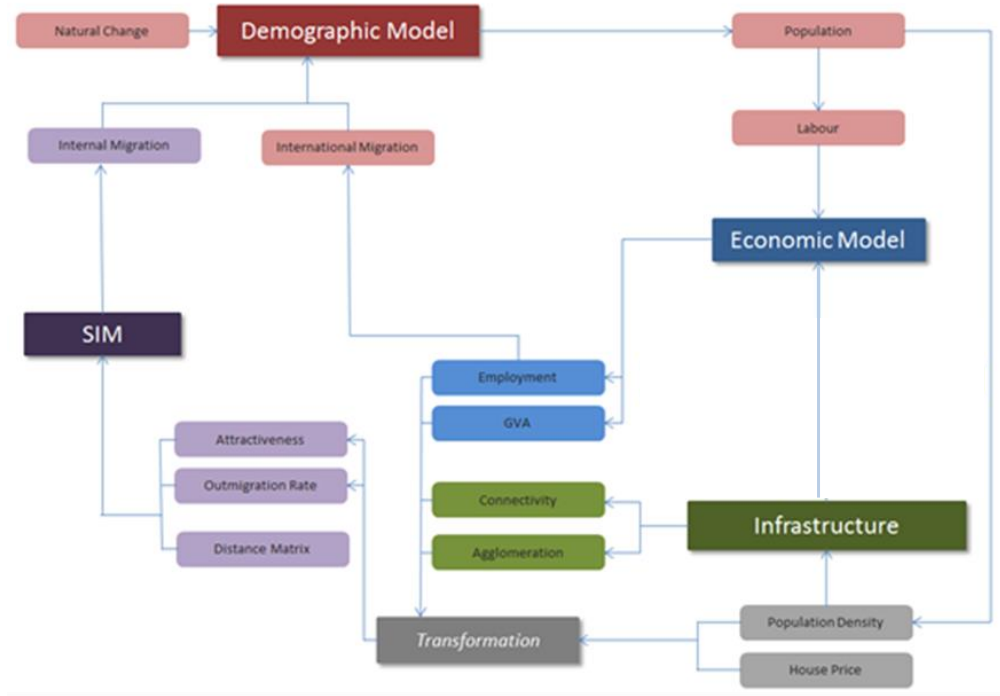


Figure 2 Modelling Framework of Simultaneous model

A standard population projection model consists of three components, which are fertility, mortality and migration, denoted as (1):

$$Pop_Y = Pop_{Y-1} + F_Y - M_Y + NM_Y \quad (1)$$

Where  $Pop_Y$  represent the population at year, F and M represent fertility (number of births) and mortality (number of deaths) separately, NM represents net migration which is the difference between immigration and outmigration. In Work Stream 1 (WS1) of the ITRC project, a series of population projection model were established to provide basic population estimations against different scenarios up to 2100. In WS1, a series of linear models were built to represent the net migration of each Local Authority District (LAD). These linear models have two main drawbacks: 1) lack of representing of detailed migration flow pattern; and 2) lack of flexibility in response to any local changes (i.e. infrastructure development in some areas). Therefore, a more sophisticated model is proposed is introduced by express migration in a more detailed way.

$$Pop_Y = Pop_{Y-1} + NC_Y + {}^oNM_Y + {}^iIM_Y - {}^oOM_Y \quad (2)$$

Where:  $Pop_{LAD}^Y$  represents the population for a LAD at year Y. NC represent natural change which the difference in number between live births and deaths. Migration terms are expressed by four variables here which are net migration from overseas ( ${}^oNM$ ), immigration from within UK ( ${}^iIM$ ),

and outmigration to somewhere else within UK (  ${}^IOM$  ).

For a given level of outmigration (OM), a production-constrained spatial interaction model can be introduced to generate inter-regional migration flows:

$$T_{ij} = A_i * {}^IOM_i * Att_j * d_{ij}^{-\beta}, \quad A_i = \frac{1}{\sum_j Att_j * f(d_{ij})} \quad (3)$$

where  $T_{ij}$  represents the migration flow from region (i) to region (j),  $Att_j$  is attractiveness of region (j) and  $d_{ij}$  is a measure of the distance or trip cost between region pairs.  $A_i$  is a balancing factor which ensures that the flows from each region are constrained to the overall level of outmigration.

$d_{ij}^{-\beta}$  is a distance friction function. So the internal immigration for each area (LAD) in this case can be calculated by sum up the immigration from  $T_{ij}$

$${}^IIM_j = \sum_i T_{ij} \quad (4)$$

There are two sets of parameters need to be calibrated, where  $\beta$  is a parameter related to the efficiency of the transport system; and attractiveness ( $Att_j$ ) is a synthetic variable, which indicates the potential to attract migration into a region (j). In practise, a goodness-to-fit statistic (Standardised Root Mean Squared Error - SRMSE) is used to calibrate the SIM and to calculate the attractiveness value of each region based on the historical migration data, which is generated from patient registration data in the National Health Service (Lomax et al, 2014).

In order to integrate the migration model (SIM) into the demographic model, it is important to understand the variation of attractiveness and out-migration rates, as these two variables are the key inputs of the SIM. Two simple linear models were built to predict these two variables based on a series of socio-economic variables (Rees et al, 2004), including Population Density (PD), Total Employment (Emp), Average House Price (HP), Gross Value Added (GVA), Unemployment (Unemp), Average Road Speed (AS). A stepwise multivariate regression technique was applied to identify the most appropriate predictive variables. A location specified error  $e$  ( $e'$  in outmigration model) was introduced. Equations (5) and (6) represent the attractiveness and out migration model respectively.

$$Att_Y = K * PD_{Y-1}^{K1} * GVA_{Y-1}^{K2} * HP_{Y-1}^{K3} * Emp_{Y-1}^{K4} * AS_{Y-1}^{K5} * e \quad (5)$$

$$OMR_Y = M * HP_{Y-1}^{M1} * PD_{Y-1}^{M2} * Emp_{Y-1}^{M3} + GVA_{Y-1}^{M4} * e' \quad (6)$$

According to these equations, GVA and Employment are needed to estimate the local out migration rate and attractiveness. These two figures can be obtained from the economic model. In this study, a modified Cobb-Douglas production function (Canning and Pedroni, 2008) is chosen as the heart of the economic model. This can be written as:

$$Y_{year} = A_{year} * L_{year}^{\alpha} * C_{year}^{\beta} * I_{year}^{\gamma} \quad (7)$$

where Y is total production (the real value of all goods produced in a year); L is labour input (the total number of person-hours worked in a year); C is capital input (the real value of all machinery, equipment, and buildings), I is infrastructure capital, A is total factor productivity (TFP) which accounts for effects in total output caused by many other factors other than labour and capital, including technical innovation, organizational and institutional changes, education level etc. (Hulten et al, 2001); and  $\alpha$ ,  $\beta$  and  $\gamma$  are the output elasticity of labour, capital, and infrastructure capital. Assuming the investment on capital and labour are all from the earning (e.g. total production) of the previous year, we have:

$$I_{year} + C_{year} + L_{year} * W_{year} = Y_{year-1} \quad (8)$$

where  $W$  represents the average annual wage for a labour during a given year in a region. Therefore, when wages and capital are fixed, the optimised job opportunities can be estimated. Considering the commuting patterns, which is highly depends on transport infrastructure, a SIM based commuting model was built to allocate the residence to the working place, so the real local employment (Emp in Eq. (5) and (6)) can be estimated based on the commuting patterns.

A linear correlation between TFP and average annual wage can be observed from historical data. Assuming this relationship genuinely exists, the local TFP for each region can be estimated based on the regional specified average wage, which can be collected from national labour market statistics. Figure 3 a) shows the disaggregated TFP by NUTS2 region, 2006. Then, the projected local TFP for each NUTS2 region can be estimated based on the trend observed from historical data.

The operation of the model can now be summarised as follows. The primary objective is to project changes in population and Gross Value Added for each region under a variety of policy scenarios. Let us suppose for the sake of illustration that laissez faire policies are adopted without restraint on migration, and furthermore that infrastructure is able to adapt swiftly and smoothly to increased demand. For a given base year, the population, wage rate, TFP and capital employed are all known. Now assuming demographic growth then productivity in the following year will increase with greater availability of labour (from equation (7)). However population growth also affects regional attractiveness and migration rates in (2)-(6), which are also influenced by productivity and employment rates. Hence the dynamics of economic performance and demographic change are interdependent and iteratively.

### 3. Results and Discussion

In this paper, a High Speed Railway (HS2) scenario was built to test the impact of infrastructure development on demographic and economic growth. HS2 is an ambitious transport infrastructure development project which will connect the biggest cities in England by a Y-shaped high-speed railway network. The project will cost 46 billion pounds and be completed in 2032, according to DfT's plan. Figure 3 shows the evolution of population and GVA by NUTS2 region level during 2006 and 2050 under the baseline model (i.e. without HS2).

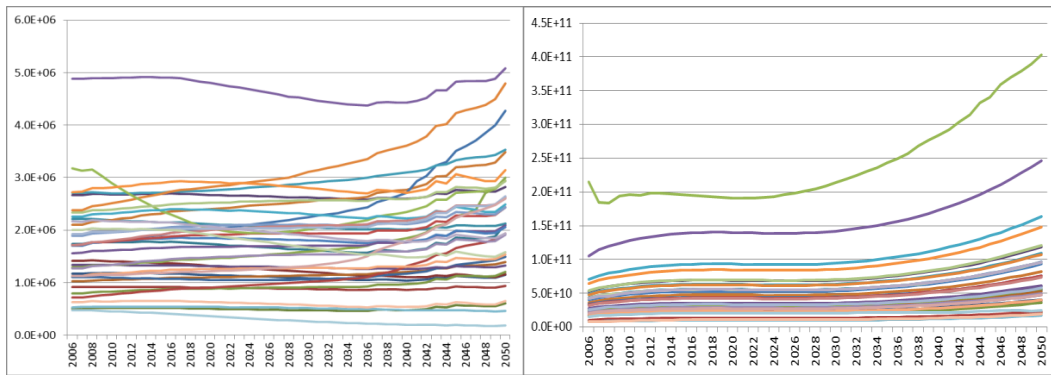


Figure 3 Time series of population and GVA by NUTS2 under baseline Scenario, 2006-2050

Figure 4 illustrates the change of commuting patterns for specific years.

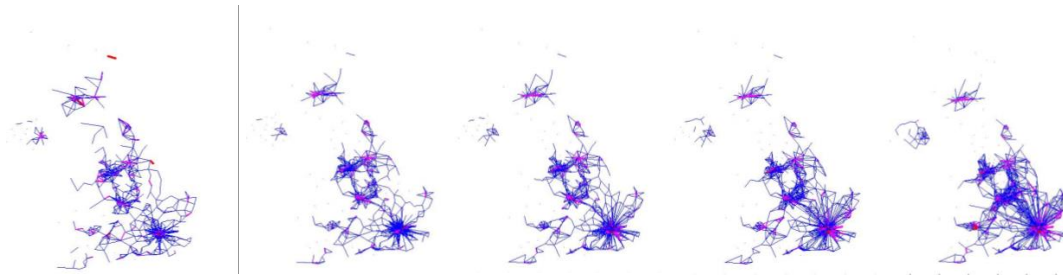


Figure 4 The commuting pattern in 2010, 2020, 2030, 2040 and 2050

The modelling results show there will be a significant increase of commuting flows between London, Birmingham, Manchester and Leeds, which indicates demand for HS2.

Figure 5 shows the impact of HS2 on the distribution of population and GVA in 2050.

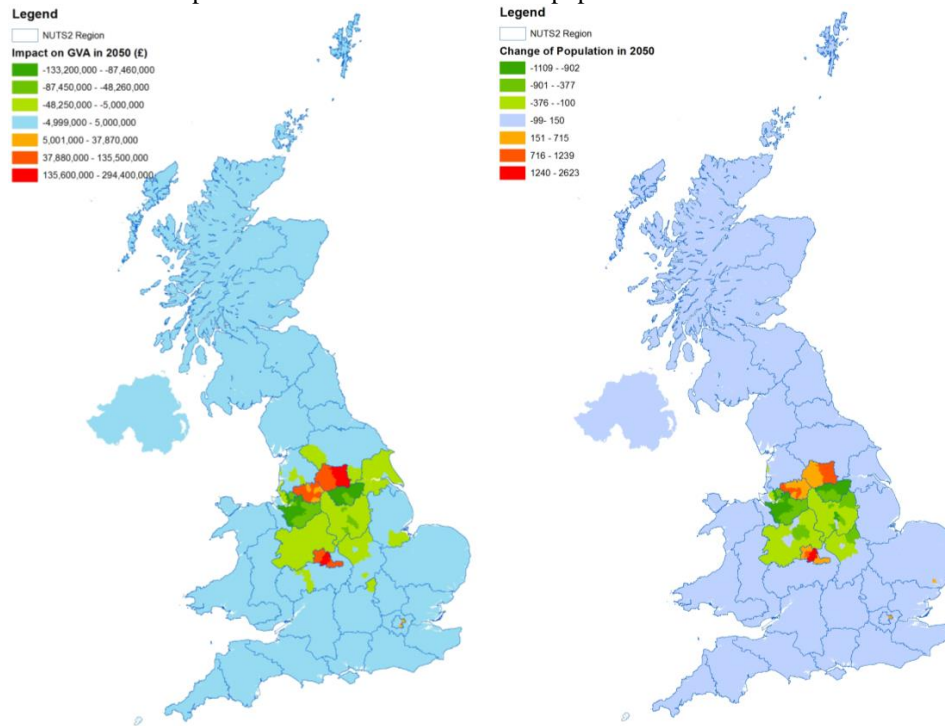


Figure 5 the impact of HS2 on GVA and Population in 2050

The modelling suggests that the HS2 will greatly increase the development of Birmingham, Manchester and West Yorkshire, while having some negative impacts on Cheshire and South Yorkshire. This phenomenon has some similarity to the agglomeration effects leading to accelerated growth of the major urban centres in China when the new High-speed rail network was made operational; (The World Bank, 2014). The reason for the phenomenon may be complicated. In our model, HS2 enables more labour working in the areas with higher economic efficiency, which generate more GVA giving rise to a higher attractiveness for migrants. Here the HS2 scenario shares a common international migration policy with the baseline, which restricts net international migration to today's level. This policy restricts population growth in the UK and also slows down the economy due to the shortage of labour.

#### 4. Conclusion and Discussion



The HS2 scenario presented here is offered as an initial proof of concept of the feasibility of implementing a co-dependency model of demographic and economic growth over a medium to long time horizon. However, the model still lacks an explicit representation of the evolution of infrastructure. The HS2 scenario is based on a rather abstract view of the role of infrastructure. Future work will explore this dimension in greater detail, for example by investigation of the relationship between total factor productivity and agglomeration, energy efficiency and knowledge exchange (Carlsson et al, 2013), as well as connectivity for which the HS2 scenario constitutes a preliminary test. In this way, a more sophisticated representation of the economy (than equations 7, 8) will be offered, while still recognising the positive relationship between labour, capital and economic performance. Alongside transportation, other key infrastructure sectors such as water, waste, ICT and energy will be represented. This programme promises a distinct and exciting perspective on the dynamic interplay of social and economic policies, regional development, infrastructure provision and prosperity.

## References

- Beaven, R., Birkin, M., Crawford Brown, D., Kelly, S., Thoun, C., Tyler, P., Zuo, C. (2014) Future Demand for Infrastructure Services, in Tran, M. et al (eds) Planning Infrastructure for the 21st Century, Cambridge University Press (forthcoming).
- Boucekkine, R., de la Croix, D., Licandro, O. (2002) Vintage Human Capital, Demographic Trends, and Endogenous Growth, *Journal of Economic Theory*, 104, 340-375.
- Canning, D., and Pedroni, P. (2008) Infrastructure, Long-Run Economy Growth and Causality for Cointegrated Panels, *The Manchester School* Vol.76 No.5, 504-527
- Carlsson, R., Otto, A., Hall, J. W. (2013). The role of infrastructure in macroeconomic growth theories, *Civil Engineering and Environmental Systems*, 30th Anniversary Special Issue, 30(3–4).
- Douglas, P.,(1976), The Cobb-Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values, *Journal of Political Economy*, 84(5), pp.903-916
- Dove, M. (1993) The coevolution of population and environment: The ecology and ideology of feedback relations in Pakistan, *Population and Environment*, 15, 89-111.
- Hall, J.W., Henriques, J.J., Hickford, A.J. & Nicholls, R.J. (2012, eds). A Fast Track Analysis of strategies for infrastructure provision in Great Britain: Executive Summary. Environmental Change Institute, University of Oxford.
- Holtz-Eakin, D., and Schwartz, A. (1995) Infrastructure in a structural model of economic growth, *Regional Science and Urban Economics* Vol. 25, 131-151
- Hulten, C., Dean, E., Harper, M. (2001). New Developments in Productivity Analysis, University of Chicago Press. pp. 1–54.
- Lomax, N., Stillwell, J., Norman, P. and Rees, P. (2014) Internal migration in the United Kingdom: analysis of an estimated inter-district time series, 2001-2011. *Applied Spatial Analysis and Policy*. Vol.7, pp25-45
- Miller, R., Blair, P. (2009) Input-output analysis: foundations and extensions, Cambridge University Press.
- Rees, P., Fotheringham, S., Champion, T. (2004) Modelling Migration for Policy Analysis, in

Stillwell, J., and Clarke, G. (eds.) *Applied GIS and Spatial Analysis*, Wiley, Chichester

Tran, M. et al (2014). National infrastructure assessment: Analysis of options for infrastructure provision in Great Britain, Interim results. Environmental Change Institute, University of Oxford.

The World Bank, (2014), Regional Economic Impact Analysis of High Speed Rail in China Main Report, Report No. ACS9734

Zuo, C., Birkin, M., Malleson, N. (2014) Dynamic Microsimulation Modelling for National Infrastructure Demand in an Uncertain Future, *GeoSpatial Information Science* Vol.17(3) 153-169