Challenges in Geocoding Socially-Generated Data

J. J. Huck[1], J. D. Whyatt[2], P. Coulton[1]

[1] School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA

[2] Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ

KEYWORDS: Geocoding, Social Networking, Twitter, Place Names, False Hotspots.

Corresponding Author:

Jonathan Huck
School of Computing and Communications,
Lancaster University,
Lancaster,
LA1 4WA

Abstract

An investigation into the difficulties facing researchers attempting to geocode data derived from social networking sites for analysis is presented. A number of issues are identified that arise from the differing levels of detail with which people describe their location, and from the reliance on place-names, which do not necessarily constitute unique identifiers for a given location. These issues include incorrectly locating data due to place name ambiguity, and the introduction of false hotspots into the data due to differing levels of detail in location data. A methodology is therefore presented in order to address these issues, and as such improve the quality and meaning of spatial analysis based upon geocoded socially-generated data, as well as other data where location information is reliant upon place names at a non-specific level of detail. The impact of this methodology is then illustrated with a simple analysis, allowing comparison between data that has and has not been 'processed' in the suggested manner.

1. Introduction

The purpose of this paper is to investigate the difficulties associated with attempting to geocode socially-generated data. This process relies upon the place-names that have been specified by users of the social network as 'their location' within their user profile, and as such raises a number of issues relating to the fact that there is not way of knowing the level of detail to which the user will have described their location. Furthermore, as place-names are not unique identifiers, they can prove to be unreliable as a description of a location, raising further issues in analysis.

1.1 Twitter

In recent years, we have seen a dramatic rise in the use of social media services such as Twitter (Lochrie and Coulton 2011). Twitter is a social networking service that allows users to share information, which is described by Twitter as "What's happening?" in the form of short text "tweets" which are limited to a length of 140 characters (Phuvipadawat and Murata, 2010). Over time, Twitter has become a tool for communication and collaboration, alerting the public, marketing, and the dissemination of news (Demirbas et al., 2010, Honeycutt and Herring 2009, Phuvipadawat and Murata, 2010). Indeed, the online monitoring of tweets often provides insight on events of several different natures, the repercussions of which can often be seen as a 'trending topic' more quickly than from regular news sources (Davis Jr et al., 2011).

As a result of this wide user-base, Twitter is generating a vast amount of data, which is developing beyond 'conversational' social interaction. These data are made accessible to the researcher through the provision of the Twitter Application Programming Interfaces (API's) (Twitter, 2011), allowing data such as tweet content and user information along with derived data such as trending topics (where a significant group of users are discussing the same topic) to be retrieved programmatically and analysed (Demirbas et al., 2010). This provides the potential for using not only the tweets themselves, but also demographic, temporal and geospatial data in

order to derive information about human feelings or behaviour in both time and space. Twitter users often will tag their message with a hash ('#') symbol, followed by keywords such as #royalwedding or #rw2011 in order to indicate the topic of the tweet, which makes it relatively straightforward to group tweets together by topic (Phuvipadawat and Murata, 2010); thus allowing trends or topics to be easily extracted and subsequently explored or visualised collectively where numerous people are tweeting about a similar subject (Field and O'Brien, 2010). With an increase in the use of GPS-enabled internet-connected devices such as smartphones, it is also possible for users to geotag their tweets with the location at which the tweet was published. Mapping such 'geotagged' data is a trivial yet powerful exercise, using the location of the tweet as a proxy for the location of the individual (Field and O'Brien, 2010).

1.2 Geocoding

Rushton et al. (2006) define geocoding as the process of assigning a geographic identifier to a computer record that lacks it, thereby tying information to geographic space. Goldberg (2011) expands upon this, describing it as "*a process critical to nearly every academic, industrial, and government field that seeks to perform any type of spatial analysis or mapping"*. Data returned from a geocoding service will generally comprise both address components and a coordinate pair, which may either be stored in a database, or used 'on the fly' and input directly into an application or analysis.

Traditionally, geocoding has taken place as a commercial service on an expensive standalone machine, with a skilled operator who is well aware of the difficulties involved (Davis Jr. and de Alencar, 2011; Roongpiboonsopit and Karimi, 2010). Nowadays, however, geocoding is used ubiquitously in modern web services (Jung et al., 2011), and is performed 'on the fly' by simply submitting a request to an online geocoding service such as the *Google Geocoding API* or *Yahoo! PlaceFinder* for little or no cost, nor any appreciation of the uncertainties involved (Roongpiboonsopit and Karimi, 2010). Each such service has developed an API, enabling a

flexibility that has made geocoding increasingly popular with both citizens and professionals alike (Davis Jr. and de Alencar, 2011).

Online geocoding services offer a number of advantages to the user, such as the removal of the requirement to prepare reference databases and algorithms. There are, however, a number of disadvantages as well, such as the removal of control over, and understanding of, the reference database; unknown data match quality; and lack of data confidentiality (Karimi et al., 2011). These problems are particularly significant at the global scale, with such a large volume of data collected from a wide variety of sources proving very difficult to maintain in terms of quality, resulting in a number of incorrect, non-official or out-dated address components being returned. For this reason, it is often better to discard address component data retrieved from geocoders where possible, and use the coordinate-pairs provided to extract address components from a more reliable or better-managed source.

Applications utilising online geocoding services vary from simply locating objects on a map (Marsh, 2010), to the provision of location-based services (Around Me, 2012), and even complex spatial analysis (ESRI Australia, 2010). As good locational data are fundamental to many applications, geocoding uncertainties and errors will be propagated through subsequent research, analysis, modelling and decision-making and, as such, will degrade the quality of the application in question (Karimi et al., 2011; Roongpiboonsopit and Karimi, 2010). A significant amount of research has already been undertaken towards the assessment of accuracy in geocoding (Zandbergen and Green, 2007; Zandbergen, 2007; Ward, et al., 2005; Bonner et al., 2003; Krieger et al. 2001) and the comparison of different geocoding service providers (Karimi et al., 2011; Roongpiboonsopit and Karimi, 2010; Whitsel et al., 2004), although this has largely been focused in the US.

Much of the literature on geocoding focuses upon "address matching", and the issues that arise from attempting to locate a given address on a street. This is because much of this work originates from the USA, where there is not a database of coordinates for all addresses, in contrast to the UK, where a number of gazetteers and the Royal Mail Postcode Address File (PAF) remove the requirement for such techniques. The rise of social media, however, has brought issues of geocoding to the fore once again, with the realisation that important insights may be gained from the use of such data in geographic analysis. For the most part, however, these data are not inherently spatial, and as such present the challenge to the researcher in terms of geocoding location based in the information associated with each datum.

When dealing with socially-generated data that do not have coordinates attached to them, the practice has traditionally been to geocode as best as is possible and then have analysis proceed based upon only those data which geocoded successfully; raising issues of bias (Curriero et al. 2010). In order to increase the amount of data which is successfully geocoded, a lot of research has also been conducted on the development of new geocoding systems (Jung et al., 2011; Charif et al., 2010; Curriero et al., 2010; Arikawa and Noaki, 2005; McElroy et al., 2003), though due to the 'black box' nature of the commercial online geocoders (Karimi et al., 2011), it is not known to what extent developments in academia have been adopted in the commercial world.

Noordhuis and Lazovik (2010) noted that the large amounts of data generated by social media platforms mean that traditional approaches to mining and processing do not scale well and are known to be expensive. This is also the case with geocoding, where online geocoding providers impose limits on the amount of data that users can geocode in order to prevent services from being exploited or overwhelmed. The *Google Maps Geocoding API* for example, has a limit of 2,500 requests per day for free (or 100,000 per day if you are a paid user) (Google, 2011),

whilst *Yahoo! PlaceFinder* has a limit of 50,000 requests per day (with the option to contact them to negotiate a greater number of results) (Yahoo, 2011).

As such, consideration has to be given to the volume of expected data collection, along with the capacity of infrastructure used for data collection, and the time and budget available for geocoding in order to ensure that data is not lost, and processing is completed in an appropriate timescale.

2. Background to Study

Both the use of social networking websites and applications using geocoding services have seen significant growth in recent years, with geocoding becoming cheaper and more accessible, and social networks such as Twitter increasingly becoming a source of useful information on daily events (Davis Jr et al., 2011). One consequence of this is that many people are attempting to locate socially generated data spatially, without fully understanding the issues associated with geocoding and the effect that these issues may have upon the quality of their analysis. Issues can arise because the actual location associated with a tweet is often rather uncertain, usually either based upon an automatic location based upon an IP address, or a user determined non-specific location in text format (Davis Jr et al., 2011).

The sample dataset used within this study is data collected from Twitter regarding the 'Royal Wedding' of Prince William and Kate Middleton, which took place on Friday 29th April 2011. This was chosen as a suitable event due to its highly emotive nature and the significant amount of media attention that it received. Data were collected for a month before and after the event (over 1.7 million tweets in total) using a PHP script that identified tweets related to the Royal Wedding through the Twitter API (Twitter, 2011), and loaded them into a MySQL relational database. Well over 100,000 tweets per day were captured in the days immediately surrounding the event, and further information on the event can be found at officialroyalwedding2011.org.

Due to the limitations of the Twitter search API at the time of data collection, the script was required to focus upon specific areas for search, which are illustrated in Figure 1 along with the geocoded location of each tweet. These limitations have since been solved with the Twitter Search API, and a 'global' search is now possible.

The spatial distribution of the data in Figure 1 is purely indicative, as the geocoding is a 'first pass' attempt using the 'Google Maps Geocoding API' (Google, 2011) that does not address any of the issues explored in this paper. There were high levels of activity in the USA and Europe, and to a lesser extent in Australia; though it should be noted that these areas coincide with the search radii that were used to capture Tweets (illustrated by red circles in Figure 1), and so may not represent the complete global distribution of Twitter activity relating to the Royal Wedding. Additionally, since the US-based Google Maps Geocoder (Google, 2011) was used to geocode the data displayed here, there is likely to be a positive bias towards the USA.

2.1 Place name ambiguity

Of the data collected from Twitter for this research, only approximately 1% contained coordinate data geotagged from a GPS-enabled device. Given this low proportion, the location for the vast majority of points in the dataset relies upon coordinates derived from geocoding the place-names that are specified within the profile information of each Twitter user as 'their location'. Place-names are described by (Longley et al., 2011) as the simplest form of georeferencing, that can be applied to any feature in the landscape (either physical or administrative), at any scale, and which may or may not be officially sanctioned.

As such, the use of place-names to locate twitter users in space is problematic, because place-names are not unique identifiers. Different places will sometimes share a common place-name. One example of this is 'Whitchurch', which when compared with the Ordnance Survey 1:50,000 gazetteer has 9 exact matches, and a further 9 approximate matches (e.g. Whitchurch-

on-Thames), all of which are located within the South of England and Wales (Figure 2). This creates an issue, as it is not possible to determine which of a number of places with the same name is the one being referred to by the twitter user, and thus a decision has to be made as to which of a number of coordinate locations will be used. This problem will generally worsen as the area included within the analysis increases, as the number of duplicate place-names will also increase. Similarly, single places often have multiple names, including vernacular or colloquial names such as '*The Big Smoke*' for London (in reference to the 'great smog' of 1952), '*Brizzle*' for Bristol (in reference to the local accent), or the politically incorrect '*Bradistan*' for Bradford (after the 1999 film 'East is East').

Further issues arise because place-names do not have any explicit 'level of detail' associated with them and as such, without any prior knowledge, there is no way to determine whether two locations are comparable based upon their place-names alone (Longley et al., 2011). For example, some users will describe in detail where they live, whereas (more commonly) others will give a very vague location such as '*Europe*' or '*Latin America*'. In addition, there will also be Twitter users who will enter extremely ambiguous text locations such as '*Whitchurch'*, or even false or humorous locations such as "*Teen World*", "*GLEE WORLD*" and "*World of chances* ☺ ". This problem persists when place-names are geocoded, as data returned from geocoders also lack any implicit scale (Whitsel, 2008).

Additional challenges arise from the comparison of geotagged tweets, and those where locations are derived from geocoded place-names. This is because data that have been geotagged provide coordinates specifying the location of the tweeter at the time they created the data; whereas those data that were geocoded using the 'location' set by the user in their profile provide coordinates specifying the location where people consider themselves to live. Whilst both would initially appear to provide a suitable proxy for the location of the twitter user, it should be noted that they are not representing the same thing: the location of the user when they published

a tweet is not necessarily comparable to the location where another user considers themselves to live. This is particularly important in the case of an event such as the royal wedding where many people travelled to London in order to experience the event. Field and O'Brien (2010) discuss this issue, and conclude that the Twitter profile location field is intended to be a permanent human-readable description, not a repository for ever-changing metadata.

2.2 False hotspots

In the case of social networking data such as tweets, where it is commonplace for locations to be specified as place-names that require geocoding to be placed on a map, it is likely that the coordinate data returned from the geocoder will relate to the centroid of an administrative area (such as a country, county or town). The 'level of geography' (county, town etc.) at which each location is returned is not known, and will vary from tweet to tweet, resulting in a dataset of locations described to a multitude of different levels of detail.

One of the major issues associated with geocoding socially-generated data is that there is neither an implicit scale associated with the data returned from a geocoder (Whitsel, 2008), nor a 'level of detail' associated with the textual representation of location given in a Twitter users profile. This presents a challenge, as geospatial analysis using such geocoded data must take place at a scale smaller than or equal to that of the data in order to avoid the introduction of false 'hotspots'.

If, therefore, such data were plotted onto a map without any knowledge of the level of detail associated with each tweet, then it is likely that false 'hotspots' will form at the centroid of administrative areas; appearing as a dense cluster of data-points on the map, but in reality being nothing more than an artefact caused by data at multiple different levels of detail being compared (e.g. a cluster of Twitter users who list their location as "England" should not be compared as like-for-like with a cluster of Twitter users who list their location as "Lancaster").

Figure 3, for example, shows a density map derived from tweet locations relating to the royal wedding, with no standardisation of the level of detail contained within each location before plotting. The resulting distribution of tweets across Great Britain (Figure 3a) exhibits three significant 'hotspots'. Two of these are expected, with one located in London: the most significant population centre and the location of the royal wedding itself; and another in, Blackburn, Lancashire; which Prince William and Kate Middleton visited three weeks before the wedding. The third hotspot, however, is located in the West Midlands, away from any significant population centres or any activity relating to the royal wedding.

The addition of bounding-boxes (Figure 3b) to the map helps to identify the cause of this hotspot. This point represents the centroid of the bounding box of 'England', and therefore this hotspot reflects nothing more than those twitter users whose location could not be resolved to a greater level of detail than 'England'. Furthermore, it is apparent that there are similar hotspots visible at the centroids of Scotland and Wales (circled in Figure 3b). These are examples of 'false hotspots'; concentrations of activity caused by the comparison of location data at different levels of detail, and not by actual twitter activity. A similar problem was reported by (Field and O'Brien, 2010) whereby their system located Twitter users that did not have a location specified in their profile at the intersection of the equator and the prime meridian (0,0), creating a false hotspot off the coast of Ghana.

3. Methodology

In order to analyse tweet data based upon user-defined locations (as opposed to geotagged tweets), whilst avoiding issues regarding place name ambiguity and the introduction of false hotspots, the following methodology is presented. The method comprises two distinct stages, firstly one to resolve which location will be accepted in the case of place names that have more

than one possible location, and secondly one to standardise the level of detail used for each location, in order that the user can avoid introducing false hotspots into their analyses.

3.1 Place Name Ambiguity

The first step in this methodology is to identify those place-names that are ambiguous (i.e. can be related to more than one location, which can be achieved by submitting all of the place-names to the geocoder and simply counting the number of results that are returned for each one. All of the returned locations are stored in a database, and each tweet classified either as: unique (the place-name resolves to only one location), ambiguous (the place-name resolves to more than one location), or invalid (the place-name can not be resolved to a location). At this stage, tweet data with invalid place-names are discarded from the analysis, tweet data with unique place-names are accepted, and tweets with ambiguous place-names are subject to further processing in order to resolve them to specific locations.

There are many different methods by which a location can be resolved from an ambiguous place-name. One such method is to apply Tobler's 'First Law of Geography', which states that; "*Everything is related to everything else, but near things are more related than distant things.*" (Tobler 1970). If this law is applied to the phenomenon of tweeting on a specific topic, it can be assumed that a tweet from an ambiguous location is likely to be close to the known locations of other tweets. The likelihood of each ambiguous location being the 'correct' location can therefore be inferred by the creation of a simple density surface of unique locations. Every potential location for each of the ambiguous tweets is assigned a value representing the density of unique tweets at that location, which can therefore be used to assess the most likely location.

Although it is not possible to define a definite 'correct' value, this method reduces the level of ambiguity, and increases confidence in the data compared to simply relying upon the 'black box' ranking value assigned by the geocoder. This method can act to remove some of the

'outlier' locations, examples of which can be seen in Figure 1 in isolated locations far away from the defined search areas. This approach changed the location of 34,850 tweets, which equates to 46.6% of the 74,722 ambiguous tweets or 2.1% of all tweets. This clearly demonstrates the importance of selecting an appropriate method for resolving ambiguous place-names, with almost half of locations resolved differently using this method, as opposed to accepting the rankings provided by the geocoder.

A decision should be made at the beginning of an investigation regarding the method that would be best for a given dataset, with the final decision likely to be dependent upon variables such as the number of tweets being examined, and the spatial extent of the analysis. It is, however, vital that some thought is given to how ambiguous locations will be resolved, and that this issue is not permitted to persist into analysis. Other methods that could be considered for resolving ambiguous place-names include the comparison of tweet timestamps at each location (following the presumption that a person may be more likely to tweet at 5PM than 5AM for example); investigation of the locations of other twitter users who 'follow' or are 'followed' by the user (following the presumption that a user is likely to be located closely to other users within their social network); or even allowing the geocoder to make the decisions with it's ranking system.

3.2 False Hotspots

The next step in the process is to determine a suitable scale at which analysis may take place in order to avoid the introduction of 'false hotspots'. The process of identifying the level of detail of each geocoded location is trivial, and the specifics will depend upon the format in which the data is returned from the researcher's chosen geocoder, but the principle involves simply counting the number of 'address components' that make up each location, compiling a list of all of the unique address components, and geocoding all of them, returning a coordinate pair for each level of detail included in each address (e.g. 'Lancaster, Lancashire, England' would provide three coordinate pairs). This process is illustrated in Figure 4.

The number of geocodable tweets for each address component can then be counted, and used in order to determine the most appropriate level of detail for analysis. Once a specific level of detail has been chosen (e.g. county level), the coordinate pairs associated with each tweet at that level of detail would be adopted for display and analysis. Any tweets with an insufficient level of detail will therefore need to be discarded from the analysis, whilst those with a greater level of detail than necessary will be located at a lower level of detail. There is, therefore, a trade-off situation, whereby analysis at a smaller scale will sacrifice detail but maximise the amount of data used, whereas analysis at a larger scale will sacrifice more data, but yield more detailed results. An example of the number of tweets in Great Britain with location information at each level of address detail is give in Table 1.

If the example given in Figure 4 is followed and the data is 'normalised' so that the county location is used for each of the tweets, then the result is as illustrated in Figure 5. The difference between this and the 'raw' data in Figure 3a is very clear, with the obvious removal of the false hotspots at the centre of each country.

4. Example Analysis

In order to demonstrate the impacts of these techniques, analysis will be performed on the royal wedding tweet dataset in order to investigate the level of tweet activity around the UK at the county scale. 613,877 tweets were collected from Twitter, geocoded using *Yahoo! PlaceFinder* and loaded into a MySQL relational database, with location data and tweet data stored in separate related tables. This approach is very flexible in that it allows tweet data and locations to be dynamically joined to each other, which is necessary in the case where there are number of possible locations for individual tweets, and where there are a number of coordinate pairs to attach to locations representing each component of an address.

This sample analysis will assess tweet activity across the various counties of the United Kingdom in terms of number of tweets per 1000 head of population of 'tweeting age'. Various studies into the demographics of Twitter users available online (such as Hepburn, 2010) suggest an age profile of 10 to 59. This analysis will take place using both 'raw' data returned by the *Yahoo! PlaceFinder* API (herein referred to as *raw*), and the same data post-processed using the techniques described in this paper (herein referred to as *processed*). Comparisons will then be drawn between the results derived from the two datasets in order to assess the impact of the additional processing described in this paper upon the outcome of the analysis.

Tweet location data within the United Kingdom were extracted from the database and mapped with ArcGIS, using both the 'default' geocoded locations for the raw dataset, and locations derived from the techniques employed in this paper for the processed dataset. County and Unitary Authority (UA) level 2001 Census data were then obtained from CASWEB (MIMAS, 2012), and the 'tweeting age population' established. In order to align the census data to 'address geography', UAs were dissolved into their respective counties (as addresses returned from the geocoder do not account for UAs). A spatial join process was then used in order to count the number of tweets within each county, and this figure was then divided by the 'tweeting population' of that region in order to generate a figure representing '*tweets per capita of tweeting age*' for both the raw and processed tweet datasets. The results were multiplied by 1000 in order to make the numbers more 'user-friendly'. This process is illustrated in Equation 1.

The results of both analyses are illustrated in Figure 6, with Figure 6a representing the results derived from the raw data, and Figure 6b representing the results derived from the processed data. The first difference to note is that the raw data represents 613,877 tweets, whereas the processed data only represents 511,470 (c.83% of the raw data), as is shown in Table 1. This reduction in data volume occurs because the locational information associated with these tweets

was too coarse for county-scale analysis (e.g tweets located at "UNITED KINGDOM or "ENGLAND").

The difference between the results generated from the two datasets is immediately obvious, with a very different spatial pattern created by each one. Much of this difference can be immediately attributed to the false hotspots at the centroid of each country (and the United Kingdom), which are clearly visible in Figure 6a. These hotspots cause high values in Northern Scotland (A), central Northern Ireland (B, Northern Ireland centroid), Eastern Northern Ireland (C, United Kingdom centroid), Western England (D) and Western Wales (E), reflecting the same pattern that was demonstrated in Figure 3a. The intensity of the patterns, however, does differ from that demonstrated in Figure 3a, notably with the increase in intensity of activity in Scotland and Wales. The reason for this is that both of these areas (the Scottish Highlands and Ceredigion in Wales) exhibit relatively low populations and, as such, the effect of the hotspots is intensified in an analysis of activity 'per head of population'.

Not only do these false hotspots give the appearance of a very high level of tweet activity in areas that actually exhibit a much lower level of activity (as can be seen by comparison with Figure 6b), but they also act to disguise the genuine levels of high of activity. An example of this effect is Lancashire (F), which stands out prominently on Figure 6b (due to the aforementioned visit to Blackburn by the royal couple), but not in 6a where the high concentrations of tweet activity attributed to false hotspots mean that real variations in tweet counts are masked by the simple cartography.

5. Discussion and Conclusions

Data published to social networking sources such as twitter provide a unique insight into the thoughts and feelings of a population, and as such are of significant value to researchers. The ability to locate these thoughts and feelings in geographic space adds further value to these rich

data, though before this value can be fully realised, confidence has to be developed that analysis will not be affected by place name ambiguity or the introduction of false hotspots, as if these phenomena are allowed to persist, then meaningful analysis of this powerful data will never be achieved.

Goldberg (2011) suggests that: "*there has never been a more clear need for georeferencing systems to be available to accurately process non-typical inputs, i.e. textual information other than traditional postal address data"*; and that: "*researchers, scientists, policy-makers, and other consumers of geocoded data must strive to understand the quality of the geocoded data they use in their research, practice or analysis*". Davis Jr. et al. (2011) agree that geocoding is no longer limited to addresses, and needs to be able to recognise and understand location from a diverse set of sources.

Attention must, however, be turned to the online mass geocoding services themselves, which are not robust. As the technology has shifted from being a specialist local activity to freely available global online services, the quality of the services has inevitably been reduced. Returned data are frequently out-dated or non-official, due to the mass data collection methods that are employed in order to maintain a global dataset of places, and the impracticalities of 'ground-truthing' and maintaining such a large dataset. One example of such an issue is that of 'Humberside'; a former county in England that ceased to exist in 1996, yet is still returned from the *Yahoo! PlaceFinder* geocoder.

The introduction of spurious locations in to the geocoding database will skew results, as tweets will be attributed to 'incorrect' as opposed to 'correct' locations, thus reducing levels of apparent activity. All tweets located at Humberside, for example, are therefore not attributed to the East Riding of Yorkshire, thus artificially reducing the level of activity shown in that county. This issue can be avoided by discarding the address information that is returned with

geocoding results, and simply using the coordinates to extract new address data from a 'standard' dataset (e.g. UK Census geography) using a spatial join. This is, however, a time-consuming process that would not be necessary if the datasets upon which the geocoding services are based were of a higher quality.

Similarly, the inflexible data structures returned from geocoders do not accommodate global variations in administrative geographies, which can lead to significant problems when analysis is taking place across a number of countries. One such example of this would be the sovereign state 'United Kingdom, which is considered a country by the geocoder. This has no negative effect upon the analysis when confined to the United Kingdom, but if addresses were to be compared with another country, the levels of address geography would be offset, as most countries do not have an equivalent address level, and as such, the geocoder data structures do not allow for one. If no adjustment were to take place, therefore, 'England', would compare to a county in another country, 'Lancashire' to a town, and so on. Once again, this can be accounted for as part of data preparation for analysis, but is an issue that would be more elegantly solved by the geocoding service. A similar issue was also experienced when place-names such as 'Middle East', 'Latin America', 'Europe', 'Benelux' and 'Caribbean' were submitted to the geocoder. None of these submissions returned location data as they are either informal, or do not fit within the pre-defined data-structure.

Part of the reason for these data quality issues with the geocoding services is that many of these systems were intended for use in online mapping applications (e.g. Google Maps, 2011) where data-quality issues would not have a significant negative impact upon the application, as opposed to scientific analysis where precision is vital. The upgrade of these systems in order to provide a more robust system suitable to support scientific analysis would be a significant step forward in the development of socially-generated data as a valuable resource for research. This is important, as even with the increase in uptake of smart-phones and other GPS-enabled

portable devices, the proportion of 'geotagged' tweets is still very low (circa 1% for the data collected for this work), indicating that place-name geocoding will likely remain the primary source of geolocation socially generated data for the foreseeable future. Even as uptake of GPS-enabled portable devices does increase over time, it is likely that the desire of users to maintain locational privacy by not publishing their specific real-time location online will prevent geotagging from replacing the need to geocode place-name data.

As the use of geocoding services and socially-generated data increases in both academia, and the media, the value of these data as a resource for gauging public interest and opinion will be increasingly recognised and exploited, allowing it to influence decision making. The spatial analysis of such data is an inevitable and already prevalent extension to this and, as such, maximising the quality of analysis is vital to ensuring that conclusions are meaningful and representative of true spatial patterns. Geocoding errors cannot be avoided completely (Karami et al., 2004; Zhang and Goodchild, 2002), but this paper has made progress towards improving the quality and reliability of analysis by demonstrating a process by which the quality of geocoded socially-generated data can be increased: both in terms of the removal of bias (by the eradication of scale-related 'false hotspots'); a reduction in the ambiguity arising from the use of free-text place names for locating tweet origin; and the avoidance of some issues arising from the poor quality of data returned from the geocoding services.

The examples given here have been restricted to the United Kingdom, but the issues identified in this paper will intensify towards the global scale, with more place-names and administrative areas increasing the scope for hotspots to form, and the probability of place-name duplication.

In any analysis where locational information is going to be derived from user-specified place-names, it is vital that false hotspots are removed, proper consideration is given to resolving ambiguous place names, and the effect of poor quality data returned from geocoders is

minimised. It is recognised that geocoding is an imperfect process, and so rather than trying to solve a problem which is inherent in the data, this paper aims to 'make the best of what we've got', and demonstrate a methodology by which analysis relying on this imperfect data source may be improved, and thus increased in reliability and value.

6. Acknowledgements

The authors thank Gemma Davies (Lancaster Environment Centre) for GIS support, and Mark Lochrie (Lancaster University School of Computing and Communications) for assistance with data collection.

7. References

Arikawa, M., and Noaki, K. 2005. Geocoding natural route descriptions using sidewalk network databases. In WIRI '05 *Proceedings of the international workshop on challenges in web information retrieval and integration.* Piscataway, NJ, IEEE Computer Society Press: 136-144

Around Me. 2007. Around Me. WWW document, http://www.aroundme.com/.

Beresford, A. R., and Stajano, F. 2003. Location privacy in pervasive computing. *Pervasive Computing Magazine*: 46-55.

Bonner, M. R., Daikwon, H., Nie, J., Rogerson, P., Vena, J. E., and Freudenheim, J. L. 2003. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* , *14*, 408-412.

Charif, O., Omrani, H., Klein, O., Schneider, M., and Trigano, P. 2010. A method and a tool for geocoding and record linkage. In *Proceedings of the Second IITA International Conference on Geoscience and Remote Sensing.* Piscataway, NJ, IEEE Computer Society Press: 356-359.

Curriero, F. C., Kulldorff, M., Boscoe, F. P., and Klassen, A. C. 2010. Using Imputation to provide location information for nongeocoded addresses. *PLoS ONE 5* (2): e8998.

Davis Jr. , C. A., and de Alencar, R. O. 2011. Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. *Transactions in GIS , 15* (6), 851-868.

Davis Jr., C. A., Pappa, G. L., de Oliveira, D. R., and Arcanjo, F. d. 2011. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS , 15* (6), 735-751.

Demirbas, M., Bayir, M. A., Akcora, C. G., Yilmaz, Y. S., and Ferhatosmanoglu, H. 2010. Crowd-sourced sensing and collaboration using Twitter. In *IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*. Piscataway, NJ, IEEE Computer Society Press: 1-9.

ESRI Australia. 2010. Australia Flooding Trends Map. WWW Document, http://www.esri.com/services/disaster-response/australia-flooding-map-2011/trends-map.html.

Field, K., and O'Brien, J. 2010. Cartoblography: Experiments in using and organising the spatial context of micro-blogging. *Transactions in GIS , 14* (s1), 5-23.

Goldberg, D. W. 2011. Advances in geocoding research and practice. *Transactions in GIS , 15* (6), 727-733.

Goldberg, D. W. 2011. Improving geocoding match rates with spatially-varying block metrics . *Transactions in GIS , 15* (6), 829-850.

Google. 2011. Google Maps API Web Services. WWW Document, http://code.google.com/apis/maps/documentation/geocoding/.

Hepburn, A. 2010. Infographic: Facebook vs Twitter Demographics. WWW Document, www.digitalbuzzblog.com/infographic-facebook-vs-twitter-demographics-2010-2011/.

Honeycutt, C., and Herring, S. C. 2009. Beyond Microblogging: Conversation and collaboration via Twitter. *Hawaii International Conference on System Sciences.* Piscataway, NJ, IEEE Computer Society Press: 1-10.

Jung, C., Knopp, S., Luxen, D., and Sanders, P. 2011. Efficient error-correcting geocoding. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* ACM, New York: 469-472.

Karami, H. A., Durcik, M., and Rasdorf, W. 2004. Evaluation of Uncertainties Associated with Geocoding Techniques. *Computer-Aided Civil and Infrastructure Engineering Computer-Aided Civil and Infrastructure Engineering , 19* (3), 170-185.

Karimi, H. A., Sharker, M. H., and Roongpiboonspit, D. 2011. Geocoding recommender: An algorithm to recommend optimal online geocoding services for applications. *Transactions in GIS , 15* (6), 869-886.

Krieger, N., Waterman, P., Lemieux, K., Zierler, S., and Hogan, J. W. 2001. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American Journal of Public Health , 91* (7), 1114-1116.

Krumm, J. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing 13* (6), 391-399.

Lochrie, M., and Coulton, P. 2011. Mobile phones as second screen for TV enabling inter-audience interaction. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology.* ACM, New York.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. 2011. *Geographic Information Systems and Science* (Third Edition ed.). John Wiley and Sons.

Marsh, B. 2010. *#uksnow Map*. WWW Document, http://uksnowmap.com/.

McElroy, J. A., Remington, P. L., Trentham-Dietz, A., Robert, S. A., and Newcomb, P. A. 2003. Geocoding addresses from a large population-based study: Lessons learned. *Epidemiology , 14* (4), 399-407.

MIMAS. 2012. CASWEB. WWW Document, http://casweb.mimas.ac.uk/.

Noordhuis, P., and Lazovik, A. 2010. Mining Twitter in the cloud: A case study. In *International Conference on Cloud Computing.* Piscataway, NJ, IEEE Computer Society Press: 107-114.

Official Royal Wedding. 2011. WWW Document, http://www.officialroyalwedding2011.org/.

Phuvipadawat, S., and Murata, T. 2010. Breaking news detection and tracing in Twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and*

Intelligent Agent Technology - Volume 03. Piscataway, NJ, IEEE Computer Society Press: 120-123.

Roongpiboonsopit, D., and Karimi, H. A. 2010. Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science , 24*, 1081-1100.

Rushton, G., Armstrong, M., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., et al. 2006. Geocoding in Cancer Research. *American Journal of Preventative Medicine , 30*, S16-S24.

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography , 46* (2), 234-240.

Twitter. 2011. Twitter API. WWW Document, https://dev.twitter.com/.

Twitter. 2011. Twitter Developers Documentation. WWW Document, https://dev.twitter.com/docs.

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., et al. 2005. Positional accuracy of two methods of geocoding. *Epidemiology , 16*, 542-547.

Whitsel, E. A. 2008. Error and bias in geocoding school and students' home addresses. *Environmantal Health Perspectives , 116 (8)* .

Whitsel, E. A., Rose, K. M., Wood, J. L., Henley, A. C., Liao, D., and Heiss, G. 2004. Accuracy and repeatabiity of commercial geocoding. *American Journal of Epidemiology , 160*, 1023-1029.

Yahoo. 2011. Yahoo! Placefinder. WWW Document, http://developer.yahoo.com/geo/placefinder/.

Zandbergen, P. A. 2007. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health , 7* (37).

Zandbergen, P. A., and Green, J. W. 2007. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives , 115* (9), 1363-1370.

Zhang, J., and Goodchild, M. F. 2002. *Uncertainty in geographic information.* London, Taylor and Francis.

Level of Detail	Number of Tweets	% of Total UK Tweets
Raw (non-normalised)	613,877	100%
Country	550,171	90%
County	511,470	83%
Town	470,565	77%
Better	60,654	10%

Table 1. The number of UK tweets containing location information at or greater than each level of address detail.

$$tweets\ per\ capita\ of\ tweeting\ age\ _{county} = \frac{tweet\ count\ _{county}}{population\ aged\ 10-59\ _{county}} \times 1000$$

Equation 1. Tweets per capita of tweeting age, used in the illustrative analysis for the effect of the processes described in this paper.

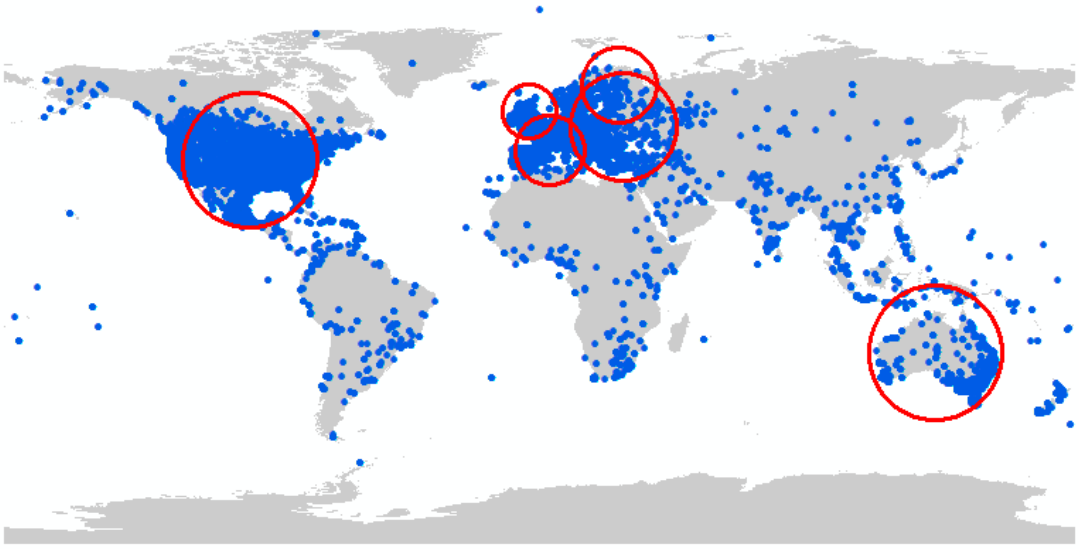

Figure 1. ‘First pass’ geocoded locations for the tweets collected within this investigation. The areas upon which the data collection focused are illustrated in red.

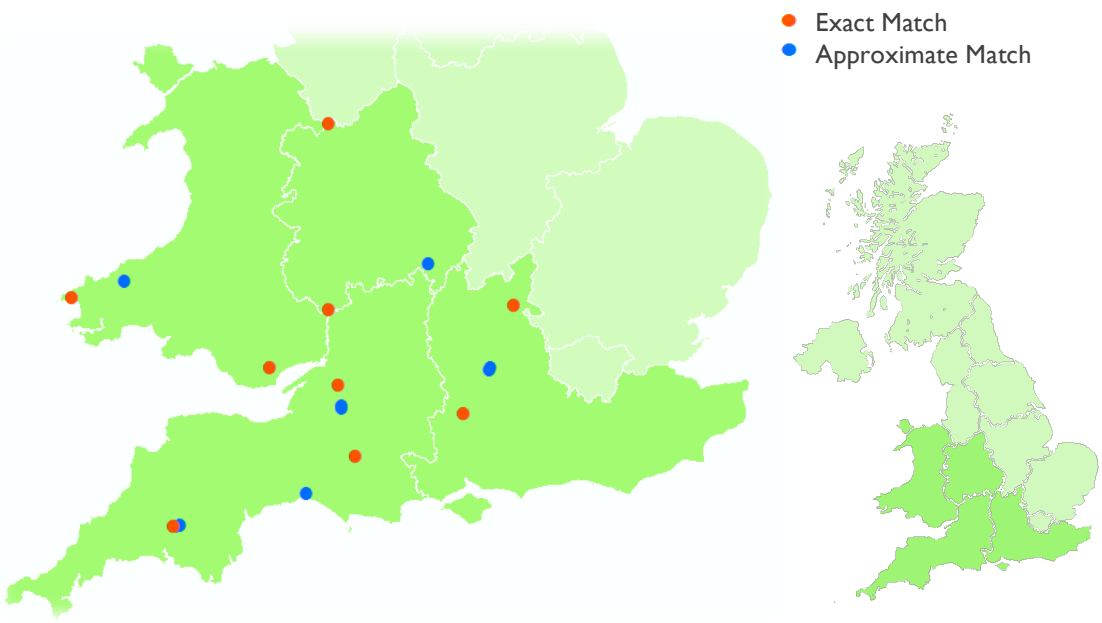

Figure 2. Places listed as being called 'Whitchurch' (orange), or similar (blue) according to the Ordnance Survey 1:50,000 Gazetteer.

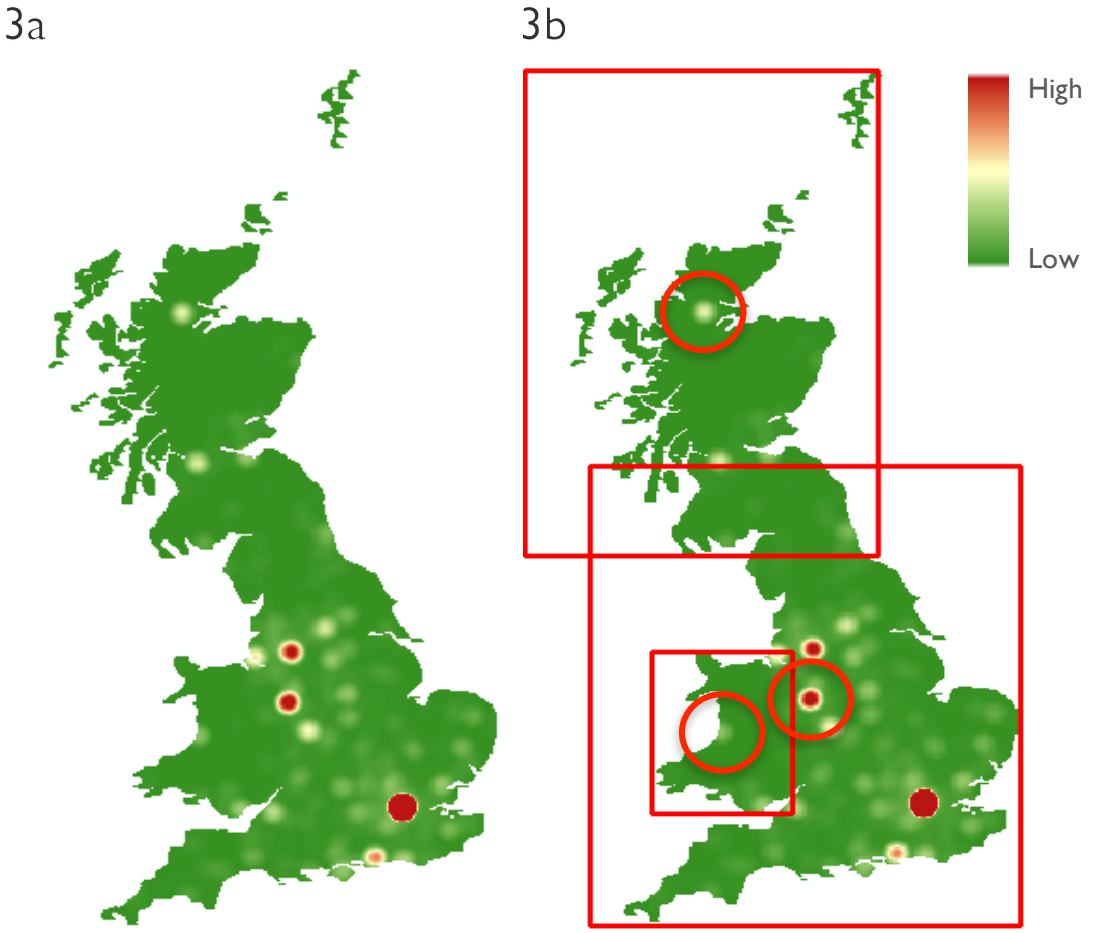

Figure 3. (a) A density map of 'first pass' geocoded tweet locations in the UK. **(b)** The same density map including bounding boxes for each country, and with the associated false hotspots circled.

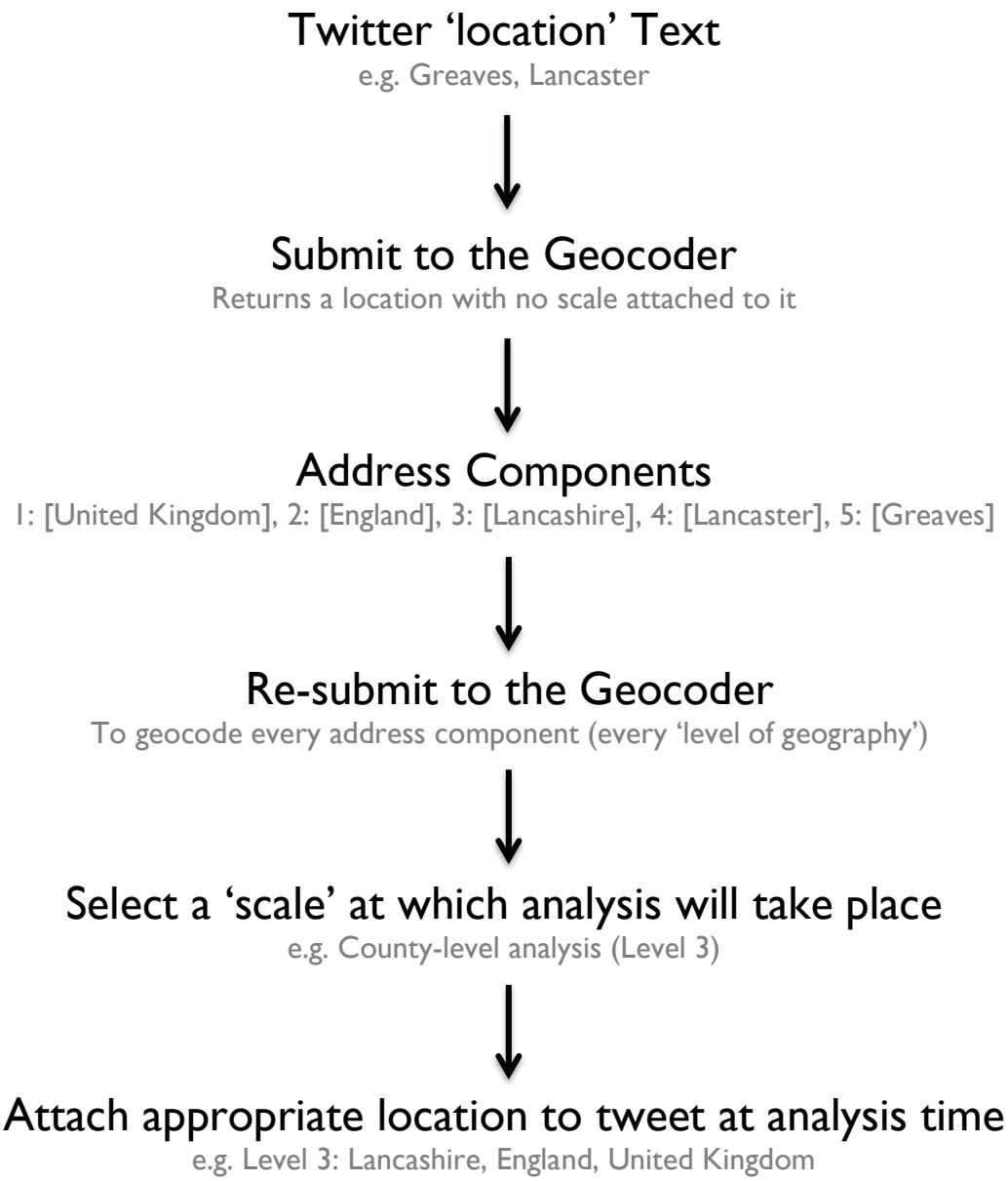

Figure 4. Flow diagram illustrating the 'normalisation' process of a tweet that has a greater level of detail than that chosen for use in the analysis.

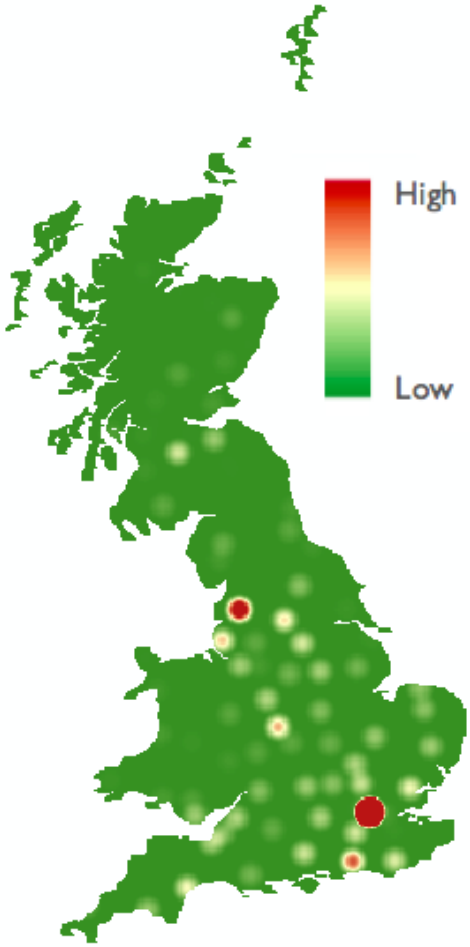

Figure 5: The royal wedding data normalised to the 'county' level of detail. The tweets are now located at the centroid of each county (as returned by the geocoder), and the false-hotspots at the centroid of each country (as illustrated in Figure 3b) are removed.

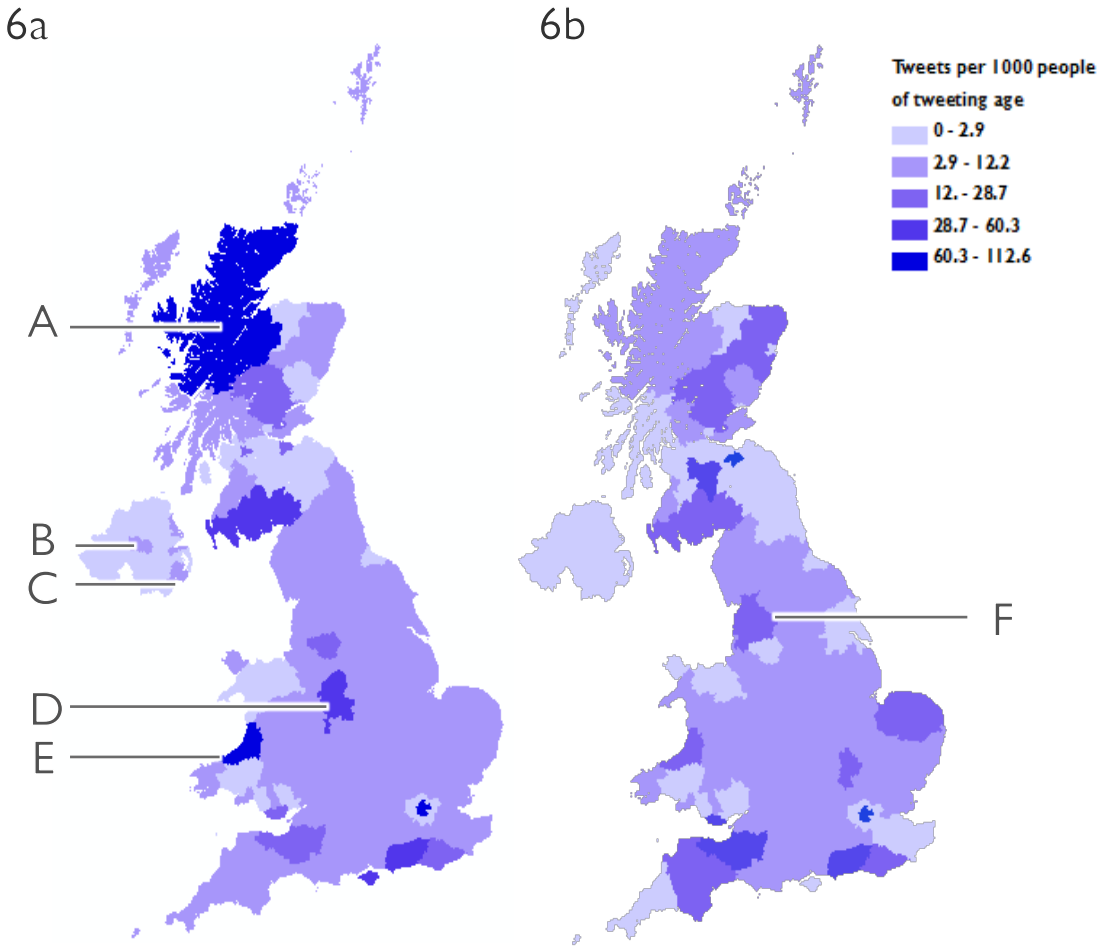

Figure 6. Tweet activity relating to the royal wedding per 1000 population of tweeting age. (a) illustrates the 'raw' data, and (b) illustrates the data having been processed according to this paper.

List of Illustrations: