Text, Images and Statistics: Integrating Data and Approaches using Geospatial Computing

Ian Gregory
Lancaster University
I.Gregory@lancaster.ac.uk

Abstract

Geographical Information Systems (GIS) originated as a quantitative technology with a social science paradigm. Its early uses in humanities disciplines followed this approach such that much of the best developed research in 'Historical GIS' is associated with quantitative statistical analysis. Technological advances mean that it has become increasingly easy to use GIS with qualitative sources such as texts and images. This in turn has led to an increasing uptake in geo-technologies across the humanities and the development of 'Spatial Humanities'. This paper explores how the different types of data can be used to deliver new research outcomes and examines how GIS allows different academic paradigms to be used in a synergistic way to bring together traditionally separate disciplines and approaches.

1. Introduction.

The arrival of Geographical Information Systems (GIS) in human geography in the late 1980s met with a mixed response. Enthusiasts welcomed its potential to "put Humpty-Dumpty back together again" [1], arguing that GIS had the potential to re-unite a fragmented discipline. Others were openly hostile and castigated the GIS approach as, for example, "the very worst sort of positivism" [2]. Twenty years on, GIS is now a well established part of human geography mainly because academics have demonstrated its ability to derive new knowledge that develop our understanding of the discipline regardless of whether one is interested in GIS *per se*.

The use of geospatial computing in the arts and humanities lags some way behind its use in geography, however the field is reaching maturity in some areas, particularly modern quantitative history, and is spreading rapidly into more qualitative sources and across the humanities more generally. This therefore seems an opportune time to explore the extent to which the use of GIS can be used to integrate different sources and cross traditional disciplinary divides.

2. GIS in Quantitative History.

The use of GIS in historical research, Historical GIS, is now a well established part of the discipline of history [3]. While early work focused on database construction and methodologies, the field has now evolved to an extent where it can be shown to have made a significant impact in delivering high-quality research in books and peer reviewed journals [4]. Many of these papers take a quantitative approach but it is instructive to consider how one of these studies made a contribution to knowledge and its limitations, as lessons can and should be learnt for approaches that are more directly concerned with the emerging field of spatial humanities [5].

A well established story in nineteenth century British social history is that rapid urbanization and industrialization led to overcrowded and insanitary urban areas which, in turn, led to horrific levels of mortality especially in the very young. Infant mortality rates were so high that one baby in five would die before its first birthday in some of the worst urban centers. The Victorians responded to this with a number of acts of parliament through the 1870s and '80s that aimed to improve sanitation and pre and postnatal care. By the 1900s infant mortality showed a marked decline which has continued ever since. The orthodoxy is thus that infant mortality was an urban problem which was solved by government intervention in public health. While this orthodoxy has been challenged, attempts to do so convincingly were stymied by having to focus on a subset of data, such as a small number of areas whose representativeness was unclear, or by using highly aggregate data that masked underlying variations.

Using a GIS enabled all data to be geo-referenced on a single set of administrative boundaries so that change from the 1850s, when the earliest comprehensive data are available, to the 1900s could be explored for every district in England & Wales. This revealed was that the biggest changes, in terms of the numbers of infant deaths averted by the declining rates, did indeed occur in urban areas, however, it also showed that the biggest percentage improvements occurred in areas that were rural and particularly found in the south and east of England. These declines started well before the legislation of the 1870s and '80s. This finding presents a direct challenge to the orthodoxy that mortality decline was caused by public health improvement. A different process drove down rates in many areas, particularly in the south and east, but its impact was masked unhealthy urban areas [6].

The strength of this study is it uses data on every recorded infant death over a 60 year period. By exploring the detailed geographies of these deaths and how these geographies changed over time it is able to challenge a long-standing orthodoxy. The limitation is that the data are limited to infant mortality rates, population density and location, so the study is unable to advance an explanation of its own. It can therefore be characterized as a wide but shallow analysis. To develop a convincing explanation would require a return to a more traditional humanities approach in which multiple sources are studied in-depth to gain a detailed understanding of particular areas. Limitations of time and sources mean that this type of analysis could not be conducted in a uniform way at national level. This may thus be characterized as a deep but shallow approach. While the two methods are very different, they are in fact complimentary – in-depth studies of a small number of locations provides an explanation which can be contextualized by the wider study which tells the researcher which areas followed the similar trajectories and thus for which the explanation may also be valid. Areas that followed different trajectories are likely to have an explanation that is at least somewhat different.

Thus the in-depth but narrow humanities-based approach and the wide but shallow social science-based approach can complement each other: one by offering explanation and nuance, the other by challenging orthodoxies and providing context. The challenges are how to apply the in-depth research using GIS, and how to apply this framework to other non-quantitative parts of the humanities such as Literary Studies.

3. Texts, Images and GIS.

Literary Studies is a discipline that might be thought to follow the equivalent of the in-depth but narrow approach as researchers traditionally carefully study a small number of texts to gain a nuanced understanding of them. Pressure of time ensures that only a small percentage of the total works within one genre can be studied by any one researcher. In his book *Graphs, Maps, Trees*, Moretti [7] argues for an alternative approach which he terms 'distant reading.' This involves the researcher using the three devises from the title to conduct a wider but shallower analysis. He uses the examples of: graphing the evolution and decline of the numbers of books published within different genres over time; mapping the locations of protagonists in Parisian novels; and using trees to explore the evolution of the role of clues in detective novels. While his use of graphs and trees is convincing, the maps section of his book, in which manual cartography is used, is less so. Producing his maps must have involved carefully reading the text and researching Parisian street plans. While the result may be an effective use of maps within Literary Studies, it is far from the ideal of distant reading.

To explore how GIS can help understand literary geographies we began the "Mapping the Lakes" project. This initially looked at accounts of two early tours of the Lake District, Thomas Gray's proto-Picturesque tour of 1769 and Samuel Taylor Coleridge's 'circumcursion' of 1802 [8]. We are currently working to extend this to include some of William Wordsworth's work [9]. The first stage of the project was to extract place-names from the texts. This was done manually, a considerable undertaking but one that forces the reader to pay close attention to the places named thus potentially making it a valuable form of in-depth reading. Once place-names were extracted they could be matched to a gazetteer and geo-referenced. The advantage of using a GIS, compared to Moretti's manual cartography, is that once the GIS database has been created it can be mapped, re-mapped, queried, integrated with other material, and manipulated in a wide range of ways. The creation of the first map in the GIS therefore represents a very early stage in the research process, it is a late stage in manual cartography.

The project produced a range of maps including: simple dot-maps of places mentioned, 'density smoothed' maps to summarize complex point patterns, and maps of emotional response to the landscape. Some maps were of the individual texts, others compared and contrasted the different texts [10]. More sophisticated forms of analysis integrated data from other sources such as a Digital Elevation Model of the Lake District and contemporary population density. Research showed that Gray followed the main valleys and based himself in towns. He rarely travelled to heights of more than a few hundred feet although the higher peaks, those of over 2,500 feet, attract considerable attention in his writing. Coleridge, by contrast, avoided the populous parts of the Lake District, staying in the Western Fells and climbing Sca Fell, the highest mountain in England. While his ascend (and hair-raising descent) of Sca Fell is well known, what is more interesting is that much of his account is also concerned with time spent in low places. Unlike Gray, he names places of all heights including those between 1,000 and 2,000 feet which Gray almost completely ignores. The two tours barely overlap, the only place where they do is Keswick, where Coleridge lived and Gray spent several nights, and the road between Grasmere and Keswick although neither account says much about this part of his journey.

GIS can thus be shown to be a useful tool for distant reading, allowing a wide but shallow approach to understanding the texts. We also wanted to explore whether GIS could help the in-depth reader. To this end we created a KML version of the GIS that we placed in Google Earth [11]. A window showing the text was put on the bottom half of the screen with a Google Earth map on the top-half. Super-imposed on the map were the locations mentioned in the texts, which can be switched on and off in various ways, and a contemporary map showing the Lakes in 1815. This allows the reader to read the text while following the locations named using either a modern or historical representation of the landscape as a backdrop. This enriches the experience of in-depth reading of the text by visualizing and contextualizing the places mentioned. This architecture also allows the user to click on a location to ask "what have the different writers said about this place?" To enrich this further, users can link from our site to the photographic website Flickr [12]. Flickr allows users to upload and share their digital photographs. These can be tagged with metadata such as 'landscape' or 'mountain' and users can also add 'geo-tags' that give latitude and longitude, providing the photo with a location. The geo-tags enable the reader to link from our texts and explore how modern visitors have photographed the landscape near to the places described by our historical writers.

Using this approach, Flickr allows us to show what the different areas of the Lake District look like today, and thus to assist in-depth reading. It is also apparent however that there are pronounced geographies within Flickr. Some areas are extensive photographed while others are ignored, and the different tags that people place on images also have pronounced geographies. We were able to extract the number of photographs geo-tagged to locations in cells of approximately 1km square across all of the north-west of England. This was done with all photos, and also with those with specific tags such as 'mountain(s)' or 'tree(s).'

Mapping all photographs produces some interesting geographies, in particular, most photos seem to be taken in the urban centers or the main valleys. Minor roads such as that over the passes of Wrynose and Hardknott, also seem to encourage photography. It may be therefore that modern visitors to the Lake District, at least as represented by people who upload geo-tagged photographs to Flickr, do so in a way that is more like the Picturesque tours of Gray than the Romantic experiences of Coleridge or Wordsworth. Further work is needed to explore this in more detail.

Geospatial computing and a distant reading approach thus allows us to integrate historical texts and modern photographs and ask new questions about representations of the landscape. Again, distant reading is better able to summarize patterns and develop questions than it is to provide explanations for the geographies it discovers.

4. Extracting place-names and context.

One limitation with the using the approach described above with texts is that extracting place-names manually is labor intensive and slow. Mapping meanings associated with these texts is even slower. If we are to harness the full potential of distant reading approaches we need to be able to geo-reference corpuses that potentially run to many millions of words. Automated extraction of place-names can be achieved by using corpus linguistics' approaches to identify proper nouns that may be place-names. Filters can then be devised to remove those that are people's names or other non-place-names. The remaining nouns can be matched to a gazetteer to see which match to the gazetteer entries. Disambiguation of multiple matches and handling of non-matches then needs to take place. While this process is not entirely automatic, it has been shown to be effective in mapping the places mentioned in an 800,000 word corpus of news-books published in London in the 1650s [13]. However we are not simply interested in *where* the corpus is discussing, we are also interested in *what* it is saying about these places. Again this can be achieved using corpus linguistics' approaches. *Collocation* finds words that are near to our place-names, while semantic tagging allocates meanings to these. In this way we were able to map, for example, locations referred to in relation to themes such as war, finance, and governance.

This approach drives distant reading to what is perhaps its logical conclusion – a largely automated process that rapidly produces maps summarizing large corpuses that the researcher need not even have read. This is thus an extreme form of a wide but shallow analysis which is unlikely to be sufficient to a researcher for two sets of reasons: first the process is inevitably error prone and we need to be able to check patterns for the impact of this, and secondly, as was established above, distant reading is good at answering *what, where* and potentially *when*, but is largely unable to say *why*. Both of these require a more in-depth approach, here again corpus linguistics can help. Our map of places tagged as 'I1,' those associated with finance, shows a cluster in Tunis. This seems strange so we can explore the original corpus for mentions of 'Tunis' near to words tagged as I1. This returns several appeals to "call the Turks to account at Tunis… for the injuries they have done unto the Christians." Clearly the word "account" has been mis-tagged as financial and this cluster can be dismissed as an error. Other clusters however prove robust. A major cluster appears associated with tag 'G3', war, in eastern Scotland. This is because there were multiple references to a rebellion that was occurring in Scotland at this time and the efforts to suppress it. This in turn could lead to further detailed research on this rebellion, representations of Scotland, and so on. The key point is that again there is an interplay between the distant and the in-depth approaches which is enabled and driven by the maps and the questions that the maps raise.

5. Conclusions: Integrating approaches

This paper has briefly explored how GIS can be used to map and analyze a wide variety of sources including statistics, texts and images. These sources can be approached in a wide variety of ways. At one extreme Google Earth enriches our understanding of a single short description of a Lake District tour. This uses a humanities-style approach to augment our understanding of the geographies in the text. At the other extreme we explore the geographies in large volumes of data such as: all infant deaths in England and Wales over 60 years, all of the geo-tagged images in Flickr for the north-west of England, or all of the surviving texts from news-books published in London in the mid-1650s. In all three of these analyses a quantitative social science-style approach was used to summarize large volumes of data and derive new questions from them. Both of the distant and in-depth methods have their limitations: the detailed approach is unable to say how representative this source or place is, the wide approach has very little ability to explain. They should therefore be taken as complementary and synergistic. In this way geospatial computing has the potential to allow researchers to use a disparate range of sources in both a wide social-science based paradigm and an in-depth humanities-style paradigm to deliver research findings that have the potential to make the kind of impact that is essential if spatial humanities are to become an accepted part of mainstream humanities scholarship.

6. References

[1] S. Openshaw, "A view on the GIS crises in geography or using GIS to put Humpty-Dumpty back together again", *Environment and Planning A*, 1991, 23, pp. 621-628.

[2] P.J. Taylor, "Editorial comment: GKS", *Political Geography Quarterly*, 1990, 9, p. 211-212.

[3] See for example Gregory I.N., and P.S. Ell, *Historical GIS: Technologies, methodologies and scholarship*, CUP, Cambridge, 2007 or Knowles A.K. ed., *Placing History: How maps, spatial data and GIS are changing historical scholarship*, ESRI Press, Redlands CA, 2008.

[4] These include the *British Medical Journal*, *Annals of the Association of American Geographers*, the *American Historical Review*, and the *Journal of Economic History.* See the H*GIS Research Network* website, http://www.hgis.org.uk/bibliography.htm. Accessed 24th September 2009).

[5] D.J. Bodenhamer, J. Corrigan, and T. Harris, "Introduction", in D.J. Bodenhamer, J. Corrigan, and T. Harris, *The Spatial Humanities: GIS and the Revolution in Humanities Scholarship*. Indiana University Press, Bloomington IN, In press

[6] I.N. Gregory, "Different places, different stories: Infant mortality decline in England & Wales, 1851-1911", *Annals of the Association of American Geographers*, 2008, 98, pp. 773-794

[7] Moretti, F., *Graphs, Maps, Trees*, Verso, London, 2005.

[8] I.N Gregory, and D. Cooper, "Thomas Gray, Samuel Taylor Coleridge and Geographical Information Systems: A Literary GIS of Two Lake District Tours" *International Journal of Humanities and Arts Computing,* In press.

[9] My thanks to Kirsten Hansen (Mt. Holyoke College, USA) for her work on this while at Lancaster in the summer of 2009 and Robert Schwartz for setting this internship up.

[10] Mapping the Lakes website. See http://www.lancs.ac.uk/mappingthelakes. Accessed 24th September 2009.

[11] http://earth.google.com. Accessed 24th September 2009.

[12] http://www.flickr.com. Accessed 24th September 2009.

[13] A. Dunning, I. Gregory, and A. Hardie, "Freeing up digital content with text mining: New research means new licenses" *Serials*, 2009, 22, pp. 166-173.