Combining Statistics and Texts Using GIS: Nineteenth Century Health Reports

Catherine Porter[1], Paul Atkinson[1] and Ian Gregory[1]

[1]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. 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,

[1] c.porter2@lancaster.ac.uk
[1] p.atkinson3@lancaster.ac.uk
[1] 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

Crowding	**Food and Waterborne**	**Respiratory**
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.

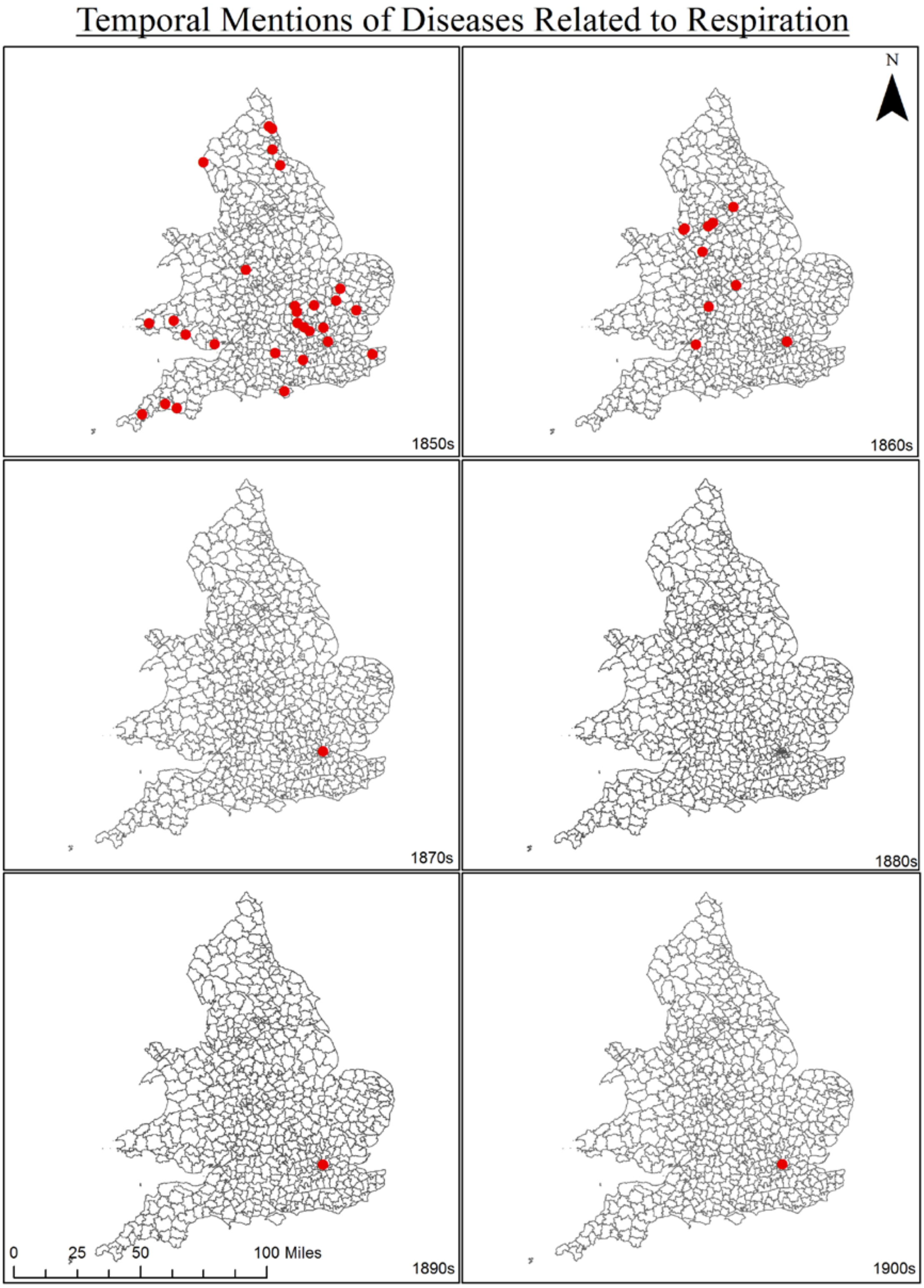

Figure 1 The Registrar-General mentions related to respiratory diseases mapped on the Registration Districts for England and Wales and displayed for each decade of the research time frame

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).

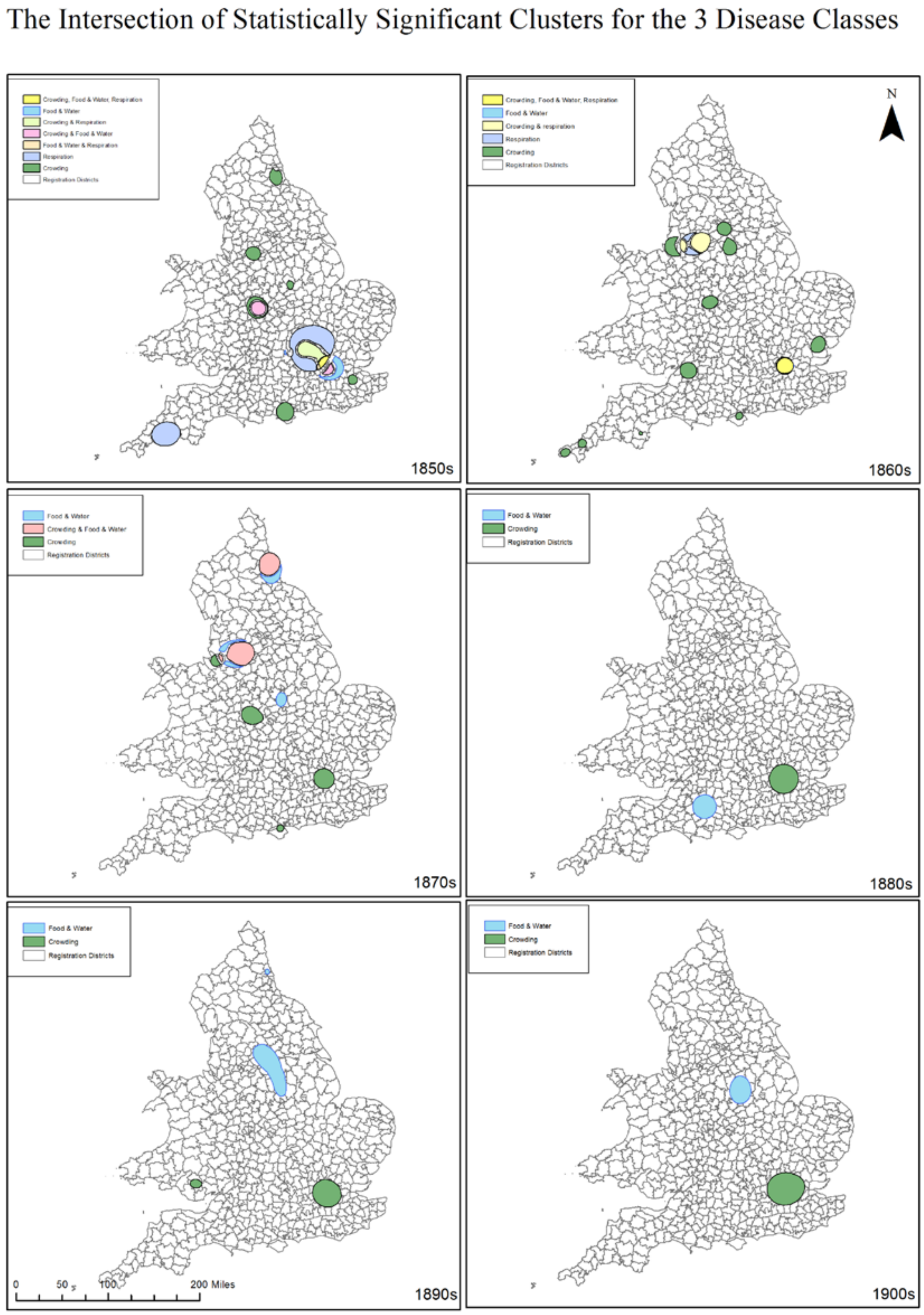

Figure 2 The 'unioned' density polygons of disease mentions for each of the three disease classifications and decades of the research time frame. The areas in yellow show where all three classifications coincide.

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.

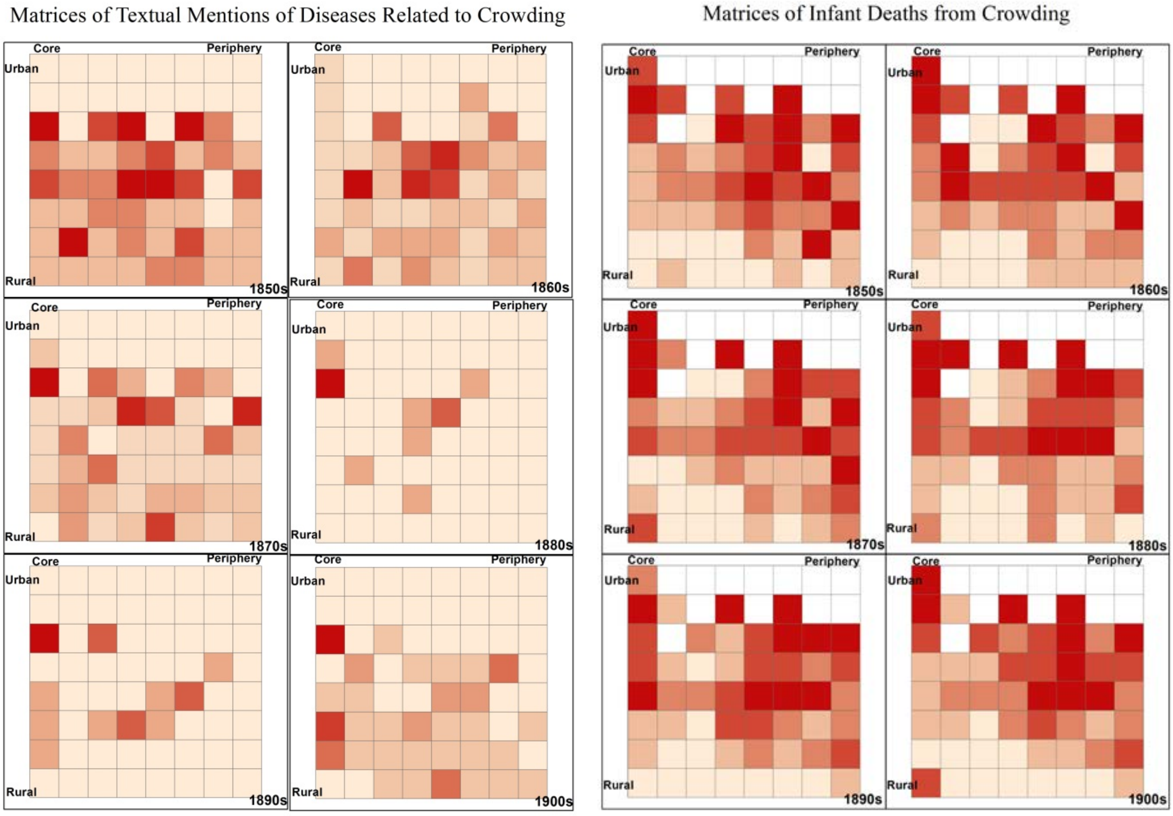

Figure 3 The Hierarchical Regional Settlement matrices for crowding diseases, the mentions by the Registrar General (left) and the mortality figures (right).

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.