Geographical Text Analysis: a new approach to understanding nineteenth-century mortality

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

[1]Department of History, Lancaster University, Lancaster LA1 4YT, United Kingdom

c.porter2@lancaster.ac.uk; p.atkinson3@lancaster.ac.uk; i.gregory@lancaster.ac.uk

http://www.lancaster.ac.uk/spatialhum/

+44(0)1524 592769

This is an author accepted manuscript of a paper that will appear in Health and Place (http://www.journals.elsevier.com/health-and-place/).

Geographical Text Analysis: a new approach to understanding nineteenth-century mortality

Abstract

This paper uses a combination of Geographic Information Systems (GIS) and Corpus Linguistic Analysis to extract and analyse disease related keywords from the Registrar-General's Decennial Supplements. Combined with known mortality figures, this provides, for the first time, a spatial picture of the relationship between the Registrar-General's discussion of disease and deaths in England and Wales in the nineteenth and early twentieth centuries. Techniques such as collocation, density analysis, the Hierarchical Regional Settlement matrix and regression analysis are employed to extract and analyse the data resulting in new insight into the relationship between the Registrar-General's published texts and the changing mortality patterns during this time.

KEYWORDS: Geographical Text Analysis; Corpus Linguistics; GIS; Infant Mortality; Registrar-General.

1.0 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). In the absence of substantial records on morbidity, the Registrar-General's reports on patterns of mortality, including cause of death, are central to any population study occupying this time period. The large and at times controversial literature on nineteenth-century changes in mortality was well summarised by Woods (2000), while Eyler (1979) studied the work of William Farr, the first superintendent of statistics at the GRO. Short treatments of the GRO's history also exist (Szreter, 1991; Higgs, 2004) but the narrative sections of Registrar-General's Reports remain an under-used resource and hitherto no project has attempted to investigate the relationship between the two and a quarter million words contained in the reports and actual population mortality figures.

In this paper we merge two methodologies that are ordinarily separate, namely Geographical Information Systems (GIS), a technology usually employed to analyse the spatial patterns

within quantitative data, and corpus linguistics, a method which is used to analyse large volumes of digital texts but which, to date, has largely ignored geography. By combining these to create a set of techniques called Geographical Text Analysis (GTA) (Gregory et al, 2015; Murrieta-Flores et al 2015) we are able to explore the geographies within the Registrar-General's Reports. This is achieved by extracting disease related keywords and associated place-names from the reports and allows us to examine which diseases the Registrar-General was most interested in, which places he associated with these diseases, and how this changed over time. By combining these with the Registrar-General's mortality figures, it allows us to compare and contrast these patterns with the actual patterns of disease mortality.

The analysis of the data is three-fold. The first stage develops and establishes the categories of disease for analysis. The second, uses Corpus Linguistic Analysis to extract the Registrar-General's mention of disease related to place-names, and the third, focuses on comparing the textual outputs with mortality statistics derived from the Registrar-General's reports. Throughout, the use of GTA provides a new view on the nineteenth and early twentieth century world according to the Registrar-General by contributing new insight into whether his published texts were directly related to changing mortality patterns during this time.

2.0 Data and background

The primary dataset used to explore the relationship between the Registrar-General's reports and mortality is the Histpop collection (www.histpop.org), which provides online access to the official population reports for Britain and Ireland from 1801 to 1937. Within this, the Registrar-General's reports and accompanying Decennial Supplements are a record of collated statistics on births, deaths and marriages published since 1837, the main material of interest often being written by the Superintendent of the Statistical Department rather than the Registrar-General himself. For the purpose of this paper, the period 1850-1911 is of particular interest because it encompasses the beginning of the major decline in mortality that characterised the twentieth century. This puts the focus on the reports of the four statisticians employed between 1850 and 1911 (William Farr, William Ogle, John Tatham and THC Stevenson) and is the earliest time period that reported on cause of death specific to local areas. These published documents, in addition to statistical tables, include a discussion of the types of diseases and related places of interest to the GRO during these decades. From these data it is therefore possible to gain not only the actual mortality figures for the time period,

but also the GRO's discussion of the related diseases and places in which these diseases largely occurred or indeed were less prevalent.

The GRO's statistical data were extracted from the Great Britain Historical Geographical Information System (GBHGIS) (Gregory et al., 2002), a database that focuses on historical statistics such as census data and GRO reported information such as births, deaths and marriages. Within the GBHGIS these data are spatially linked with the Registration Districts from the time, these administrative boundaries the basis on which the GRO collected and collated population data. Registration Districts were established in 1837 on the basis of the recent Poor Law Unions, each having its local registrar and registry office (Higgs, 2, 24). As such, the 635 Registration Districts in vector polygon format were utilised as the basis for the analysis and provide the primary spatial structure to this paper.

As the paper primarily assesses the link between mortality and the GRO's discursive work, infant mortality (population aged under one year) was chosen as a focus because of its salience as an indicator of overall health conditions. Titmuss (1943) called it 'a measurement of human progress', adding: 'The toll of infant deaths is today, just as it has always been, a broad reflection of the degree of civilisation attained by any given community.' Decennial infant mortality figures related to specific diseases were extracted from the GBHGIS and the Infant Mortality Rate (IMR) summarised by Registration District for use in the analysis.

2.1 Classifying disease

Researchers are familiar with the difficulties created by the Registrar-General's changing classification of cause of death (Hardy, 1994). Woods and Shelton's *Atlas of Victorian Mortality* (1997) discusses how far causes of death may be linked in equivalent groups from one Decennial Supplement to the next. Woods (2000) proposed three analytical groups of disease linked to different aspects of the environment, each offering hints about different possible factors affecting change in mortality: (I) diseases of crowding, related to housing adequacy and population density; (II) those related to food and waterborne disease, relating to the effectiveness of Victorian sanitary reform, and; (III) respiratory diseases (excluding tuberculosis), linked to air quality. These categories can be used to classify the recorded deaths in this paper under the three main headings of Crowding, Food and Waterborne, and Respiratory. The implementation of this scheme is shown in Table 1 and includes prevalent

causes of death in the nineteenth-century such as measles, scarlet fever, cholera, diarrhoea, dysentery and pneumonia. It should also be noted that spelling and nosology of diseases varied with alternates such as, 'Scarlatina' and 'Scarlet Fever' (Hardy, 1993), 'whooping cough' and 'hooping cough', being common, so these were also taken into consideration in the creation of this scheme.

Crowding	**Food and Waterborne**	**Respiratory**
Diphtheria	Cholera	Bronchitis
Measles	Diarrhoea and Dysentery	Diseases of Lungs
Scarlatina	Enteric Fever	Diseases of Respiratory System
Scarlet Fever	Simple Continued Fever	Influenza
Small-pox	Typhoid	Pneumonia
Typhus	Diseases of the Digestive System	
Whooping cough		

Table 1: The three disease categories based on Woods (2000).

2.2 Statistics on infant death

The GRO's discussion of disease is key to this paper but also of interest is the infant population that died from said diseases during the study period. Initially these data may be described by using a graph of Infant Mortality Rate (IMR) for each decade and for each of the disease categories previously described (Figure 1), the IMR being infant deaths per thousand live births and 'infant' meaning children under the age of 1. First to be exemplified in the graph is the decline in infant deaths related to Crowding, a pattern most likely linked to the decline in virulence of scarlet fever and diphtheria during the nineteenth-century. The Food and Waterborne plot is also unsurprising as it illustrates the sharp rise in infant deaths in the 1890s due to the number of warmer than average summers during this decade and the related prevalence of diarrhoea in the population, a key cause of death in infants. Of particular interest, is the pattern of IMR for respiratory related diseases as this exhibits a rising rate of infant respiratory deaths from the 1850s through to the 1890s, despite the introduction of measures by the Victorian government to assist in public health (Smith, 1979). Woods and Shelton (1997) also recognised this pattern which they believed may show a true rise in the rate of infant mortality from these causes, but they were equally troubled by the possibility that this was an artefact of a change in the Registrar-General's nosology, or indeed in cause of

death registration practice. This theory stems from the fact that during the nineteenth-century doctors did not record the cause of infant deaths with much precision (Hardy, 1994) and no data is available about the changing breakdown of diseases within this Respiratory group over time.

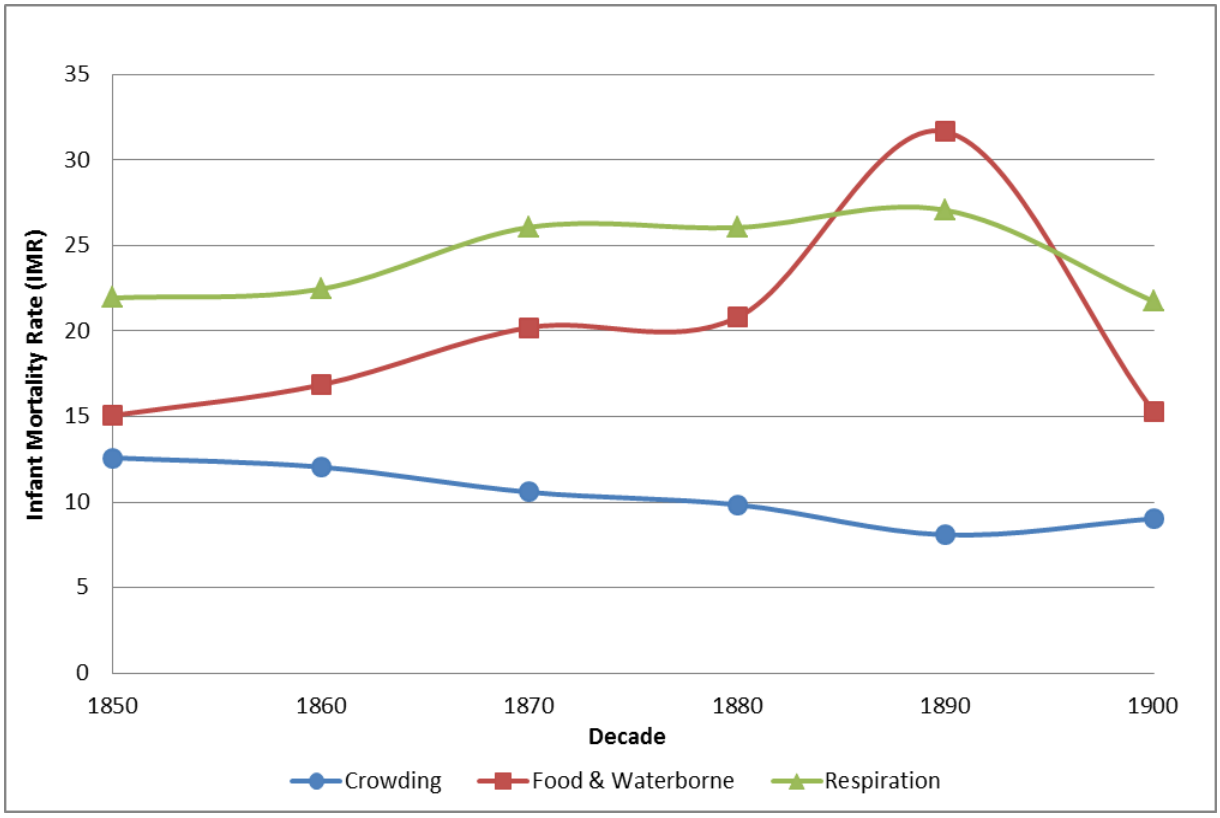

Figure 1: Infant Mortality Rate (IMR) of each disease classification per decade of the study, 1850-1900 (Source: GBHGIS (Gregory et al, 2002)).

2.3 Texts on causes of death

As well as publishing tables of statistics on causes of death, the Registrar-General's reports from 1851 to 1911 also contain two and a quarter million words of text in which he discusses the evidence presented in the tables. While we are familiar with analysing statistical data, analysing large volumes of digital text is less familiar, particularly when geography is one of the main themes of interest. Corpus linguistics provides the tools to allow a *corpus*, a large body of digital text, to be explored using a combination of quantitative and qualitative approaches (McEnery & Hardie 2012). One basic corpus linguistic technique is to search the corpus for occurrences, or *instances*, of a search term. As well as providing quantitative evidence of how common the search-term is, and potentially how this varies over time. This also allows the *concordances*, the text around the search term, to be explored to build up a

qualitative impression of what is being said about the search term (the online software CQPweb (Hardie, 2012) was used to implement the corpus linguistic analysis). For example, a search for "measles" reveals that there are 684 instances in the corpus, a frequency of 304.26 instances per million words. Example concordances include: "…The Epsom district suffered from scarlatina; Guildford from small-pox and *measles*; Farnham from fever, measles, hooping cough, and diarrhoea. The deaths for the first time…"; "…district suffered from scarlatina; Guildford from small-pox and *measles*; Farnham from fever, *measles*, hooping cough, and diarrhoea. The deaths for the first time exceed the…"; and "…somewhat 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, …" Even from this very simple list we can see that a common theme is the Registrar-General describing measles and other diseases in relation to place. Additionally, the discourse being used is very descriptive, acknowledging that the disease is occurring but not exploring its causes or methods for reducing it.

The concordances above clearly show that the Registrar-General associated disease with place, a finding that may be examined by exploring concordances more widely; however, traditional corpus linguistics does not explicitly handle geography. Geographical Text Analysis is designed to overcome this limitation. It relies on first *geo-parsing* the corpus, a two-stage process in which place-names are first identified and are then allocated to co-ordinates using a gazetteer (Grover et al, 2010). The results are incorporated into the text using XML tags. One of the simplest implications of this is that the results of a concordance can be mapped to explore, in this case, which places the Registrar-General was associating with particular diseases.

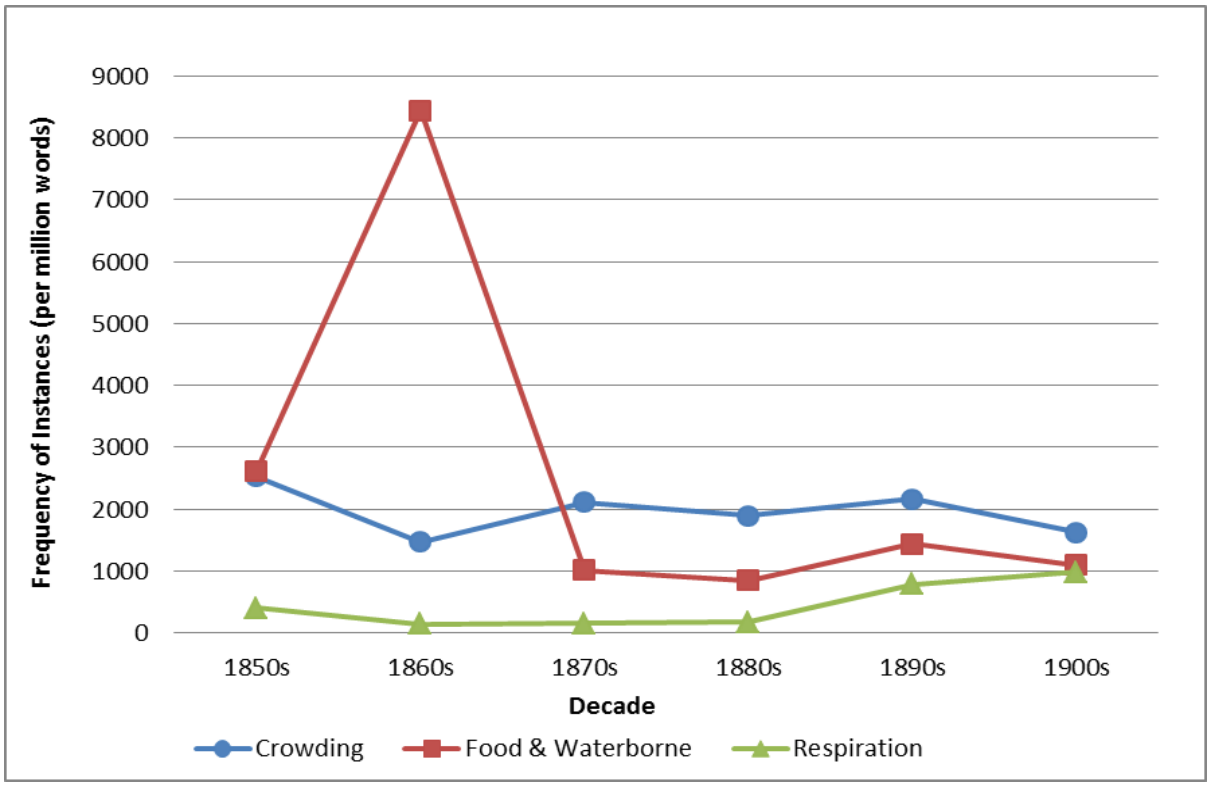

Figure 2: Instances (per million words) of each disease classification in the corpus, per decade of the study, 1850-1900.

Before turning to geography, however, the first stage is to explore some of the more general trends in the text. Figure 2 shows the frequency of instances of diseases in the three main disease categories. Comparing these patterns derived from the Registrar-General's texts with evidence from his statistics from Figure 1 reveals an interesting conundrum: respiratory diseases had the highest mortality rates among infants in five of the six decades but have consistently low rates of mention in the reports. In addition to this, as mentions of this disease classification increase near the end of the period (1890-1900) the mortality rate for these diseases starts to decline. Opposing this, crowding diseases have consistently the lowest IMRs but are generally the ones that the reports concentrate on the most. Food and waterborne diseases typically lie in between the two. The peak in instances in the 1860s can be explained by the supplement to the 1868 *Report on the cholera epidemic of 1866 in England: Supplement to the twenty-ninth annual report of the Registrar General* which contains over 3,000 instances of "cholera," however the surge in deaths in the 1890s is not matched by a corresponding surge in interest. The general pattern this reveals is that there is little evidence that the diseases the Registrar-General was most interested in were actually the ones that had the highest mortality rates.

3.0 Respiratory diseases

According to the previously described instances of disease (Figure 2) and the IMR recorded per disease category (Figure 1), the Registrar-General's discussion of respiratory-related diseases in particular warrant further investigation. The IMR figures showed that respiratory-related deaths were the highest of the three classes for five of the six decades and that they show little evidence of decline over the period. Despite this, the Registrar-General seemed noticeably less interested in respiratory disease than in the other two classes. Why then were so few overall mentions by the Registrar-General recorded in this disease category and do these correspond with the places where respiratory related disease was at its highest? To analyse these questions further a spatial picture of the respiratory related discussion was created by mapping the Registrar-General mentions (for all decades) that collocate with place-names (Figure 3). In addition to this, the number of collocates were mapped in three classes as low, medium or high to provide a first view of which places were of greatest interest or focus to the Registrar-General.

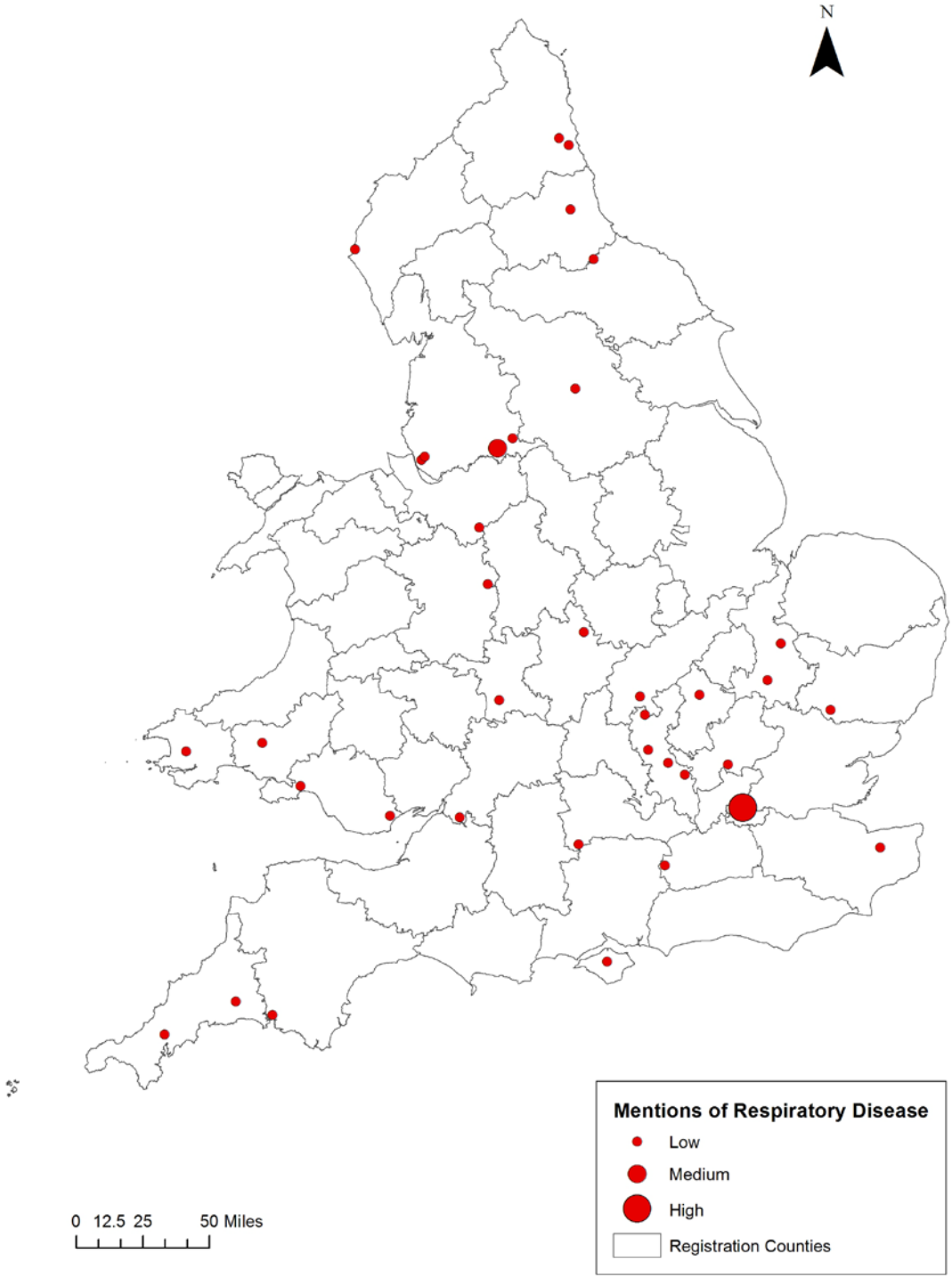

Figure 3: The geographic spread of the Registrar-General mentions of Respiratory related diseases that collocate with place-names in England and Wales, 1850-1900. The mentions are classed as Low, Medium and High depending on the number of mentions made by the Registrar-General of a particular place-name.

This spatial picture of respiratory related interest by the Registrar-General shows that the mention of this disease category was geographically widespread between the 1850s and 1900s, and included discussion of many of the major industrialised settlements in England and Wales such as London, Manchester, Liverpool and Newcastle (London and Manchester having the greatest number of mentions) in addition to a range of more rural places. However, what this spatial depiction does not provide is a breakdown of the change in mentions related to places over time. To aid this temporal investigation the disease mentions were mapped again by dividing the data into the decades of study (Figure 4).

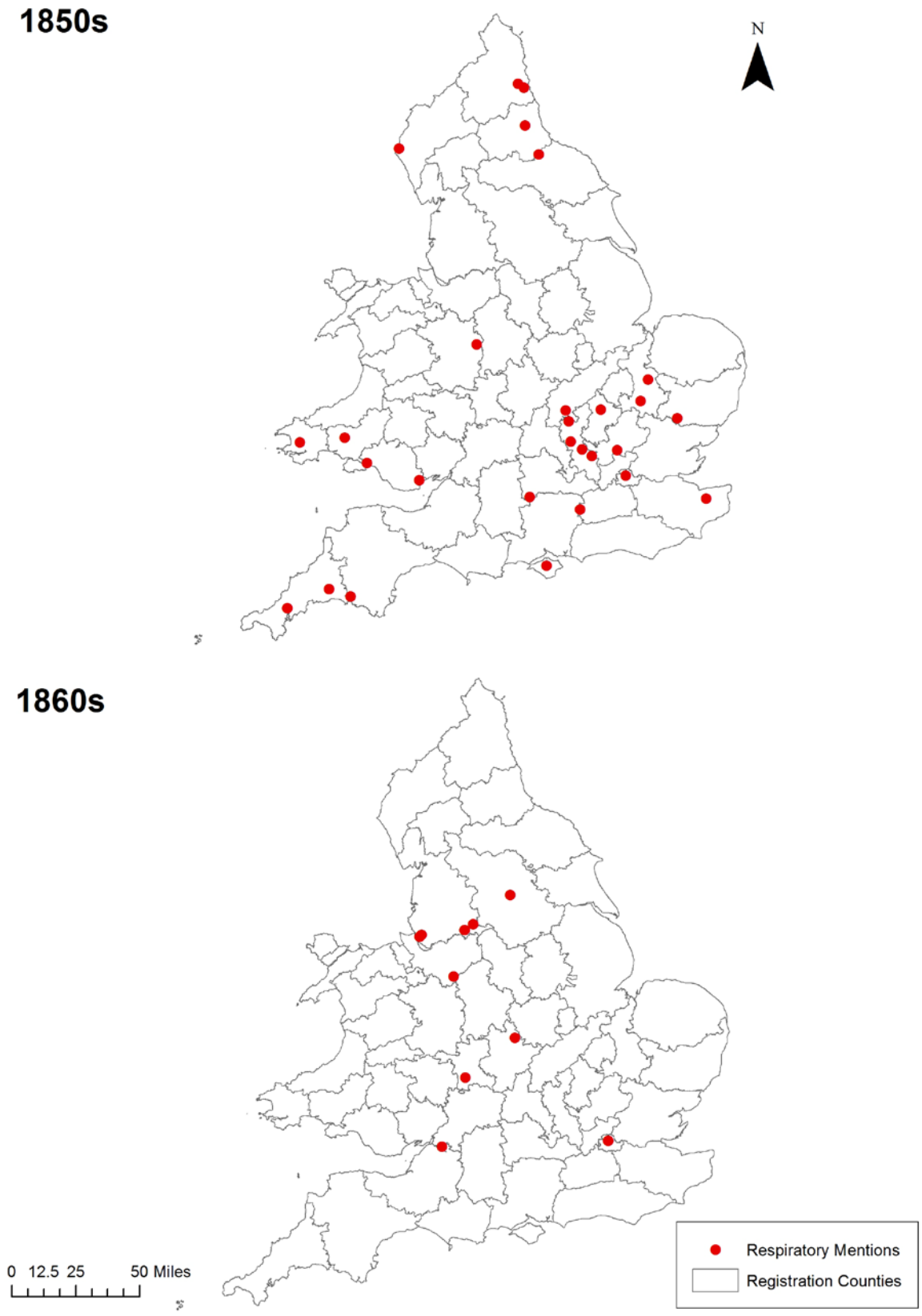

Figure 4: A temporal and spatial depiction of Registrar-General mentions of Respiratory diseases, 1850s and 1860s.

This temporal and spatial depiction of respiratory disease shows that the height of the Registrar-General's interest lay in the earlier decades, the 1850s and 1860s. There was also little or no mention of respiratory disease, other than in London, from the 1870s through to the 1900s. It therefore does not necessarily correspond with the rate of deaths from respiratory disease mentioned previously in Figure 1 which showed an increase in the rate of respiratory deaths up to and including the 1890s.

To further refine the geographical context to these mentions, and in order to highlight the places where the Registrar-General was most interested in respiratory disease, the count data shown in Figure 4 were subjected to a density smoothing process (Figure 5). This form of analysis highlights the areas of most significance in the data (1% significance). As the density smoothing process requires a minimum number of points to run, the latter decades with the fewest mentions show no significant polygons and the greatest concentration of mentions are again shown to be in the 1850s and 1860s. For the 1850s, this includes, firstly, a region bordering Cornwall and Devon, places that are not major urban areas. By revisiting the corpus it informs us that this is the result of the Registrar-General discussing an epidemic of influenza in Saint Agnes and Truro in Cornwall, of which the report stated: "Plymouth and the surrounding districts are still in an unsatisfactory sanitary state". Secondly, the large polygon to the north west of London including Bedfordshire, Buckinghamshire and Oxford, and including the places, Chesham, Wendover, Waddesdon, Leckhampstead, Towchester and Bedford, the reports mention a "higher mortality than average, chiefly from fever and bronchitis". The 1860s outputs are less surprising, producing smaller polygons of 1% significance that, as with previous analyses, focus respiratory discussion on urban centres such as Greater London and the Manchester area.

These significant polygons most likely result from a habit of the statistician Farr of summarising local Medical Officer of Health reports, and from this making places appear salient on the basis of his analysis of these local *texts*, rather than completing his own spatial analysis of *data* (though that is something he also did on a large scale). Since his very limited staff had to use pencil and paper for every calculation, this short cut is understandable, and indeed offers an insight into the effects of Victorian technologies on epidemiological methods (Higgs, 2004).

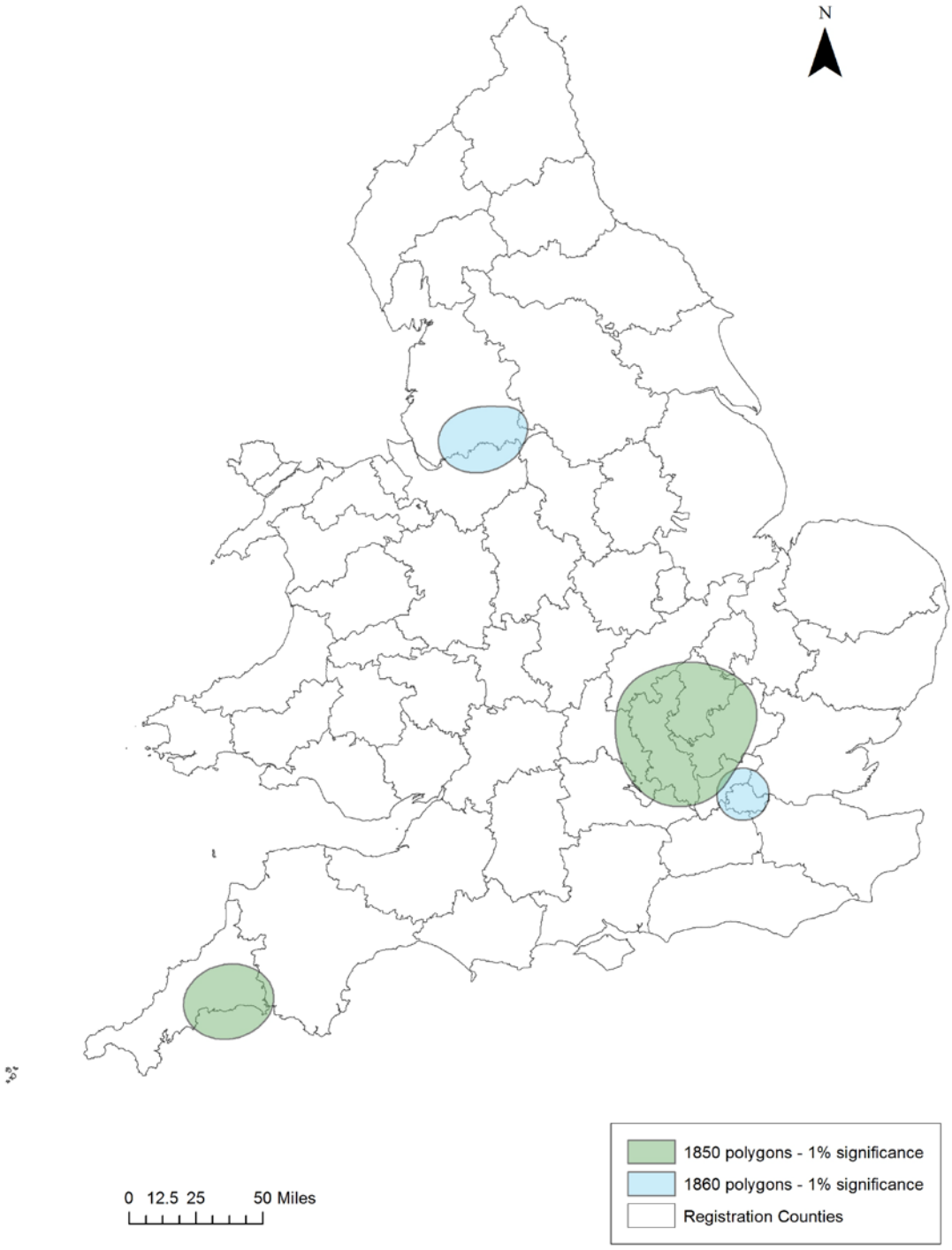

Figure 5: The output from the density smoothing procedure completed on the Respiratory mentions by the Registrar-General in the 1850s and 1860s. The data are calculated to 1% significance.

3.1 Comparing the geographies of instances and infant deaths

It is now clear where and when the Registrar-General discussions concentrated on respiratory disease but what is still unclear is how well this corresponds with actual infant deaths from respiratory causes during this time. To explore this further the IMR data discussed in Figure 1 is utilised again. Figure 1 illustrated that there was a clear rise in the rate of respiratory related infant deaths but it is also of interest to question where these deaths were occurring and whether those deaths relate to the same places where the Registrar-General was concentrating his discussion. In other words, was the GRO targeting discussion and highlighting those places in greatest need? To do this, the Hierarchical Regional Settlement matrix (HRS); a method based on that used by Gregory (2008), is employed as this better enables the exploration of the contrasting patterns of text and mortality between the rural and urban and between the core and peripheral regions of England and Wales. The HRS matrix classes Registration Districts using two measures. The first is a measure of 'urbanness' and is based on population density. The second measures peripherality according to the distance the Registration District is from London. These two measures are then brought together on a matrix such that cells in the top left are core urban places, ie. those in London. Cells in the bottom right are peripheral and rural and include places in Cornwall, Anglesey, the Lake District and Northumberland. Cells in the bottom left are rural places near to London, while those in the top-right are peripheral urban centres such as Liverpool, Manchester and Newcastle. As these places do not have high population densities as with Registration Districts in inner London, the top rows of cells away from London tend to be empty.

The Registrar-General's mentions of respiratory disease were mapped to the HRS matrices (Figure 6, left). From the outset it is clear that, as with the counts shown previously, spatial discussion focused on larger urban settlements such as Birmingham, Liverpool and Manchester, but primarily London, other than in the 1880s, a decade when no spatial discussion has been recorded. However, there was also focus on other places such as those in Bristol, Norwich and Sheffield, as well as other more rural and peripheral areas.

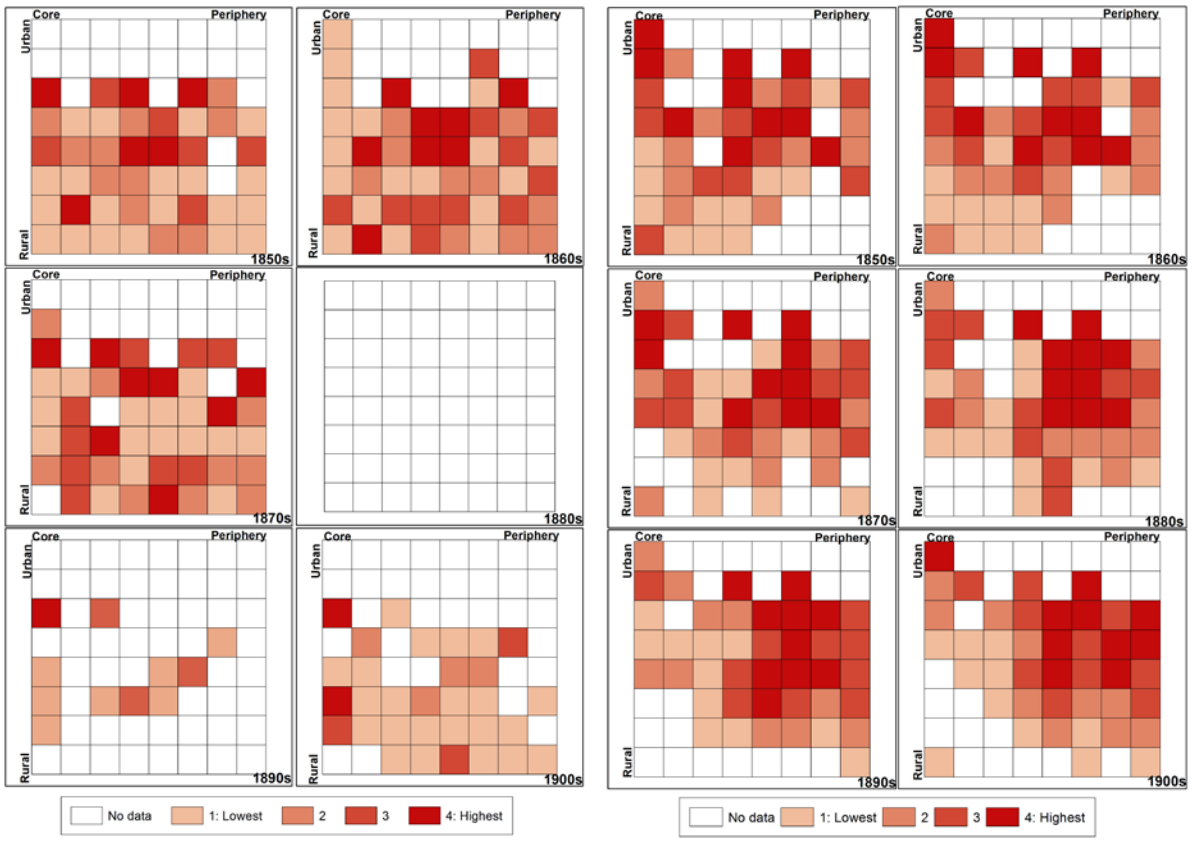

Figure 6: The Hierarchical Regional Settlement matrix (HRS) for the Respiratory category; the textual mentions by the Registrar-General (left) and the infant mortality (right). The least number of mentions/lowest IMR are shown in the lighter shading (1,2) and higher values symbolised using darker shading (3,4).

The IMR data were also mapped to the same HRS matrices (Figure 6, right) allowing for comparison with the Registrar-General mentions. The IMR data show high rates of infant mortality in the larger settlements such as London, but are more precise and actually indicate the east end of London (Registration Districts such as Whitechapel and Hackney) as having high levels of infant mortality in the 1850s, 1860s and 1900s, something which the Registrar-General discussions fail to highlight.

What these matrices offer is the first evidence of a possible poor correlation between the spatial distribution of the Registrar-General's mentions of disease in his official reports and the incidence of mortality from those diseases in the population. This relationship was further assessed by the use of regression analysis. The coefficient of determination (r^2) and the minimum and maximum residuals were derived in order to determine the strength of correlation and to pinpoint the greatest outliers in the datasets. For respiration this process

revealed for the relationship of total mentions versus infant mortality an r^2 value of 0.0004 providing statistical evidence of no meaningful correlation at a significance level of 0.01. The residuals derived from the same calculations showed the minimum values corresponding with Matrix ID 9, places in east London such as Whitechapel and Hackney, indicating they were mentioned far less than their mortality would predict, whilst the maximum residual corresponded with Matrix ID 17, other sections of London such as 'London City', Greenwich, Hampstead, Strand and Wandsworth. London is one of only a handful of places where the focus of the Registrar-General's discussion of disease coincided with a high infant death rates, however, the residuals illustrate that London had many more mentions than the death rate may predict.

3.2 Crowding and Food and Waterborne diseases

We now turn to the other two primary disease categories highlighted in Table 1. The Crowding and Food and Waterborne mentions were mapped temporally and spatially to allow for investigation of change in mentions over space and time. As with the Respiratory category, it is clear that the Registrar-General had great interest in the earlier decades of the study, the 1850s-1860s, when, for instance, he discussed crowding related diseases widely throughout England and Wales with little or no obvious focus. However, unlike respiration, this time frame extends to include the 1870s too. For both classifications the decade of the 1880s shows little interest in these disease groups from a spatial perspective, mirroring the spatial respiratory data described earlier (Figure 4) and coinciding with Higgs' 'age of inertia' at the GRO (Higgs, 2004, p, 90-128), the latter two decades illustrating a more focused concentration on larger settlements and the coal mining communities of middle England and south Wales.

Figure 7: The Registrar-General mentions related to Crowding (left) and Food and Waterborne diseases (right) mapped on the Registration Counties for England and Wales and displayed for each decade, 1850s-1900s.

These two disease categories were also subjected to the Hierarchical Regional Settlement matrices analyses based on Registration Districts. For Crowding (Figure 8, left), as noted in the respiratory analyses, the outputs present the 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 in the more core-rural and rural-peripheral sections of the matrix such as parts of Shropshire, Staffordshire, Devon, Northumberland, the Lake District, Anglesey and Cornwall, places where infant mortality was lower.

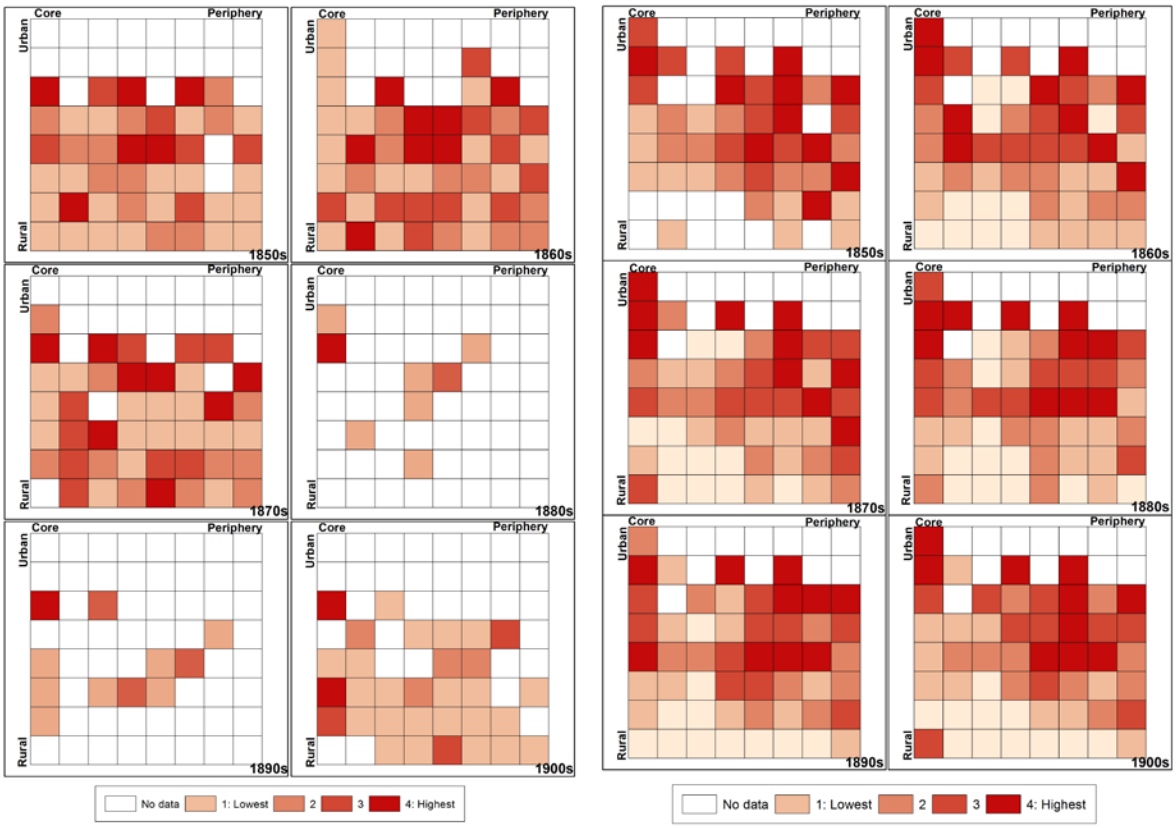

Figure 8: The Hierarchical Regional Settlement matrix (HRS) for Crowding; the textual mentions by the Registrar-General (left) and the infant mortality (right). The least number of mentions/lowest IMR are shown in the lighter shading (1,2) and higher values symbolised using darker shading (3,4).

From the corresponding IMR matrices (Figure 8, right) it is immediately clear that for Crowding the matrices frequently highlight different groups of Registration Districts to the Registrar-General mentions discussed previously (Figure 8, left), the majority of infant deaths being located in the core-urban and peripheral-urban regions. There is however some similarity in the matrices which include places like London, Birmingham, Liverpool, Manchester, Leeds and parts of Durham and Staffordshire, but many other places do not coincide, such as a large swathe of Registration Districts that make up the periphery of London, the Home Counties, as well as places in Northumberland, the Lake District, Anglesey and Cornwall. To add to this, there are also places with high infant deaths such as Bradford, Plymouth, Newcastle-upon-Tyne and parts of London, including the east end of the city, which the Registrar-General reports (Figure 8, left) do not consider specifically in any of the decades under analysis. Regression analysis revealed a correlation of 6% between the Crowding mentions and actual deaths confirming the visual patterns shown previously by the

matrices but also providing evidence that the Crowding mentions and the IMRs had the strongest correspondence out of the three disease groups analysed.

The Food and Waterborne matrices (Figure 9) show a similar pattern to the Crowding example with a high concentration of infant deaths in places the Registrar-General did not mention as much in his texts. Again, the discussion of disease was also more general with London being mentioned but areas within the capital, where infant deaths were high, not specifically referred to. The regressions confirmed this with an r^2 value of 0.0005. As such, it can be said that the recording of infant deaths was more spatially precise than any discussion of related diseases.

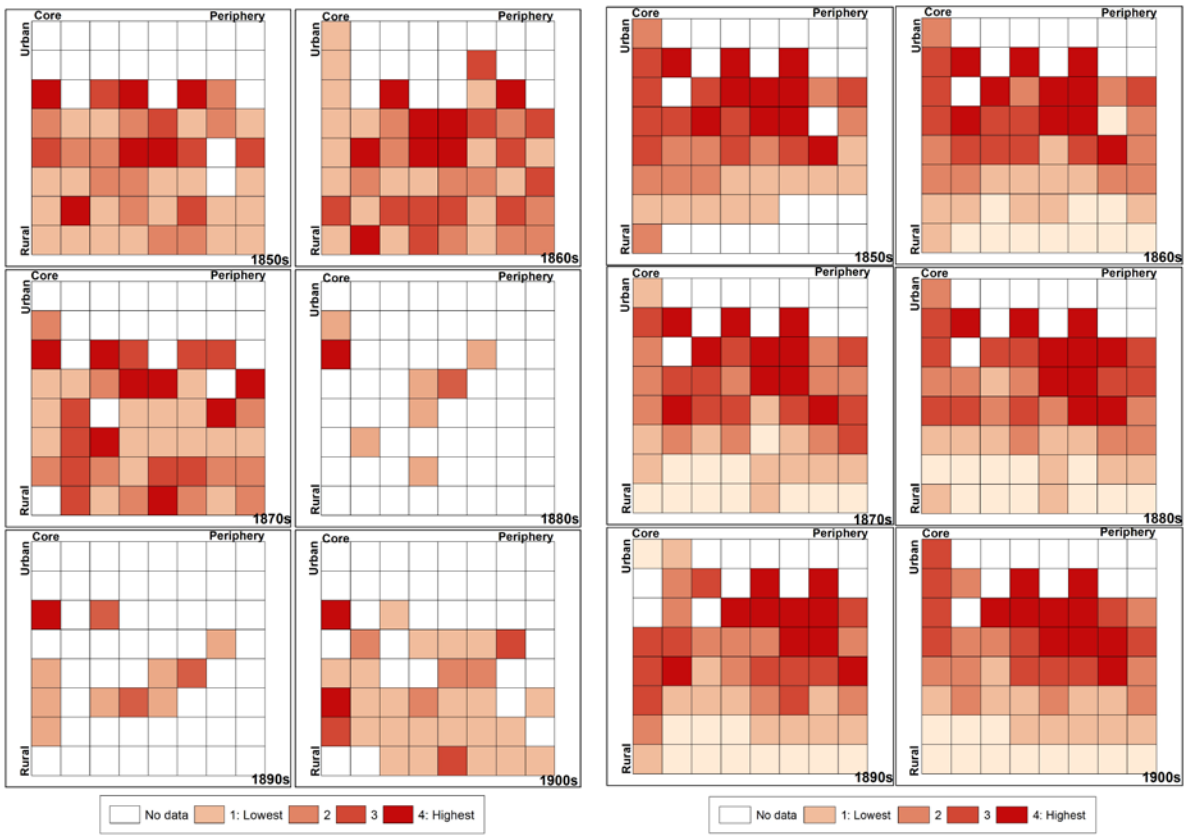

Figure 9: The Hierarchical Regional Settlement matrix (HRS) for Food and Waterborne diseases; the textual mentions by the Registrar-General (left) and the infant mortality (right). The least number of mentions/lowest IMR are shown in the lighter shading (1,2) and higher values symbolised using darker shading (3,4).

4.0 Discussion and conclusions

This novel approach, which has linked the quantitative analysis of text and statistical tables, leads to four principal findings. First, the authors' level of interest in the major diseases of infancy and childhood was fairly consistent. The exceptional circumstances of the 1866 cholera epidemic led to the publication of a special report which fully explains the otherwise startling 'spike' in mentions of this disease group in the 1860s (Figure 2). Apart from this, Figure 2 shows a picture of surprising consistency of attention across the careers of four Superintendent Statisticians and spanning a period of substantial change in the expert understanding of disease, not to mention significant changes in the burden of disease itself (Woods, 2000). Perhaps this is a sign of intellectual conservatism at the GRO: alternatively it could reflect a political need for published reports to address issues which their audiences considered the most important. It may have been the audiences rather than the Superintendent Statisticians who were the intellectual conservatives, though this label would fit George Graham's Victorian and Edwardian successors in the role of Registrar-General (Higgs, 2004, Szreter, 1991).

Second, attention to diseases paid little regard to the burdens of mortality from different disease groups or their change over time (again, the special cholera report, published 1868, is an exception). However, care is required to avoid anachronistic criticism of historic writers for failing to address a modern agenda: 'burdens of disease' is a late twentieth-century analytical concept (Murray and Lopez, 1996). Nevertheless, it is interesting that only nine expressions which even resemble this concept were recovered from the whole corpus, the earliest from 1887: the authors were simply not thinking this way. Farr, a passionate campaigner for disease prevention through these reports, based his appeals on the scandal of overall mortality and of 'preventable deaths' in a generic rather than a disease-specific sense, missing the analytic opportunity to show which preventive actions might prevent which disease-specific death tolls. This is of course partly a reflection of the preventive technologies available to contemporary health authorities, which were more often broad-spectrum, as with clean piped water, than disease-specific.

With the sole exception of the cholera report, the frequency of disease mentions in the reports did not correlate with the incidence of those diseases, as the neglect of respiratory disease shows. The GRO's primary emphasis on crowding and food and waterborne disease probably reflects the Chadwickian public health agenda of sanitary reform (Smith, 1979), where

medical men knew they could make a difference, whilst the GRO felt there was little anyone could do to prevent mortality from bronchitis and pneumonia, at least among the worse-off. In addition to this, another factor that may have promoted a focus on food and waterborne disease was the growing emphasis, from the 1880s, on preventing infant mortality by tackling diarrheal disease (Dwork, 1987).

Third, there was a tendency towards a less spatial approach to analysis, as the Geographical Text Analysis of disease instances reveals. Interestingly, this does not correlate with author as the trend began in the 1860s or 1870s while Farr served from 1839 until 1880. A likelier explanation is that when the GRO came under the control of the new Local Government Board in 1871 this blunted its enthusiasm for a campaigning approach, pointing out local authorities' weaknesses, and steered it into a more diplomatic relationship with local government (Higgs, 2004, p, 111-115). It was probably also true that as the GRO took on work of other kinds during this period and the inclusion of detailed spatial analysis in reports became more difficult. Indeed, Higgs shows how the size of the Annual Report shrank from about 50 pages before the mid-1870s to about 20 for the next 25 years, and the Superintendent's letter, which had been the principal place for detailed analysis, was completely subsumed into the Registrar-General's section of the Annual reports from 1879 to 1901 (2004, p, 54-55).

When looking at the concordance analysis, it is clear that as spatial accounts declined over time, they were replaced by non-spatial discussion. One example is Ogle's and Tatham's references to occupational lung disease, which make little use of geographical based discussion. This may represent the GRO concentrating on fields they felt they could influence more, via national legislation and the (national) Factory Inspectorate, at the expense of disease groups whose remedies lay with local government and now had to be managed using more 'distant' mechanisms.

Fourth, when there was an interest in geography the selection of places considered was not based on the spatial distribution of disease – any more than, as previously remarked, temporal trends in reporting reflected trends in disease burdens. Our regression analysis shows only a weak correlation between the spatial distributions of reporting and incidence of disease. Further work could usefully examine what 'place' meant to these authors: was it primarily a reference to an environment or shorthand for the 'type' of people who lived there, and did this

alter as the GRO responded to the arguments of the eugenic movement, who stressed heredity over environmental causes of disease? Overall, we have shown quantitatively that the Registrars General and their Superintendent Statisticians were tied to a fairly static view of which childhood diseases were important, paid little attention to changes in disease-specific mortality, paid declining attention to its spatial distribution, and were not concerned to write about those places where mortality was greatest. Their professional and political interests lay elsewhere.

In conclusion, the methodological aspects to this paper are as significant as the findings just stated. For one, Geographical Text Analysis may be used to further enhance our understanding of the geography of both historic and contemporary public health. For instance, the techniques may be used to compare the geographies emerging from texts with the known spatial distribution of events and could be extended to incorporate different classifications of disease as well as other population demographics of differing temporal origins. Applications might also include epidemiological studies (for example where official health surveillance data are not available) and the analysis of the places most important to public discourses about a disease. The identities of the authors of health related reports could also be explored in order to quantify their interests in different topics, including cause of death, and indeed possible geographical biases. However, public health analysis is only the first of many subject areas and research interests to which such a methodology might be applied. Alongside previous work on the digital analysis of text and geography (Gregory & Hardie, 2011; Cooper & Gregory, 2011, Murrieta-Flores et al, 2015) the present study points towards the great potential of these methods to deepen our understanding of a wide variety of historical and contemporary documents including newspapers, official reports, literature and internet sourced data.

References

Cooper, D. & Gregory, I.N., 2011. Mapping the English Lake District: A literary GIS. Transactions of the Institute of British geographers 36. 89-108.

Dwork, D., 1987. War is Good for Babies and Other Young Children: A History of the Infant and Child Welfare Movement in England, 1898-1918. Tavistock Publications, London.

Eyler, J.M., 1979. Victorian Social Medicine: The Ideas and Methods of William Farr. Johns Hopkins University Press, Baltimore.

Gregory, I.N., Bennett, C., Gilham, V.L. & Southall H.R., 2002. The Great Britain Historical GIS: From maps to changing human geography. The Cartographic Journal 39, 37-49**.**

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., & Hardie, A., 2011. Visual GISting: bringing together corpus linguistics and Geographical Information Systems. Literary and Linguistics Computing 26, (3), 297-314.

Gregory, I.N., Cooper, D., Hardie, A., and Rayson, P., 2015. Spatializing and analysing digital texts: corpora, GIS and places. In: Bodenhamer, D., Corrigan, J., Harris T. (Eds), Spatial Narratives and Deep Maps. Indiana University Press, Bloomington, pp. 150-178.

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.

Hardie, A., 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics 17, 380-409.

Hardy, A., 1993. The Epidemic Streets: infectious disease and the rise of preventive medicine, 1856-1900. University Press Oxford, Oxford.

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.

http://www.histpop.org

McEnery T., Hardie A., 2011. Corpus Linguistics: Method, theory and practice. Cambridge University Press, Cambridge.

Murrieta-Flores P., Baron A., Gregory, I.N., Hardie A., Rayson P., 2015. Automatically analysing large texts in a GIS environment: The Registrar General's reports and cholera in the nineteenth century. Transactions in GIS 19, 296-320. DOI: 10.1111/tgis.12106.

Murray, C.J.L., & Lopez, A.D., 1996. Evidence-Based Health Policy Lessons from the Global Burden of Disease Study. Science 274 (5288), 740-43.

Smith, F.B., 1979. The People's Health, 1830-1910. Croom Helm, London.

Szreter, S., 1991. The GRO and the Historians. Social History of Medicine 4(3), 401-414.

Szreter, S., 1991. The GRO and the Public-Health Movement in Britain, 1837-1914. Social History of Medicine 4(3), 435-63.

Titmuss, R., 1943 Birth, Poverty and Wealth: A Study of Infant Mortality. Hamish Hamilton, London.

Woods, R.I., Watterson, P.A. & 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. & 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. & 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.

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).