Health & Place 36 (2015) 25–34

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

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

Catherine Porter*, Paul Atkinson, Ian Gregory

Department of History, Lancaster University, Lancaster LA1 4YT, United Kingdom

ARTICLE INFO

Article history:
Received 27 March 2015
Received in revised form
6 August 2015
Accepted 28 August 2015

Keywords:
Geographical Text Analysis
Corpus linguistics
GIS
Infant mortality
Registrar-General

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.

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). 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, 1991a, 1991b; 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. 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

* Corresponding author.
E-mail addresses: c.porter2@lancaster.ac.uk (C. Porter), p.atkinson3@lancaster.ac.uk (P. Atkinson), i.gregory@lancaster.ac.uk (I. Gregory).
URL: http://www.lancaster.ac.uk/spatialhum/ (C. Porter).

http://dx.doi.org/10.1016/j.healthplace.2015.08.010
1353-8292/

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 GIS (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, 2004). 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, 11) 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 (1997) "*Atlas of Victorian Mortality*", 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, each of which were also taken into consideration in the creation of this scheme.

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

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		

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 (Fig. 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 practise. 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.

2.3. Texts on causes of death

As well as publishing tables of statistics on causes of death, the

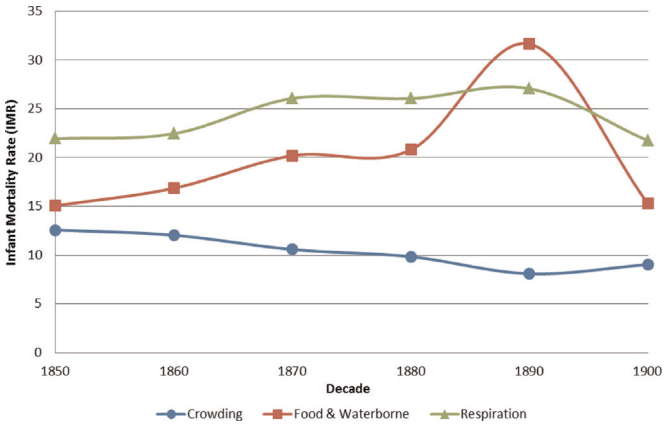

Fig. 1. Infant Mortality Rate (IMR) of each disease classification per decade of the study, 1850–1900 (Source: GBHGIS (Gregory et al., 2002)).