UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

WORKING PAPERS SERIES

Paper 149 - May 09

Family names as indicators of Britain's changing regional geography

ISSN 1467-1298

Centre for Advanced Spatial Analysis University College London 1 - 19 Torrington Place Gower St London WC1E 7HB
Tel: +44 (0)20 7679 1782 casa@ucl.ac.uk www.casa.ucl.ac.uk

FAMILY NAMES AS INDICATORS OF BRITAIN'S CHANGING REGIONAL GEOGRAPHY

James A. Cheshire, Pablo Mateos, and Paul A. Longley

Centre for Advanced Spatial Analysis and Department of Geography, University College London.

james.cheshire@ucl.ac.uk

www.jamescheshire.co.uk

Department of Geography, UCL, Pearson Building, Gower Street,

WC1E 6BT, London, UK.

7th October 2009

CONTENTS

"It may be thought by some that the investigation of the distribution of names is an idle amusement, productive of no utility of man. I have come to think, however...that it is a matter of much importance to the antiquarian, the historian the ethnologist and also to the more practical politician"

Henry Guppy, 1890:vi.

1. INTRODUCTION

A sense of regional identity remains important to the British population. Devolution, such as in Scotland and Wales, may be the most obvious means of enhancing regional identity, but there are many other manifestations of regional difference, such as the campaign for Cornish independence (BBC News, 2001), the "North/ South Divide" (Duranton and Monastiriotis, 2001), "Clone Town Britain" (nef, 2005), and different patterns of sports teams affiliations at the national, regional and local scales. There is much historical, linguistic, anecdotal and genealogical evidence for the existence of cultural and ancestral heartlands within Britain. However, much of the research into the nature of these regions has focussed on single events, serendipitous datasets, or specific regional case studies, without regard to robust measurement and comprehensive coverage across Britain. This study will attempt unearth many of the underlying population structures, real and imagined, in Britain by harnessing the wealth of data provided by family names.

The linking of Geographical Information Systems and Census information has created an unprecedented volume of geo-referenced data (Batty and Longley, 1996). Family names, or surnames, and the geographical locations of people who bear them, are frequently recorded in population registers such as the Electoral Roll or Health registers, and names recorded in the Census of Population are made public 100 years after collection. Many surnames can be used to infer the geography of a range of linguistic, historical, genetic, social or environmental characteristics about their bearers at the time of creation (Hey, 2000). Individuals with similar, or identical, surnames may share, or have shared, similar characteristics such as a common original geographical location (Hey, 2000). Despite the wealth of information now available, the study of names in Geography is still in its pioneering phase (Zelinksy, 1997). The importance of location in surname production and reproduction, combined with the genetic and cultural links over generations, outlined below, forms the fundamental premise of this work. This work seeks to create a regional geography of Britain based on the surnames of current and historical populations. No previous study has been attempted on this scale in Britain, although similar studies have been undertaken in other European countries at coarse levels of granularity (e.g. Colantonio et al., 2003). Much of the analysis is developed using an enhanced version of the British Electoral Register that identifies the names and locations of 45.6 million people in 2001. In addition, this work will provide a direct historical comparison with the regional geography of the surnames from the approximately 29 million people enumerated in the 1881 Census. The results presented show clear patterns of subpopulation structures within Britain, forming a strong basis for hypothesis generation relating to population dynamics and migration in future studies.

TABLE 1: A CATEGORISATION OF BRITISH SURNAMES. ADAPTED FROM BARKER ET AL., 2007.

Category	Example	Explanation
Occupational (Metonyms)		
Profession	Smith	Blacksmith/ metal worker
Office/ Trade	Reeve	Chief magistrate/ overseer
Rank/Status	Knight	A knighted person
Occupation Features	Falconer	One who kept/trained Falcons
Local Surnames (50% of surnames)		
Toponymic (from landscape)	Rivers	Dweller near river
Toponymic (from village/ region)	Cornwall	Man from Cornwall
Habitation (residence)	Gate	Habitation at/near a gate
Habitation (work)	Hall	A worker at the hall.
Surnames of Relationship		
From personal name (patronymic)	Johnson/ Jones	Son of John
From personal name (metronymic)	Margaretson	Son of Margaret
Personal name from other relative	Also: Johnson	Related to John
Personal name from diminutive	Dickens	Son of Dick (Richard)
Clan or tribal names	MacBain	Related to the MacBain clan.
Nicknames		
From animals	Fox	Slyness or other attributes
From characteristic traits	Careless	Free from care/ responsibility
From objects	Shorthose	Someone who wore short boots
From physical features	Little	A small person
From times and seasons	Pasque	Person born at Easter
From iconic description	Drinkwater	Heavy drinker

2. NAMES AND ORIGINS

Whilst it is unclear precisely when surnames became formalised and hereditary in Britain (Barker et al. 2007), there is agreement that the Doomsday book of 1085 made surnames a necessary, but not compulsory or hereditary, method of distinguishing individuals (Barker et al., 2007). The lack of any legal basis to surname adoption has led to the view that surnames were acquired gradually across the population. In the 13th Century surnames closely allied to locality were being regularly recorded (McClure, 1971); however, these were unlikely to be hereditary (McClure, 1979). By the 15th Century hereditary surnames became generally adopted in England (Lasker and Mascie-Taylor, 1985), but it was not until the 16th Century that the Scottish fully adopted them (Barker et al., 2007).

Fortunately, we have a much clearer understanding of the detail of surname naming conventions. As Table 1 demonstrates, surnames can be categorised into local surnames, occupational surnames, surnames of relationship, or nicknames (Barker et al., 2007). Several centuries may have passed since many contemporary surnames were created, but it is highly likely that the areas of conception remain the areas of highest concentration (Jobling, 2001). This is important as it points to enduring social and genetic commonalities within these populations.

2.1. Surnames in Human Population Biology

Surnames are a useful geographical data source since they are ascribed to unique individuals or households and are recorded in diverse and sometimes easily accessible population registers. Surname registers are a relevant data resource in that geographical distributions of surnames have been shown to closely match gene frequency distributions (Mascie-Taylor and Lasker, 1990). As such, they have facilitated a number of studies within population biology over the last century or so.

George Darwin, son of Charles Darwin, initiated the use of surnames to investigate family lineage in 1875. He was interested in the frequency of first cousin marriages and whether their offspring experienced any adverse health effects as a result of this consanguinity (Darwin, 1875). Darwin's and subsequent studies have taken marriages to be consanguineous if they were isonymous. Isonymy, in this context, can be defined as the presence of identical surnames in the ancestors of a couple (Lasker 1968).

Surname studies within genetics and more widely in human biology are based on the principle that to the extent that two individuals with the same surname are ultimately to share the same lineage, isonymy indicates biological relatedness (Lasker, 1985). The hereditary nature of surnames and their tendency to remain highly concentrated in their areas of origin are the two traits most utilised by geneticists and population biologists. Hereditary surnames contain information about relatedness within populations because patrilineal surnames should correlate with the Y chromosome inherited from a male's father (Sykes and Irven, 2000). This relationship depends on the assumption that the founding population was small, genetically diverse and comprised of families with unique surnames (Rogers, 1991). Historical evidence suggests this is unlikely as most founding populations are characterised by small, often familial, groups originating from the same region. These groups were likely to share a small gene pool and exhibit high levels of isonymy (Jobling, 2001). It is acknowledged that for these reasons some tolerance is required when using surnames to make genetic inferences (Lasker, 2002). However, the impracticality of collecting the genetic information from a complete population, past or present, makes proxy data, such as surnames, the only alternative in large-scale studies. Additionally, studies of extinct lineages have shown that many lines of descent quickly disappear so that the remaining individuals are much more likely to be related through a common ancestor (Lasker, 2002).

The availability of population registers in digital form, combined with a maturation of methods has led to a thousand-fold increase in the published use of surnames in population biology (Colantonio et al., 2003). A major breakthrough in the effective utilisation of surnames in genetics was made by Crow and Mange (1965), who formalised a Coefficient of Inbreeding from Isonymy (Crow, 1979) (Equation 1). Lasker (1977) advanced this measure by developing the Coefficient of Relationship by Isonymy (R_i) (Equation 2), that was later extended to the Lasker Distance (Rodriguez-Larralde et al. 1994). This measure, outlined in more detail below, forms the basis for many comparative studies of regions and their surnames (Colantonio et al., 2003).

3. Surnames, Regions, and Geography

In spite of the inherently spatial patterning of surnames, studies of them have been rare in the geography literature (Zelinsky, 1997). One of the earliest, and most

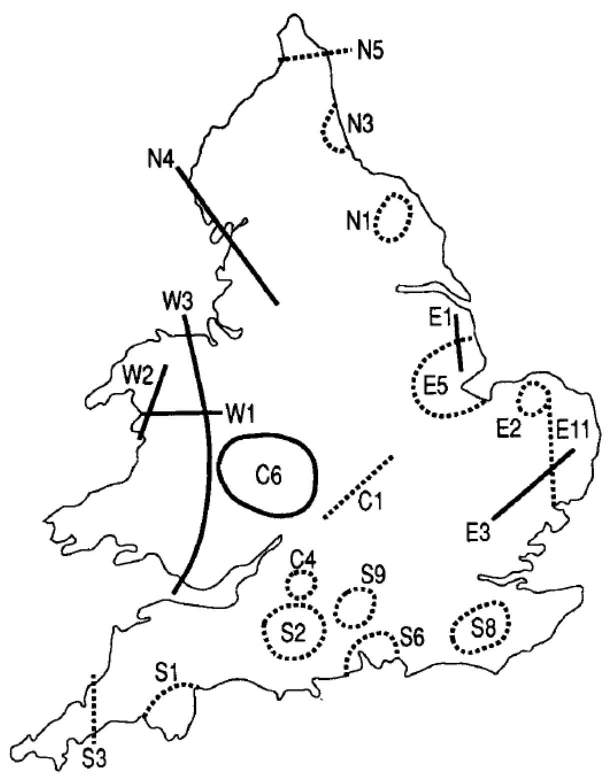


FIGURE 1: SURNAME-FREQUENCY BOUNDARIES DETERMINED BY THE WOMBLING PROCEDURE. THE SURNAME BOUNDARIES OBSERVED IN SOKAL ET AL.'S OVERALL ANALYSIS AND THOSE PRODUCED FROM INDIVIDUAL ANALYSES ARE ABSTRACTED AS THICK SOLID AND DASHED LINES RESPECTIVELY. SOURCE: SOKAL ET AL., 1989, PAGE 467.

thorough, attempts to define surname regions was undertaken by Guppy (1890) in his book "Homes of Family Names in Great Britain". Many of the issues, such as whether the Welsh border defines the extents of Welsh communities, or whether Parliamentary areas "are not political or artificial" in their determination (Guppy, 1880), remain important today. Table 2 contains the surname categorization developed by Guppy:

TABLE 2: GUPPY'S CLASSIFICATION OF BRITISH NAMES. THESE CATEGORIES ARE STILL APPLICABLE TO MANY CONTEMPORARY SURNAMES. SOURCE: GUPPY, 1880. PAGE 11.

Classification	Occurrence
General Names	30- 40 Counties
Common Names	20- 29 Counties
Regional Names	10 - 19 Counties
District Names	4- 9 Counties
County Names	2 – 3 Counties (principle home in one of them)
Peculiar Names	1 County (and generally to a specific parish/ division within it.)

From his classification Guppy established that clear regions existed with, for example, South West England's inhabitants possessing 40% of all 'peculiar' (Table 2) names. Inspired by the regions of Anglo Saxon Britain, Guppy suggested that the regionality of surnames could be sufficient to restore "the heptarchy to our land".

More recently, Zelinsky (1970) used forenames as a data source to investigate cultural variation across 16 counties in the Eastern United States but no further work appears in the geographical literature until Porteous (1982). This study suggests a multi-operational method for investigating the spatial origins and subsequent diffusion of rarer English surnames at a national and regional scale (Porteous, 1982). Despite Porteous' assertion that "names have been neglected by geographers" (Porteous, 1982 , P395), and his attempt to reintroduce surname studies to geography, the article failed to stir much interest. Zelinsky (1997) also unsuccessfully encouraged geographers to use people's names in the study of population and regions. A more recent study, published in the Annals of American Geographers, was Longley et al. (2007).

The dearth of name studies within geography has been countered by the growing number of spatial studies from Human Biology and linguistics. Studies such as "The Present Distributions of Some English Surnames Derived from Place Names" (Kaplan and Lasker, 1983) and "Geographical distribution of Common Surnames in England and Wales" (Mascie-Taylor and Lasker, 1984) were published in *Human*

Biology and the *Annals of Human Biology* respectively. This demonstrates the emphasis on the genetic links to surnames. Whilst many of the practitioners of this research incorporate sound geographical analysis in the studies, there is a lack of breadth and critique from a geographical perspective.

3.1. Mapping Surname Regions

Studies of the geographical distribution of surnames, predominantly by investigating isonymy, have been undertaken for the population of several countries. These cover the populations of Switzerland, Italy, Germany, England plus Wales, Scotland, Austria, the Netherlands, Venezuela, and the United States (see Colantonio et al., 2004 for a literature review). Further studies of Western Europe (Scapoli et al., 2006), the Azores (Branco and Mota-Vieira, 2003, 2005), Belgium (Barrai et al., 2003), Argentina (Dipierri et al., 2005), Spain, (Rodriguez-Larralde et al., 2003) and Siberia (Tarskaia et al., 2009) have also been undertaken. These national and regional studies, with the exception of Scotland, demonstrate the effect of geographic distance on the patterns discernable from surname frequency distribution data. The studies employ similar methods of analysis and visualization.

In England, Kaplan and Lasker (1983) found almost twice the expected number of surnames located in areas sharing their namesake (for example Baths from Bath). Although some of the surnames (taken from 1981 English telephone directories) only partially originated from the studied areas, a tendency of association appeared to remain, despite the long period since surname establishment (Lasker and Kaplan, 1983). Moreover, their findings support the claim that places closer together have an increased likelihood of commonality of surnames (Lasker and Kaplan, 1983). This observation conforms to Tobler's First Law of Geography (Tobler, 1970).

The appendix to Lasker's (1985) book "Surnames and Genetic Structure" contains maps and diagrams for 100 surnames in England and Wales. This represents one of the first attempts to comprehensively map and compare the distributions of English/Welsh surnames: the maps depict surname frequency alongside plots of the probability of local excess/ deficiency from north to south and east to west (Mascie-Taylor et al., 1985). In addition, these maps were used as an approximate comparison to the descriptions provided by Guppy (1890). Access to the 1881 Census places this study in a more fortunate position, as it is able to make more accurate quantitative comparisons between contemporary surname distributions and those in the 19th Century. The maps in Mascie-Taylor et al. (1985) demonstrated the utility of representing surname frequencies spatially, while other publications such as the "Atlas of British Surnames" (Lasker and Mascie-Taylor, 1990) and more recently "An Atlas of English Surnames" (Barker et al., 2007) have continued in this vein. The most recent developments in mapping name frequencies originated from UCL Department of Geography and Centre for Advanced Spatial Analysis with the creation of the Great Britain Surname Profiler (http://www.nationaltrustnames.org.uk/) and the WorldNames Profiler (http://www.publicprofiler.org/worldnames/). These websites enable anyone with Internet access to produce personal name maps as well as of groups of name origins. In addition, the former site includes 1881 Census of Great Britain data, from which users can compare surname distributions between this date and 1998. Worldnames includes data from 26 countries showing users the distributions of their names in a selection of countries around the world.

Whilst useful from a genealogical perspective, simple choropleth maps portraying individual surnames are of little use when characterizing regions or groups of names. Sokal et al. (1992) undertake surface analysis on 100 surnames in England and Wales. Amongst other analyses, they undertook surface wombling (see Barbujani et al., 1989) to produce the surname frequency boundaries shown in Figure 1. This confirms the existence of surname regions within Great Britain and was one of the first studies to quantitatively do so through a variety of surname frequency aggregation procedures. It also introduced the idea of discrete regions that can be distinguished through drawing boundaries at points of abrupt change in population surname structure. This concept is well known to geneticists (for example Barbujani and Sokal, 1990).

The notion of abrupt changes in population surname structure was extended to suggest that these areas represent barriers to population gene flow (and therefore surname flow). Monmonier's Barrier Algorithm has been also used to represent such barriers (Manni and Barrai, 2001, Manni et al., 2004, Manni et al., 2008). The only published examples of the application of Monmonier's Algorithm to surname data apply to Italy (Manni and Barrai, 2001) and the Netherlands (Manni et al., 2004, 2008). More analysis is therefore required on the appropriateness of this work in the context of surname studies, not least because it holds much potential in identify the effects of topography (such as high mountains), for example, on the movement and mixing of populations.

The use of distorted geographical maps as a means of mapping surname distributions has been suggested by Mourrieras et al. (1995). This novel technique uses the distance matrix provided by the Coefficient of Isonymy to distort the outline of France around 90 reference points placed relative to each other in two dimensional space according to observed similarities in their Lasker Distance measures (Mourrieras et al., 1995). The magnitude of displacement of each reference point from its geographically correct position to its new position according to the isonymy values is represented by isolines linking points of equal displacement intensity. Grey-level shading along these isolines facilitates the segmentation of the geographical and surname maps into 'homogenous surname zones' (Mourrieras et al., 1995). The distorted maps are challenging to interpret and understand, especially for the uninitiated. The results produced for isonymy in France by Scapoli et al. (2005) are simpler to interpret and also demonstrate the clear relationship between surnames and dialects within the country.

3.2. Regions in Geography

The lack of research undertaken by geographers (as opposed to linguists, geneticists and historians) into surname, linguistic, and genetic regions has left the methodologies employed lacking context from the long tradition of debate, revisited here, that surrounds regional studies in geography.

The term *region* is used in a variety of ways to denote "spatial compartments" of formal, functional, or perceptual significance (Murphy, 1991). Massey (1995) defines it simply as a distinct area on the earth's surface.

In their classic paper, Brown and Holmes (1971) classify regions as either functional or uniform. The former is composed of areas that have more interaction within each other than with outside areas (Brown and Holmes, 1971). In the context of surnames, interactions could include the movement of one individual to marry another from a different area of origin. The datasets in this study provide only two snapshots of the population. The inherently dynamic nature of functional regions renders the cross-sectional 1881 and 2001 datasets in their present state inadequate longitudinal study. However, a subtle, but distinctive, regional geography exists for naming conventions that are suggestive of function, for example 'industrial' versus 'agricultural' names. On this basis, it is possible for names to suggest historic regional functions that could indicate a geographical reconfiguration of names into the functional regions of today. By assessing the degree of mixing between names, a more obvious use emerges; that is to establish a region's level of integration into the national and international economy. Supplementary longitudinal investigations of population interaction between areas could, for example, come from apprenticeship records from multiple years (for example, Patten (1976)).

The quantitative paradigm in Geography of the 1960s and 1970s paid scant attention to the historical and geographical variability of regional development or to the genealogy of regional formations (MacLeod and Jones, 2001). These approaches to spatial science were heavily criticised throughout the 1970s, prompting many geographers to turn to more theoretical disciplines for insights into spatial patterns

(Pudup, 1988); many of these allied to the radical political and intellectual climate of Western Europe and North America at the time (MacLeod and Jones, 2001). A parallel development was the move towards a variety of approaches centred on humanism. Cloke et al. (1991) assert that geography, and therefore the study of regions, was becoming increasingly irrelevant because spatial scientists failed to take seriously the complexity of human beings. This new approach registered a deeper concern with the

"social construction of places and with experiential meanings, interpretations, and emotional repertoires of human subjects- not least those relating to their surrounding environment, sense of place, lifeworld, and attachments to their place of dwelling" (MacLeod and Jones, 2001: 673).

Many of these concerns, such as relationship with the surrounding environment, relate closely to the inspirations behind surname formation. Humanistic approaches therefore have their place in regional research surrounding surnames- especially at the local scale. The new regional geography of the 1980s as outlined by Gilbert (1988) provides the following classification of regions:

- A local response to capitalist process.
- A focus of identification.
- A medium for social interaction.

(Gilbert, 1988: 209-213)

The first distinction is arguably the most influential and originates from much of the quantitative work of the 1960s and 1970s, especially with reference to functional regions. The latter two are most relevant here as they refer to the processes, outlined earlier, that contributed to surname creation. It should be noted that the 1881 data used by this study is likely to be of an enduring geography as there appears to have been relatively limited population movement before this occurred (Guppy, 1890). Following from the work of Gilbert and others in the 1980s, there have been calls for regional studies to become a central component of the whole of Geography and not treated as a sub-discipline (see Johnston, 1991 and Thrift, 1994). According to McLeod and Jones (2001), the most recent incarnation of regional study should have both the regional formations as objects of analysis; thus bestowing on the researcher an "ontological coherence" to engage in a serious attempt to make sense of "this world of intellectual disorientation". This aligns well with Murphy's call for the nature, extent, and character of the regions examined in empirical studies, to become part of our conceptualization of social processes that take place in those regions (Murphy, 1991). This approach also requires a social theory that does not treat regional settings as unsubstantiated abstractions or a priori spatial givens, instead treating them as the results of social processes that reflect the shape and ideas about the organisation of the world (Murphy, 1991). Murphy's call fits well with the potential of surnames to illustrate those precise social processes that shape regions, rather than a spatial given or abstraction.

The contemporary debates between critical and quantitative geography have been distilled into the Focus section of a recent *Professional Geographer* edited by Kwan and Schwanen (2009). In this Barnes argues that the binary between critical and quantitative geography emerged, in part, from an obligation felt by critical geographers to "excise everything that went before" (Barnes, 2009). Kwan and Schwanan's (2009) reflection that many quantitative geographers are concerned with "critically inspired" issues, such as segregation, health disparities and income inequalities (topics not too far removed from this work), but are critiqued on the grounds of undertaking abstract mathematical theorization is fair one. In addition the increasingly data rich nature of contemporary research has reduced the level of abstraction from reality that characterised many earlier quantitative studies.

The quantitative approach to the study of regions taken by this research undoubtedly suffers from some of the limitations outlined above. However, much debate has surrounded the explanation and analysis of regions that have already

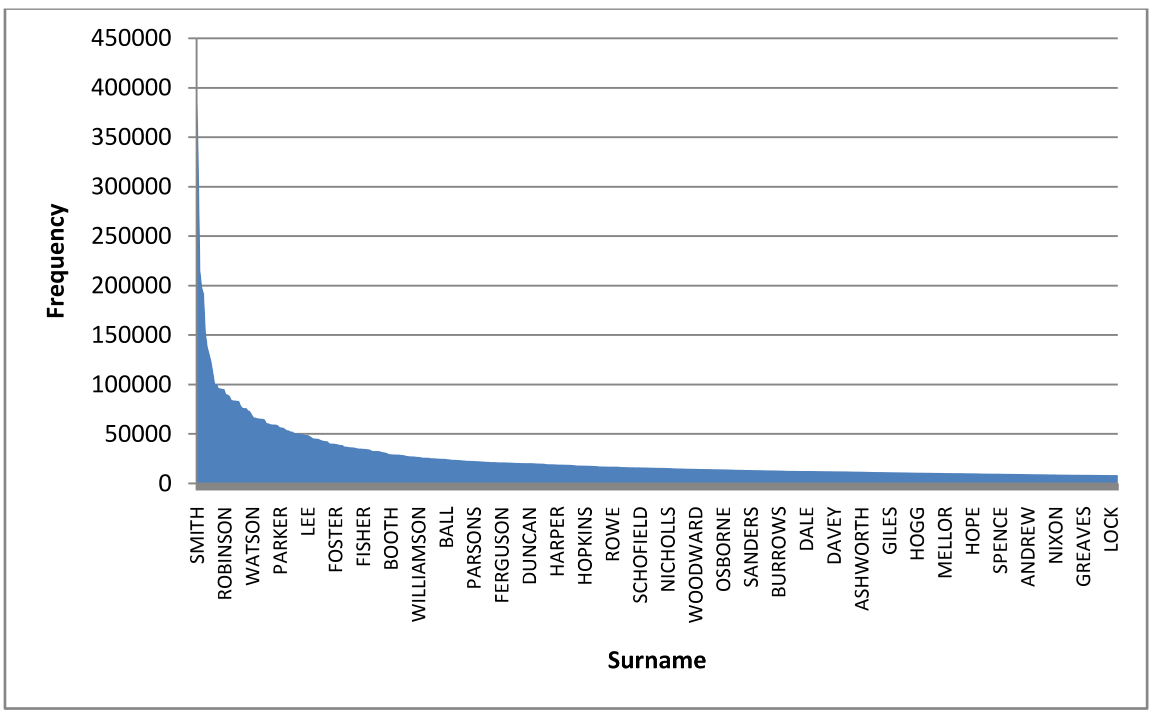


FIGURE 2: A PLOT SHOWING THE POPULATION OF EACH SURNAME (X AXIS) AGAINST THE TOP 500 SURNAMES IN BRITAIN FOR 1881 (Y AXIS). IT IS CLEAR THAT OF THE 425,793 SURNAMES IN BRITAIN THE MAJORITY HAVE LOW FREQUENCIES COMPARED TO THE MOST POPULAR 100. THIS CREATES AN EXTREMELY LONG-TAILED DISTRIBUTION.

been identified either through study, legislation (such as administrative borders), or well-known social discourse (such the North/ South Divide). Although the methods cannot be entirely absolved from the critiques of spatial science and quantitative geography from the 1960s to present day, an alternative approach remains to be found that is able to deal with large volumes of data. It is clear from the Critical Quantitative Geographies edition of *Professional Geographer* that geographers should be pragmatic when applying their critical and/or quantitative methods. Quantitative methods can only begin to be used on quantifiable attributes; they are not, for example, capable of representing complex human experiences or social realities (Kwan and Schwanen, 2009). If however the stated intention is the depiction of generalised trends from large datasets then quantitative methods are extremely appropriate. Pooley and Turnbull (1998) argue that the refocusing away from "mechanistic and quantitative" approaches to those better suited to identifying processes "of social and cultural change affecting both individuals and communities" has been detrimental to generalisations. This is because the atypical aspects of migration (and therefore the processes surrounding region building) have dominated at the expense of "the everyday and commonplace dimensions of population movement" (Pooley and Turnbull, 1998. P330).

This section has sought to demonstrate the importance and relevance of regional research within geography and the growing interest of surname regionalization by human biologists and geneticists. The latter would benefit from closer interaction with the former in order to improve the quality of geographical analysis, visualization and regionalization methods employed. The genetic focus of previous research has overlooked many theoretical considerations familiar to geographers. Aside from the theoretical, there are many important practical contributions from geographers and spatial scientists to be made to the study of surnames.

4. Research Aims

To date, no study has attempted a comparative study surname regions in Great Britain between 1881 and 2001. The intention here is to create a generalized perspective on the persistence, or otherwise, of surname regions between the 19th and 21st centuries by examining the coherence of 'what was' and looking at 'what is' to evaluate the extent to which previous patterns have changed.

In addition, unlike other spatial surname studies, the largest available datasets containing 29 million and 45.6 million individuals respectively are used. Previous research has focused on smaller geographic areas or sampled groups of names. The resulting methodological framework will be applicable at a range of spatial and temporal scales and spaces, assuming the availability of appropriate data. This study's intentions move away from genetics to investigate physical, political and social regions. On this basis the role of surnames can evaluated in relation to the notions of functional, uniform and perceived regions in Great Britain.

5. Data Sources and their Geographic Integration

5.1. The 1881 Census

Returns from the 1881 census are preserved for England, Scotland and Wales. The data provide the names and place of enumeration (Parish and Registration District) for 29 million people, with a total of 425,000 unique surnames (approximately 49,000 of which have occurrences of more than 20 people, see Figure 2). When digitising the census records, volunteers from the Church of the Latter Day Saints reproduced surnames exactly as transcribed on the original with the following exceptions: double barrelled names had dashes removed, spellings with unusual punctuation were excluded, spaces in Mc and Mac names have been removed and

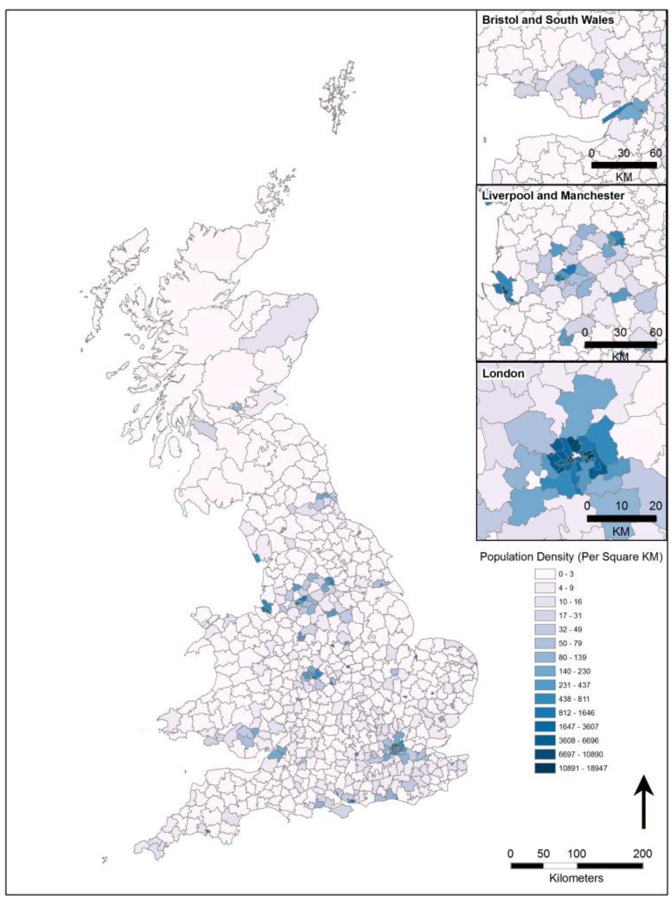


FIGURE 3: A MAP SHOWING THE POPULATION DENSITY OF EACH 1881 CENSUS REGISTRATION DISTRICT. AS CAN BE SEEN MOST DISTRICTS HAD A LOW POPULATION DENSITY, WITH ONLY A FEW URBAN DISTRICTS POSSESSING HIGH POPULATIONS. SOURCE: BOUNDARY DATA UK BORDERS.

those names only surviving as initials or only containing two letters were removed (Barker et al., 2007).

There is likely to be human error in the digitising process and a largely illiterate population in 1881 would have forced census enumerators to interpret verbal information (Barker et al., 2007). The data and documentation are available from the UK Data Archive (2000).

The geography of the 1881 census is complex due to confusion over some of the administrative boundaries used. Indeed, the census report states that the boundaries used "overlap and intersect each other with such complexity that enumerators and local registrars in a vast number of cases failed altogether to unravel their intricacy" (Census of England and Wales, 1881. In Woolland and Allen, 1999: P49). From the available boundaries, it was thought sensible to use registration districts, as opposed to parishes or counties in this study. Registration districts are much less coarse than counties but coarser than parishes and provide the best balance between spatial resolution and a sufficient population size to obtain a representative population of surnames within each geographical unit of analysis. Analyzing registration districts also makes pragmatic sense as their boundaries have been digitized and are available for download from the UK Borders website (http://edina.ac.uk/ukborders/). It should be noted that if an individual's

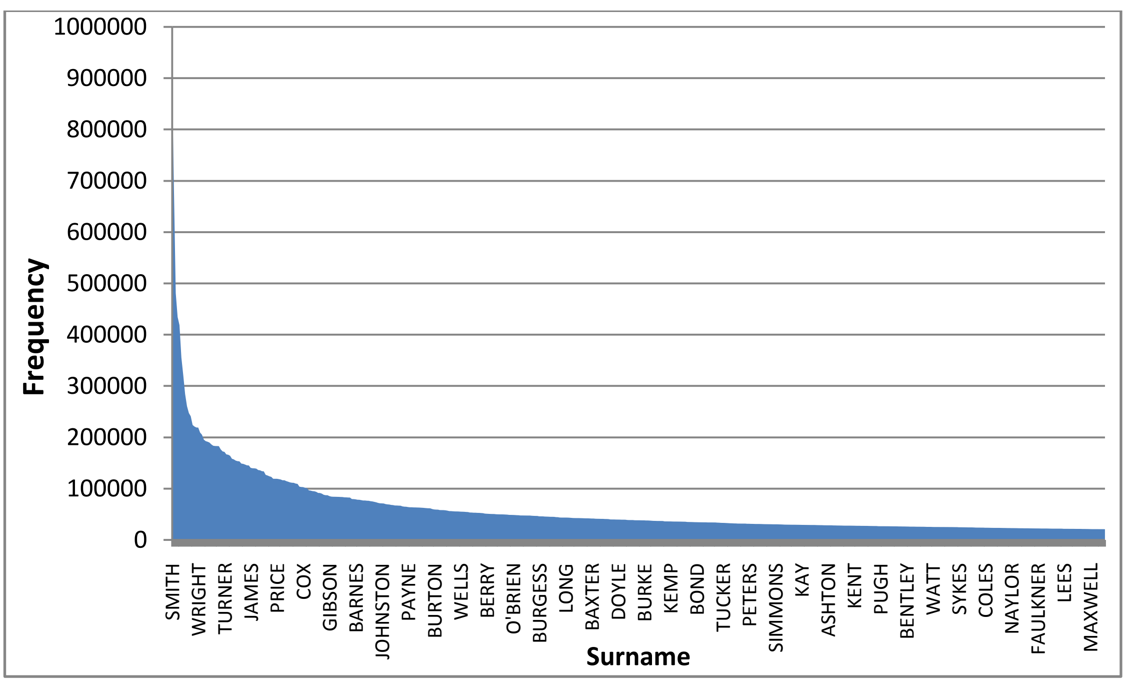


FIGURE 4: A PLOT SHOWING THE POPULATION OF EACH SURNAME (X AXIS) AGAINST THE TOP 500 SURNAMES IN BRITAIN (Y AXIS) IN 2001. IT IS CLEAR THAT OF THE 1,597, 805 SURNAMES IN BRITAIN THE MAJORITY HAVE LOW FREQUENCIES COMPARED TO THE MOST POPULAR 100. THIS CREATES AN EXTREMELY LONG TAILED DISTRIBUTION.

parish straddled a registration county they may have been registered in both. The inconsistent policy of the database creators towards this issue makes the numbers of people duplicated hard to quantify (Wooland and Allen, 1999).

It is acknowledged that the limitations of the 1881 census are much greater than those of the 2001 electoral roll. However, they are not considered sufficient to undermine the utility of the comparison between time periods. Interpretation of the results from the 1881 data will be tempered by an awareness of the limitations. The generalized perspective, and preliminary nature of this study limits the space and validity of a comprehensive appraisal of the recording of 1881 census geographies as this can be covered in further work.

In this study 662 Registration Districts (EDs) were mapped; of which 658 have surname data, with the remaining classified as common land or missing data. These latter districts were removed by enlarging the neighboring districts contiguous with them. The average population in each district is approximately 4900 inhabitants. Figure 3 shows a population density map in 1881 by Registration District.

5.2. The Enhanced 2001 Electoral Roll

The contemporary surname frequencies used in this project come from the enhanced 2001 UK Electoral Register purchased from the company CACI (London, UK). This dataset includes the names and addresses of UK residents aged 17 or over who are (or are about to become) eligible to vote in UK or European elections. This is enhanced by data, sourced from commercial surveys and credit scoring databases, on individuals not registered to vote or who opted out of the public register. The data represent 45.6 million people resident in the UK in October 2001, with a total of 1,597, 805 surnames (see Figure 4). The British, not UK, focus of this study means that only those resident in Britain are analysed from this dataset.

The 2001 enhanced Electoral Register records can be aggregated to unit postcodes that can in turn be easily linked to the 2001 Census of Population geography using the National Statistics Postcode Directory (NSPD) (available from http://www.ons.gov.uk/). From each unit postcode, the data may be aggregated to one or several of the following 2001 Census administrative boundaries available (from smaller to larger areas): Output Area (OA), Lower Super Output Area (LSOA), Middle Super Output Area (MSOA), Super Output area (SOA), Local Authority District, or Government Office Region (GOR). A balance needs to be struck between computational time, data storage and handling, sufficient populations within each unit (to avoid the small number problem) and units of similar size for reasonable comparison with the 1881 dataset. With these considerations in mind, Local Authority District level units were considered the best level of geography to use. The

District represents an administrative area corresponding to the Local Authority level in the hierarchy of the UK local government. There are 410 Districts Great Britain, including 354 in England (32 of which are London), 22 in Wales, and 34 in Scotland with an average population of approximately 105,000 inhabitants. Figure 5 shows a population density map using the 2001 Electoral Register by District.

Initial calculation of the Lasker Distance and mapping of the clustered results, as described below, produced highly fragmented results for the 2001 dataset, caused by the atypical composition of surnames in the 32 London districts. London districts, as part of a long established global city and centre for immigration, contain an atypical surname composition with, for example, the highest numbers of unique surnames when compared with the rest of Britain (McElduff et al. 2008) (see Figure 6). Aggregating the 32 London districts into a single district created more stable and plausible regions. In the final analysis, therefore the Lasker distance was calculated for 379 districts.

6. METHODS AND THEIR THEORETICAL FOUNDATIONS

6.1. COEFFICIENT OF RELATIONSHIPS BY ISONYMY AND LASKER DISTANCE

On the premise that the likelihood of a gene being shared by first-degree relatives is one in two, Crow and Mange (1965) proposed the Coefficient of Relationship by Isonymy (*Ri*) to be half the proportion of isonymy:

$$R_i = \frac{\sum p_i q_i}{2} \tag{1}$$

where p_i is the frequency of *i*th surname in fathers and q_i is the frequency of the same surname as the maiden name of mothers. The Lasker coefficient of isonymy is widely used for surname studies and extends the idea of monophyly (sharing a single common ancestor) between two populations. Lasker (1985) defines the measure as:

"The probability of members of two populations or subpopulations having genes in common by descent as estimated from sharing the same surnames" (Lasker, 1985:142).

It is calculated as:

$$R_i = \frac{\sum (S_{i_1} S_{i2})}{2 \sum S_{i1} \sum S_{i2}} \tag{2}$$

where S_{i1} is the number of occurrences of the ith surname in a sample from Area 1 and S_{i2} is the number of occurrences from the same surname from Area 2 (Lasker, 1985). The resulting value can be considered as the proportional correspondence in terms of a shared surname pool between a particular place and all others in the country (Schürer 2004).

Whilst the Lasker Coefficient of Isonymy remains a dominant measure in surname research, it has been extended to create a distance measure between two geographical areas, the Lasker Distance (Rodriguez-Larralde et al. 1994, 1998, Barrai et al., 1987, 1996), the formula for which is below:

$$L_{ij} = -\ln(2R_i) \tag{3}$$

where L is the Lasker distance and i and j are two separate populations. The logarithmic transformation of Lasker coefficient of isonymy often shows a strong relationship with the logarithmic transformation of geographic distance (Rodriguez-Larralde et al., 1994). On this basis one can think of the Lasker Distance as a measure of similarity, or difference, between two populations in surname space (Rodriguez-Larralde et al., 1998). The greater the Lasker Distance the less similar the composition of surnames between the two. Scapoli et al. (2006) suggest this can identify the link between genetic and cultural inheritance as two populations that are genetically homogenous but different from each other are likely exhibit subtle differences in cultural behavior.

Doubts surrounding the validity of isonymy studies are based on the fundamental assumptions they entail. For example the assumption that in some previous generation each male had a unique surname (monophyletic surnames) implies that not only each surname was monophyletic but also that all surname origins occurred in the same generation (Rogers, 1991). As outlined above, we know this not to be the case in the Britain as surnames were acquired gradually and for a multitude of reasons that often reflected commonalities in a variety of populations. Smith, for example, reflects the prevalence of smith occupations within every community. However, even if two populations with a very similar surname distribution are not directly related to one or a few ancestors, they are much more likely to be genetically related between themselves than with a different group that has a significantly different surname makeup.

The large size of the 2001 data required the use of Oracle Database software for storage and the calculation of the Coefficient of Isonymy (Equation 2). The 2001 dataset is significantly larger than the 1881 data, and required approximately 15 minutes processing to complete the Isonymy calculation.

The SQL query produced a table for each of the two time periods, with the R_i values comparing each district with every other district in Britain (ie. a matrix of 658 by 658 in 1881 and 379 by 379 in 2001). This reduced the data volume sufficiently for

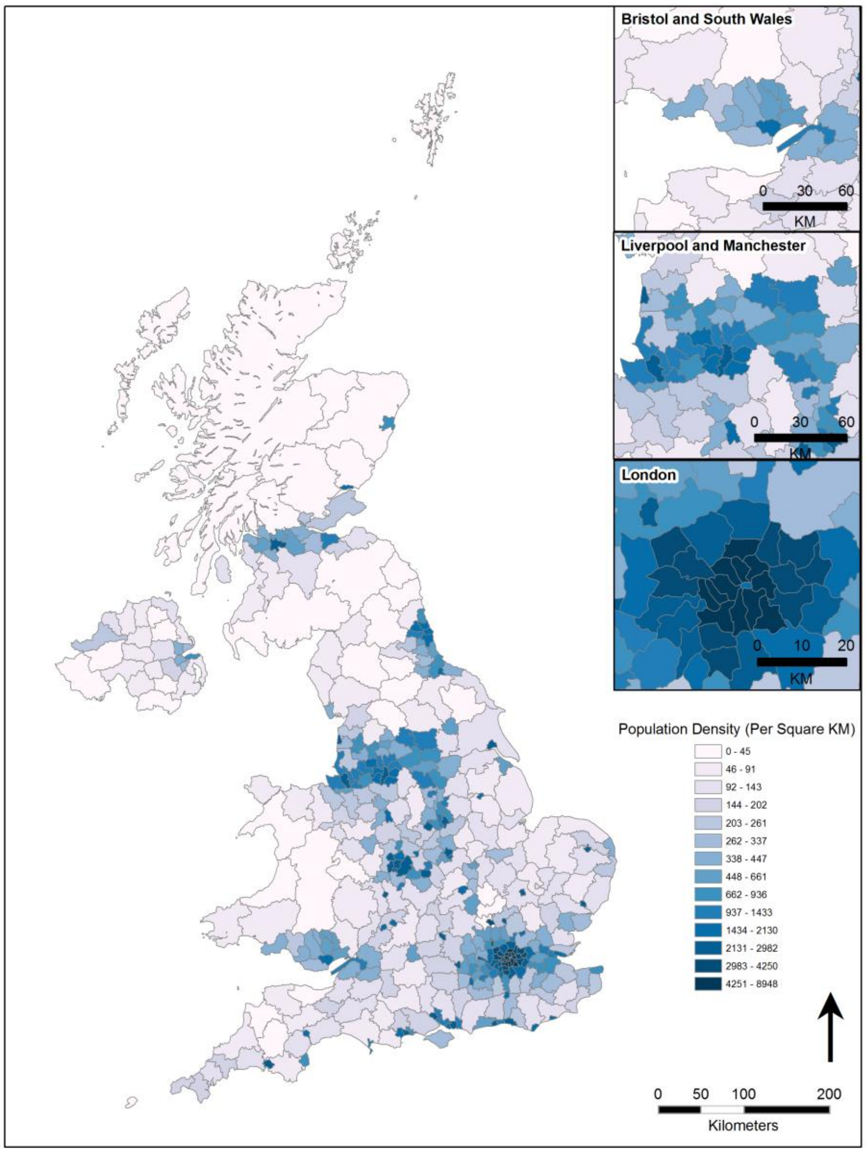


FIGURE 5: A MAP SHOWING THE POPULATION DENSITY FROM THE 2001 ELECTORAL ROLL OF EACH LOCAL AUTHORITY DISTRICT. THE DISTRICTS ARE DESIGNED TO CONTAIN APPROXIMATELY THE SAME NUMBER OF PEOPLE. URBAN DISTRICTS ARE THEREFORE SMALLER IN AREA AND HAVE A HIGHER POPULATION DENSITY AS A RESULT. BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

each year to be manageable as a single object in the R package, which was selected because its thriving open source community has facilitated the development of a number of packages for clustering and spatial analysis. The first step in R was to calculate the Lasker Distance (Equation 3) and create the data matrices. Appendix 1 provides a summary of the methodological steps undertaken to calculate the Lasker Distances.

6.2. Regionalization Methods and their Origins

Throughout the 1960s cartographic techniques dominated the discovery and visualisation of regions. These techniques were, and remain, effective for illustrating areal groupings at a glance enabling differentiation between regional characteristics (Claval, 1998). They are, however, limited to portraying/ differentiating regions based on a single characteristic. Cartographic representations, such as the use of contours, of a particular surname's frequency would highlight the surname's regions; much less effective however would be the representation of multiple names in this way without some prior aggregation.

The limitations of the cartographic approach, combined with a revolution in computing power, have led to an increased interest in automatic regionalization algorithms. From Grigg's (1965, 1967) initial work, classification utilises two methodologies: agglomerative procedures and divisive procedures (Spence and Taylor, 1970). To be effective, these methods require an assessment of the degree of similarity between observations. This is achieved by calculating measures of coefficients of association, correlation coefficients and distance measures. Of these, the most commonly used are distance measures (Lankford, 1969). Distance measures utilise the Pythagoras Sum of Squares equation to calculate the distances between points in n-dimensional space (Spence and Taylor, 1970). The Lasker Distance is classed as a distance measure as it produces a similarity matrix of the coefficient of isonymy between two populations or areas. These methods became popular amongst many geographers and regional scientists as they offered the prospect of a classification based on numerical techniques (Johnston, 1968).

Three subjective decisions need to be made that threaten to undermine objectivity of the resulting regions/ classifications (Johnston, 1968):

1. Whether to use an agglomerative of divisive procedure.
2. The agglomerative/ divisive method employed.
3. How to define group membership.

Since Johnston's article there has been over 40 years of research on which to base these decisions, but consensus is yet to be reached on deriving the optimal number of clusters for a dataset when there is no information regarding the expected number of clusters (Vickers and Rees, 2007). The existence of a number of quantitative methods to inform the decision about the number of clusters (see Gordon, 1999, pages 60-65), Everitt (1972, Everitt et al., 2001) maintains that user evaluation informed by a number of "informal" measures is the best criterion on which to base a decision.

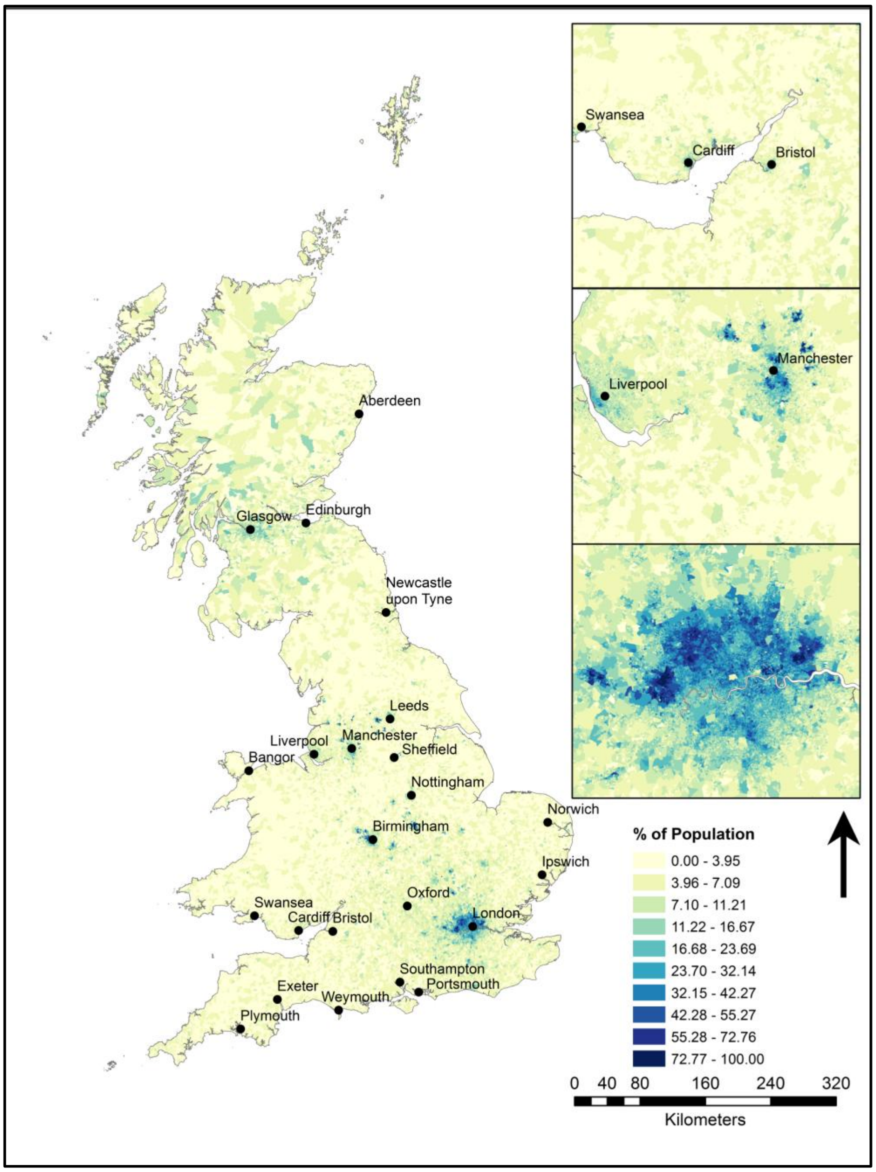


FIGURE 6: MAP ILLUSTRATING THE HIGH NUMBERS OF "NON-BRITISH" NAMES IN LONDON COMPARED WITH THE REST OF BRITAIN AT OUTPUT AREA LEVEL. NAMES CLASSIFIED USING THE ONOMAP CLASSIFICATION (MATEOS, 2008). BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

6.2.2. Agglomerative Procedures

Agglomerative hierarchical methods are amongst the most popular (Everitt et al., 2001). They produce a series of partitions in the data, starting with *n* single-member 'clusters' and finishing with a single group containing all individuals (Everitt et al., 2001). Of the agglomeration procedures, clustering is the most widely used within regional research and neighbourhood classifications (Harris et al, 2005). The ultimate aim of cluster analysis is to produce groups of individuals in which within group variance is minimised and between group variance is maximised (McQuitty, 1957). However, as suggested earlier, the potential to apply one of the following three definitions to group membership can confound the researcher when choosing a clustering algorithm (Johnston, 1968):

The individual to be assigned to the group should be closer to:

1. one member of the group than it is to any other member of another group.
2. all members of that group than to any member of another group.
3. some reference item to the group than to any group's reference item.

Applying the first definition, a classification would group individuals by their nearest neighbours, whilst applying the second definition they would be grouped by a rank order process (Johnston, 1968). The third suggests one of two hierarchical options:

1. Centroid replacement.
2. Assuming the distance between an individual and a group is the greatest distance between an individual and any of the individuals in the group.

6.2.2.1. Ward's Grouping Algorithm

Ward's (1963) grouping algorithm is a popular method of hierarchical agglomeration. The procedure forms hierarchical groups of mutually exclusive subsets that contain members of maximal similarity in terms of the specified characteristics (Ward, 1963). Ward's takes *n* groups (the initial number of observations in the first iteration), reducing them to n-1 exclusive sets by considering the union of all possible $n(n-1)/2$ pairs for the functional relation that matches an objective function chosen by the investigator (Ward, 1963). As with other hierarchical classifications (see Gordon, 1987), Ward's hierarchical clustering produces a dendrogram that can be analysed to establish the relationship between each of the observations. Each time two observations are joined a new node is introduced with branches to the joined observations, the length of which are known as the cophenetic distance. This indicates the strength of the relationship between the observations (Kleiweg et al., 2004).

The clustering was performed with the *hclust* function in R. This function performs a hierarchical cluster analysis using a set of dissimilarities (R Core Team, 2008), provided in this case by the distance matrix of Lasker Distances. The distances between clusters are computed iteratively by the Lance–Williams dissimilarity update formula according to the Ward's clustering algorithm (R Core Team, 2008).

6.2.3. *K*-MEANS CLUSTERING ALGORITHM

K-Means (MacQueen, 1967) is widely used within Geographical Information Science (Bação et al., 2005), and has been especially successful within geodemographics (Vickers and Rees, 2007; Harris et al, 2005).

K-Means is an iterative relocation algorithm that assigns each data point into one of *K* clusters until convergence to a local minimum of its objective function (Bação et al., 2005, Singleton and Longley, 2008). Here the objective function is the sum of squared Euclidean distance (square error distortion or within sum of squares) between each data point and its nearest cluster centre (Bação et al., 2005). The algorithm requires initial seeds to be allocated, around which the clusters will form for the first iteration. Of the variety of initialization methods available the Forgy method is the most widely used (Peńa et al., 1999). This method selects *K* observations (seeds) from the data at random then provisionally assigns the remaining observations to the nearest seed (Peńa et al., 1999). The stochastic nature of this approach reduces the algorithm's sensitivity to outliers (Bação et al., 2004); this is important to reduce the impact of anomalous districts with a large proportion of non-Anglo-Saxon surnames from migration. In subsequent iterations each data point is considered for reallocation to other clusters based on the objective function (Singleton and Longley, 2008). Where reallocation occurs, the cluster centroids are recalculated until the within sum of squares, is minimized or a specified number of iterations is reached (Singleton and Longley, 2008).

R has an in-built function for clustering by *K*-means. The algorithm works on the principles outlined above utilising the Hartigan and Wong (1979) algorithm (R Core Team).

Unfortunately, *K*-means does not guarantee reaching the global optimum as the final groupings rely on the initial groupings (Fotheringham et al., 2007) around the locations of the initial seeds (Milligan, 1980). It is therefore prudent to repeat the process multiple times, 10,000 in this case, and select the optimal objective function from these (de Smith et al., 2007). In addition to selecting the lowest within sum of squares (that is the result with the tightest clusters), the clustering results were mapped and assessed subjectively at every 100th iteration to get an idea of the levels of inconsistency between each run.

6.2.4. Monmonier's Barrier Algorithm

Monmonier's Barrier Algorithm (Monmonier, 1973) is a divisive procedure that includes spatial contiguity in its calculations. The objective of the algorithm differs from clustering as it does not seek to establish maximum internal homogeneity when regionalizing (Monmonier, 1973); instead it seeks boundaries where the differences between pairs of observations on either side are largest (Manni et al., 2004). The algorithm best applied to situations where the boundaries, or barriers, between regions are of greater interest than the areas covered by the regions themselves (Monmonier, 1973, Manel, 2003). It operates on a matrix of observations that have been located on a map according to their relative geographic position (Manel, 2003). Mapping the observations requires Delaunay triangulation (Brassel and Reif, 1979); this is the quickest method of connecting a set of point observations/ localities on a map with a set of triangles that fills a two dimensional space completely (Manni et al., 2004). If conceived as a network topology, the localities are the vertices and the edges are the connections between localities. Each edge is then assigned a distance derived from the data matrix (Manel, 2003). In this study Lasker Distance is used as the distance measure between locations. The first boundary is perpendicularly traced to the edges of the network, equidistant from each pair of observations, starting from the edge with the maximum distance value and continuing until the forming boundary has reached the limits of the triangulation (that is, the edge of the map) or loops back to its origin (Manni et al., 2004). Where edges have the same value, the one followed by a triangle with higher values is included in the boundary (Manni, 2004). The process is illustrated in Figure 7.

The difference between Monmonier's algorithm and the clustering methods outlined above should be emphasized. Whilst the clustering helps to define the regions, Monmonier's algorithm may inform an *explanation* of them by highlighting where strong boundaries exist between regions. In Italy, for example, Monmonier's algorithm has identified barriers between populations based on genetic and linguistic data that match topographical barriers (Manni and Barrai, 2001).

Monmonier's Barrier Algorithm can be implemented with the standalone *Barriers* software (Manni et al., 2004) or the *adegenet* package for R (Jombart, 2008). In this case the *adegenet* package was used.

6.2.5. Multidimensional Scaling

Multidimensional Scaling (MDS) is a well established method of reducing the dimensionality of a data set into an m x n matrix of similarity values (Everitt, 2001). It belongs to the same family of data reduction methods as Principal Components Analysis (PCA). This method is well suited to studies where the distance measures arise directly from previous analysis methods (Everitt et al., 2001). Here, MDS is not used to simplify the data but represent it in a geographical model in three-dimensional coordinate space with Euclidean distance representing the proximities derived from the chosen measure. Each of the combinations of coordinates can be visualized in two or more dimensions to provide a visual (but not geographical) method of detecting cluster structure (Everitt et al., 2001).

The MDS implementation used here creates three dimensional coordinates that can be visualized as three two-dimensional scatter plots or a single, interactive, three-dimensional cube. To map these each of the three coordinates are converted to values between light and dark (0-255) of the three colour components in the spectrum: red, green and blue (Spruit et al., 2009). This is achieved using Kleiweg's (2006) *iL04* R package, originally devised to map linguistic regions. Thus each geographical unit has a unique colour assignment. Similar colours/ shades are produced when districts share similar MDS coordinates and therefore must be closer together in 'surname space'; likewise more colours/ shades indicate a greater surname disparity between regions. The centroids of each of the observations are mapped and then enlarged until they border each other to fill the remaining

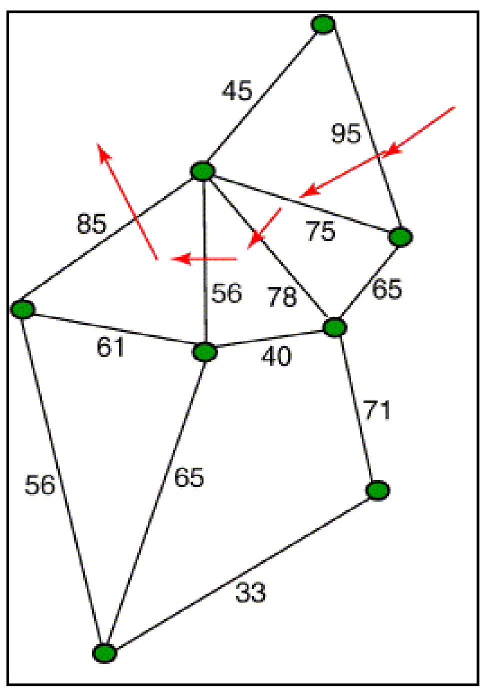


FIGURE 7: HYPOTHETICAL EXAMPLE OF DELAUNAY TRIANGULATION. THE TOPS OF THE TRIANGLES CORRESPOND TO THE GEOGRAPHICAL POSITION OF THE OBSERVATIONS. AN EXAMPLE OF DISTANCE BETWEEN OBSERVATIONS IS ILLUSTRATED BY THE NUMBER INDICATED ON EACH EDGE OF THE TRIANGLES. IN THIS STUDY THE ALGORITHM OBTAINS THIS VALUE FROM THE MATRIX CONTAINING THE LASKER DISTANCE BETWEEN EACH DISTRICT. THE ARROWS REPRESENT THE PATH OF THE FIRST ITERATION OF THE ALGORITHM. STRONGER BARRIERS BETWEEN CENTROIDS ARE CAN BE REPRESENTED WITH THICKER LINES. SOURCE: MANEL ET AL., (2003) PAGE 6.

uncoloured space (Spruit et al., 2009). Group membership from MDS can be established by proximity of a districts the three dimensional coordinates to others or final colour allocation in the MDS maps. Appendix 2 summarises the regionalization steps completed in this methodology.

7. RESULTS

7.1. WARD'S HIERARCHICAL CLUSTERING

Appendix 3 contains an example dendrogram produced from the Ward's Hierarchical clustering. Maps of the resulting cluster outcomes (Figures 8, 9, 10) show that Ward's creates compact, homogenous regions from the Lasker Distance data. To establish the geography of each cluster division multiple maps were produced by increasing the number of dendrogram divisions (and therefore number of clusters). When comparing 1881 to 2001 the first cluster division is one of the most interesting as it suggests that Wales has increased its relative similarity to England, and Scotland has become more different as the first split in 1881 forms between England and Wales, whereas in 2001 this split occurs between Scotland and England. It is not until the fourth split that Scotland is partitioned from the rest of Great Britain, suggesting a greater difference between North and South England in 1881 than Scotland and Northern England. The North/ South split in 2001 occurs at the fourth split and slightly further North of its position than in 1881. The level with five clusters differentiates the far North of England from the combined Northern/ Midland areas in both years, although the partition is located further north in 2001.

The cities in the North West and London create the 6th cluster in 2001; a position occupied by the Southwest in 1881. The former, excluding London, appear at the 7th cluster in 1881 and an enlarged Southwest area, including along the Welsh borders and Bristol Channel are distinguishable by the 7th cluster in 2001. By reviewing the cluster results when dissecting the tree into between 2 and 20 clusters, 15 clusters provides a good balance between capturing the general trends, conformity with prior expectations and ease of interpretation. Whilst, it is acknowledged that there are likely to have been a greater number of natural regions in 1881 due the smaller, more fragmented, population, it was favorable to use the same number of clusters for ease of comparison. It was hoped that the MDS and Monmonier's algorithm would highlight the greater regional variability likely in 1881. At 15 clusters the surname regions of 1881 and 2001 represent very similar patterns. Notable exceptions include the division of Scotland between the highlands and lowlands (including the Scottish Islands), the spread of the Welsh region into England in 2001 and greater differentiation within the South West in 1881.

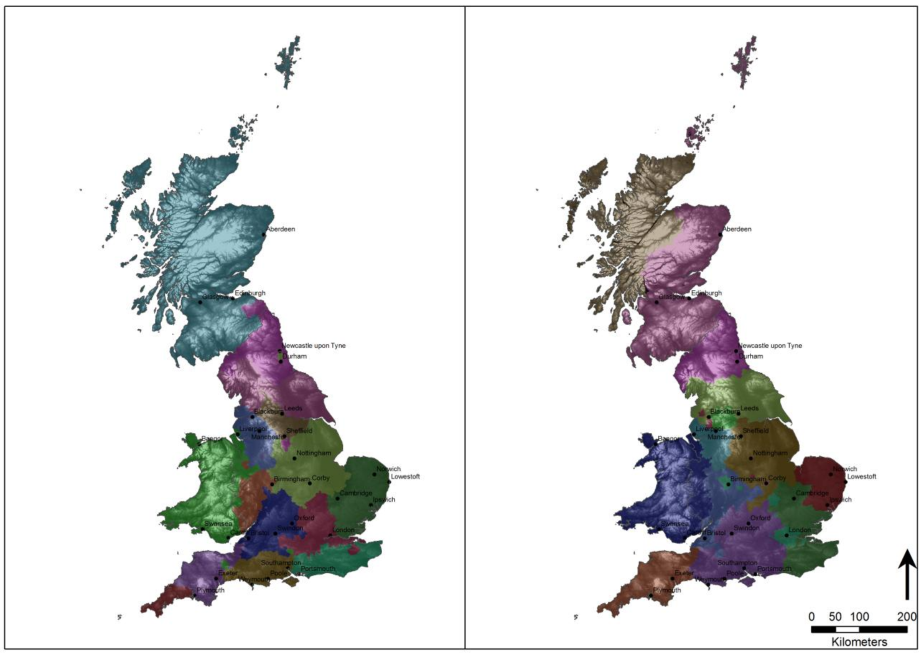


FIGURE 8: FINAL WARD'S CLUSTERING MAPS SHOWING THE 1881 (LEFT) AND 2001 (RIGHT) SURNAME REGIONS AT K=15. THE CLUSTER ALLOCATIONS, IDENTIFIED BY UNIQUE COLOURS, ARE OVERLAIN ON A SHUTTLE RADAR TOPOGRAPHY MISSION (SRTM) IMAGE OF BRITAIN- GIVING AN IMPRESSION OF TOPOGRAPHIC INFLUENCE ON SURNAME REGIONS. BOUNDARY AND SRTM DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

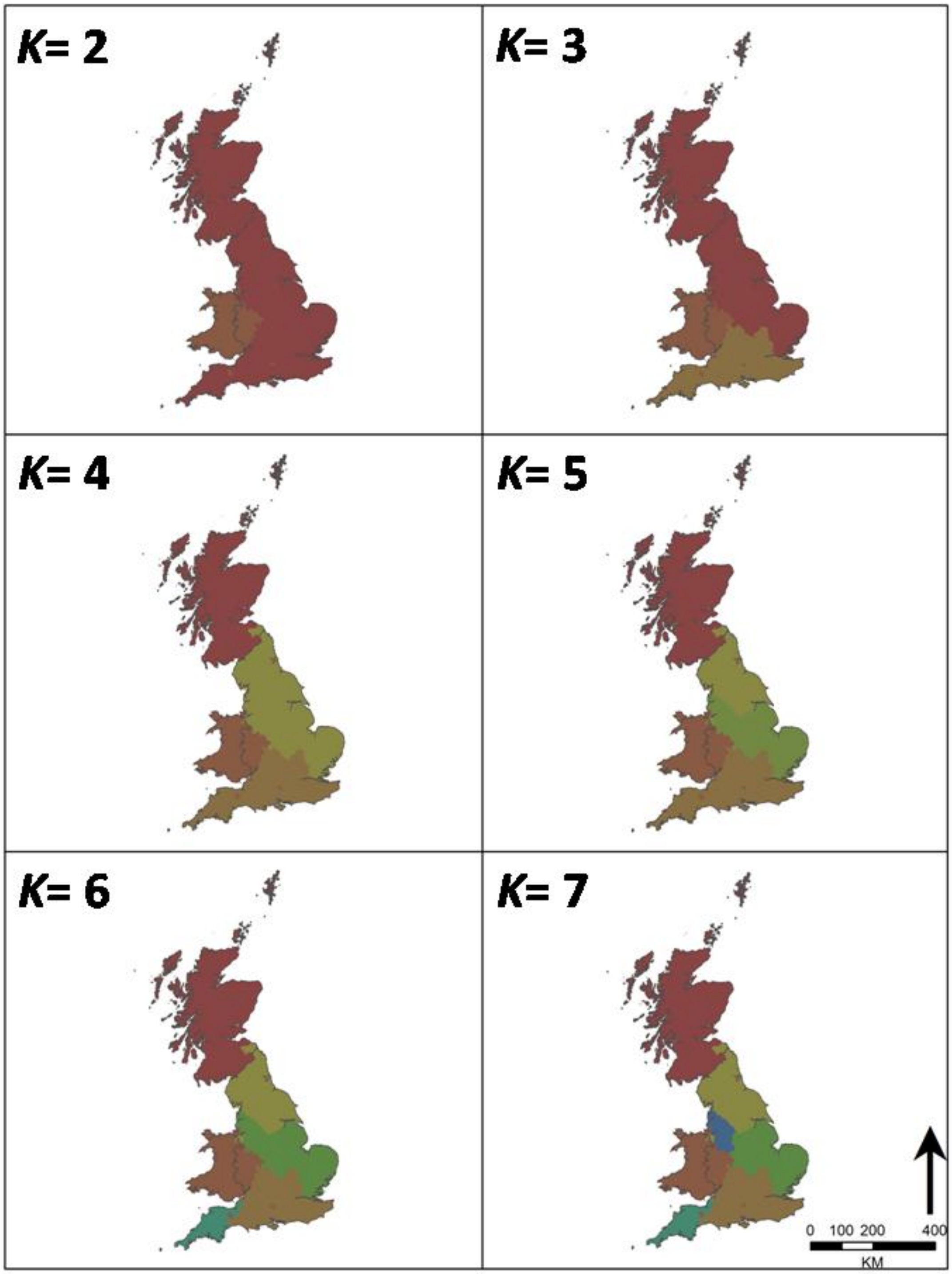


FIGURE 9: MAPS OF *K*= 2 TO *K*= 7 WARD'S CLUSTERS OF THE 1881 LASKER DISTANCES. WALES BECOMES DISTINCTIVE AT *K*= 2 CLUSTERS, THERE IS A NORTH/ SOUTH SPLIT IN ENGLAND BEFORE SCOTLAND BECOMES HIGHLIGHTED AT *K*= 4 CLUSTERS. SOUTHWEST ENGLAND IS DISTINGUISHABLE AT *K*= 6 CLUSTERS. BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

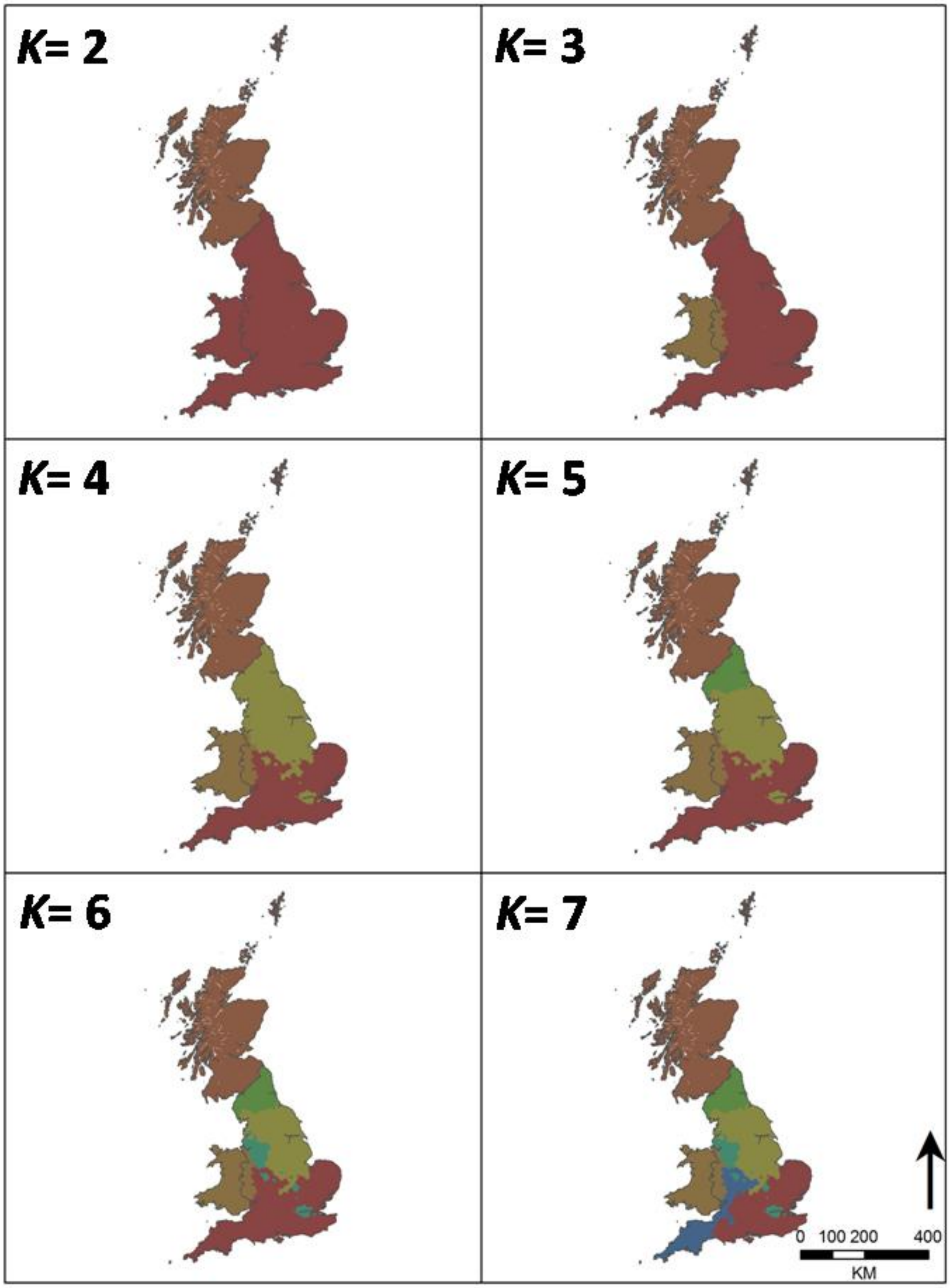


FIGURE 10: MAPS OF *K*= 2 TO *K*= 7 WARD'S CLUSTERS OF THE 2001 LASKER DISTANCES (WITH LONDON AS A SINGLE DISTRICT). SCOTLAND BECOMES DISTINCTIVE AT *K*= 2 CLUSTERS, WALES APPEARS AT K=3 BEFORE A NORTH/ SOUTH SPLIT IN ENGLAND OCCURS AT *K*= 4 CLUSTERS. SOUTHWEST ENGLAND IS DISTINGUISHABLE AT *K*= 7 CLUSTERS. BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

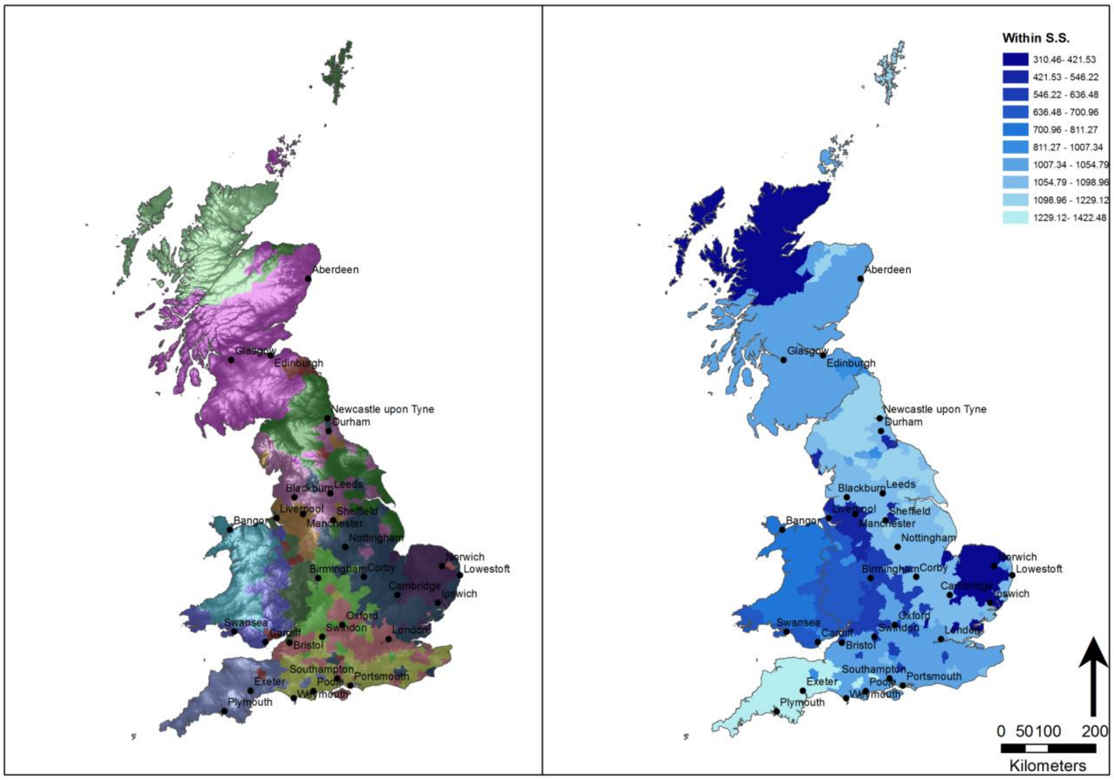


FIGURE 11: 1881 *K*-MEANS CLUSTERING MAPS SHOWING THE SURNAME REGIONS AT *K*=15. THE CLUSTER ALLOCATIONS (LEFT) ARE REPRESENTED BY UNIQUE COLOURS AND LOWER WITHIN SUM OF SQUARES (WITHIN S.S.) VALUES (RIGHT) ARE REPRESENTED WITH DARKER COLOURS TO IDENTIFY TIGHTER CLUSTERS. BOUNDARY AND SRTM DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

7.2. K-MEANS

From Figures 11 and 12 it is clear that the *K*-means clustering algorithm produces smaller, more fragmented, regions. The procedure appears to identify groupings that are more sensitive to variations within Scotland and Wales. Unlike the Ward's algorithm, *K*-means distinguishes three regions within Wales. In both years, the Western tip of Wales (Pembrokeshire) has more in common with the Welsh border regions that extend along the Bristol Channel, including Newport and Cardiff, than central areas of the country. West of Cardiff into the County of Swansea and inland to the Welsh mountains region there are commonalities with the border regions, differentiating it from the bulk of the Welsh land area. Finally the North West of the Wales (the County of Gwynedd and Isle of Anglesey) appears different. The within sum of squares ('withinss') values associated with these observations suggest that the border region of Wales, Central Wales, South Wales and Pembrokeshire are more tightly clustered than the North West region of Wales; one could therefore infer that the degree of difference between this region and central Wales is less profound than between the other Welsh regions highlighted.

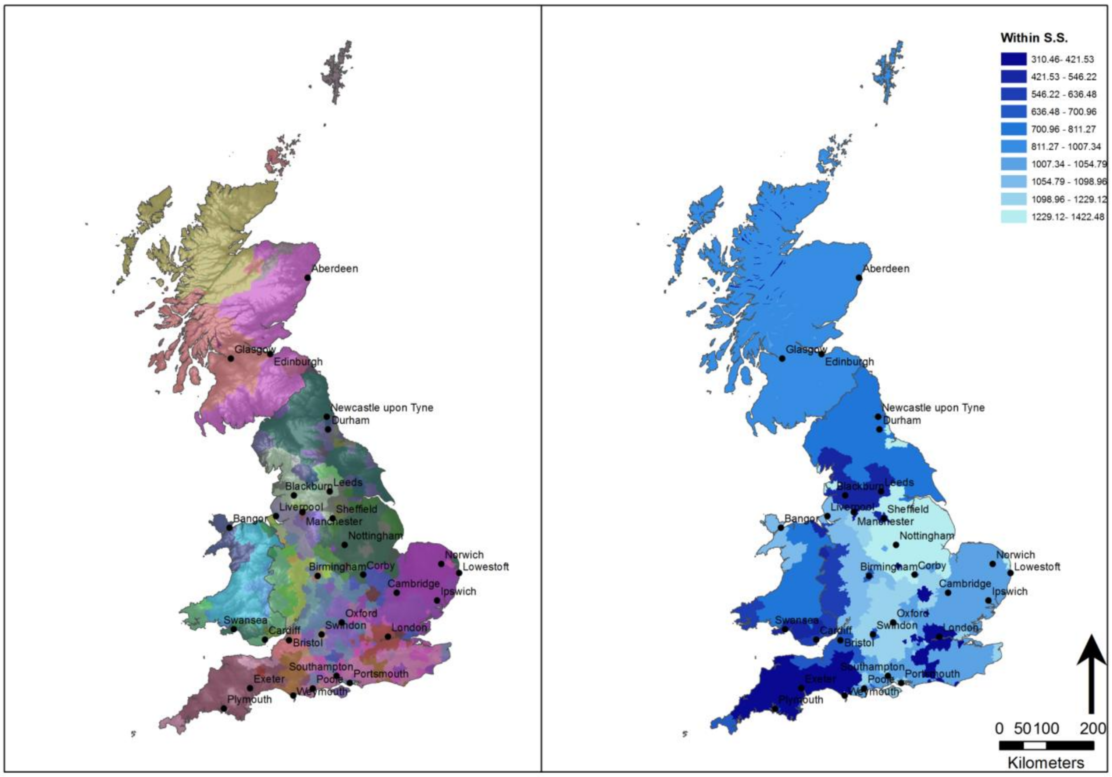


FIGURE 12: 2001 *K*-MEANS CLUSTERING MAPS SHOWING THE SURNAME REGIONS AT *K*=15. THE CLUSTER ALLOCATIONS (LEFT) ARE REPRESENTED BY UNIQUE COLOURS AND LOWER WITHIN SUM OF SQUARES (WITHIN S.S.) VALUES (RIGHT) ARE REPRESENTED WITH DARKER COLOURS TO IDENTIFY TIGHTER CLUSTERS. BOUNDARY AND SRTM DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

In England, the three tightest clusters have created a region for the South-West of England, the North-West conurbation of Liverpool and Manchester and London with suburbs.

Scotland can be approximately divided into highlands and lowlands, with the Shetland Islands sharing a greater affinity with the latter. This split is interesting as it does not appear to be present in 1881 to the same extent, with only the far north of Scotland differentiated from the rest of the country – and creating its own tight cluster. The 1881 results show the Shetland Islands and Moray Firth to share more in common with the far North of England than with Scotland.

The commonality between Southern Wales and the Welsh borders seems to have persisted since 1881, although the pattern at that time is much simpler. Gwynedd and Anglesey remain firmly grouped with central Wales. The extent of the Welsh border region into Wales and England remains largely unchanged.

The within sum of squares highlights an additional change in the likeness between districts that share a region between 1881 and 2001. In 1881, East Anglia is tightly clustered, suggesting relative isolation from its surroundings, yet the cluster disappears altogether by 2001 with the region becoming grouped with the Eastern side of England more generally. The South West cluster enlarged between the years and became significantly stronger, suggesting an increasingly distinctive region compared with the rest of Great Britain. This expansion has not included the Southern tip of Cornwall as it appears to have broken away from the rest of the South West.

K-Means clustering of English Lasker Distances in 1881 produces a noisy map, suggesting a much greater degree of diversity between English districts at that time, or quite possibly a greater variation in data quality. Central England is especially muddled, but discernable regions exist for the Southwest and Cornwall, the South Coast, East Anglia and the Far North.

7.3. Multidimensional scaling

The maps produced from MDS data (Figure 13) agree broadly with the clustering outputs. 1881 presents a much noisier picture with many regions standing out from those contiguous with them. Both maps, especially the 2001 data, illustrate a gradual change from the North to the South or East to the West of the country, with the most abrupt changes occurring at the present national boundaries between England and Scotland and England and Wales. Northern England appeared more similar to Scotland in 1881 compared to today where it exhibits a strong difference from both Scotland and the rest of England. Based on the colour changes in the 2001 map, one can split Great Britain into the following regions in 2001:

1. Northern Scotland
2. Southern Scotland
3. Far North England
4. North West England
5. Wales and England/Wales border region
6. East Anglia
7. Central England.
8. Cornwall and the South West.

These regions appear much less clear in 1881. Great Britain could be split into:

1. Scotland and the Far North of England
2. Wales and England/Wales border region
3. South West England and Cornwall
4. North/Central England

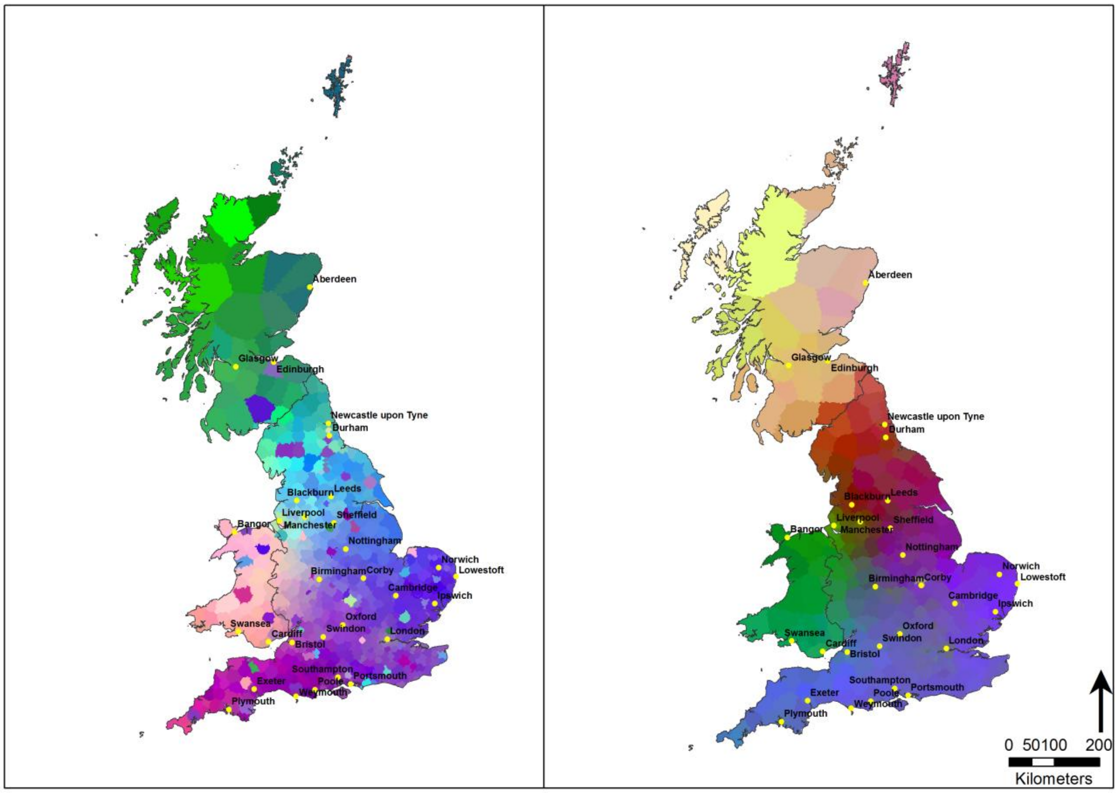


FIGURE 13: MDS MAPS SUGGESTING A MORE GRADUAL TRANSITION OF SURNAME REGIONS IN 1881 (LEFT) AND 2001 (RIGHT). BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

5. Southern, Eastern and Central England
6. Many relatively unique districts throughout Great Britain.

The MDS scatter plots (Appendix 4) attempt a more literal representation of the data used to produce the maps described above. The 1881 MDS plot in the XZ dimension (Appendix 4b) shows how Scotland (light blue) and Wales (purple) appear at opposite ends of the distribution and appear with few other districts in their point cloud. The ZY plot of the 2001 MDS coordinates (Appendix 4c) highlights the clustering of districts from the North West, Wales and West Midlands. All plots show that districts closer to each other are likely to have more similar Lasker Distances.

7.4. Monmonier's Algorithm

The barriers resulting from Monmonier's algorithm, shown in Figures 14 and 15, present a complex picture. One of the most noticeable differences between the datasets is the concentration of barriers around London and the South in 2001, compared with a more even spread in 1881. Commonalities in the results include the Scottish border region, especially prominent in 2001, and a barrier delineating South West England.

7.4.2. 1881 Barriers

What follows is an outline of some of the barriers of interest created from the 1881 data. In the Southwest there are 3 major barriers: one splitting it from the rest of England starting from North Somerset and going South around Poole and the second barrier tracking some way along the Devon/Cornwall border before heading East and stopping short of Exeter. A third barrier excludes Plymouth from the rest of the Southwest. In the far north a barrier splits north and south Scotland, whilst in England a strong barrier forms between the City of Durham and the rest of its county in addition to the division between the north eastern coastal towns and Newcastle Upon Tyne. Moving south, Greater Manchester appears to have a number of barriers surrounding it, suggesting a number of differences between the urban area and its more rural outskirts in 1881. In Wales there is agreement with the *K*-means results for 1881 as the large settlements along the south coast (Cardiff, Swansea, Newport) and Pembrokeshire have barriers differentiating them from the rest of Wales. The islands of Sheppey in Kent and Anglesey in Wales have weak barriers delineating them from the rest of mainland Britain. In addition many rural areas have had barriers drawn around them. This could be due to a lack of social mixing or data artifacts. Finally, unlike today, London and its suburbs, do not appear different from surrounding areas as it has relatively few boundaries around it. One barrier extends through central London, roughly following the Thames, suggesting a North/ South split in the population composition of the areas.

7.4.3. 2001 Barriers

Barriers derived from the 2001 data suggest an east west division in Lasker England. A barrier originates between Manchester and Blackburn tracks south, west of the Peak District, east of Derby and West of Leicester. Using this barrier Liverpool, Manchester, Stoke-On-Tent and Birmingham can be classified as western cities, whilst York, Leeds, Sheffield and Leicester are eastern cities. A barrier further south continues the East/ West split with Oxford and Basingstoke to the East and Swindon and Andover to the West.

On a regional, rather than national, scale other interesting barriers exist. For example, two barriers between Nottingham and Derby in 2001, imply a major change in surname structure. In Northamptonshire, Corby is a town that has been isolated from other areas by a barrier; this division is supported by the other methods utilized in this study. Elsewhere, the coastal fringe of East Anglia creates a strong barrier from the rest of Eastern England. Suggesting the towns of Ipswich, Lowestoft and Great Yarmouth share more commonalities with each other they do

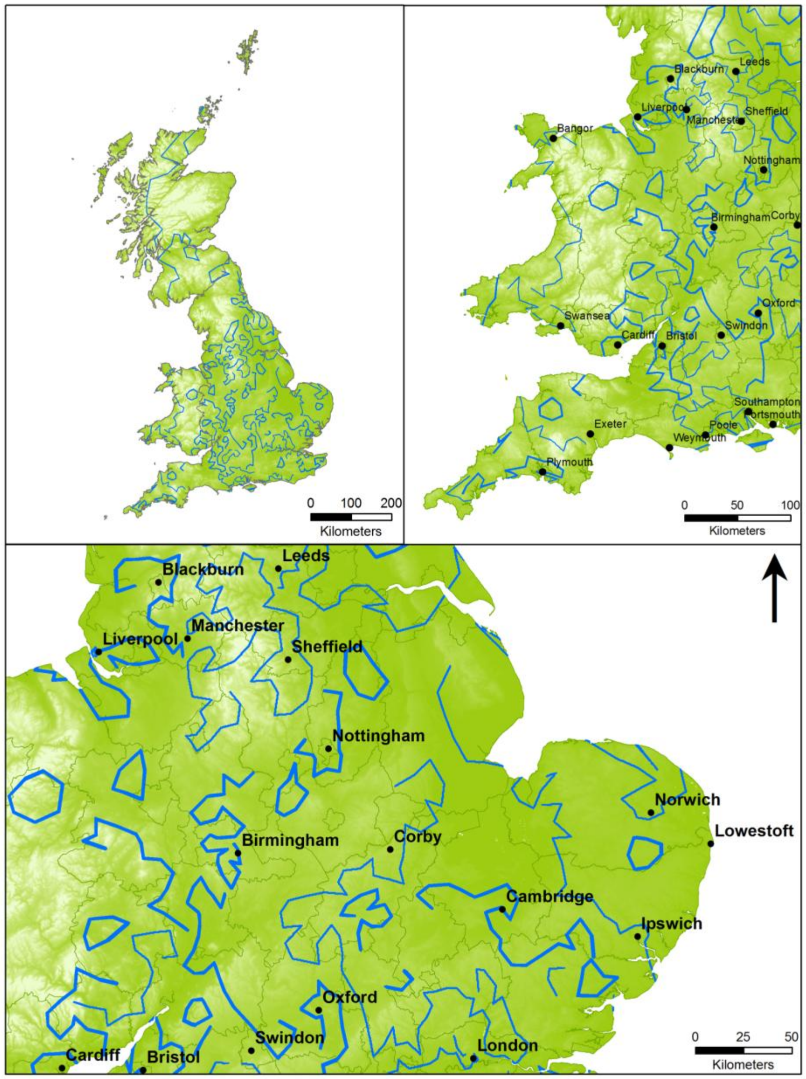


FIGURE 14: 1881 SURNAME BARRIERS CREATED FROM THE MONMONIER'S ALGORITHM MAPPED WITHOUT THE UNDERLYING DELAUNAY TRIANGULATION AND OVERLAIN ON SRTM DATA. CONTEMPORARY COUNTY BOUNDARIES ARE SHOWN IN DARK GREEN. BOUNDARY AND SRTM DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

with the city of Norwich. Cambridge also appears an isolated city in Eastern England with a strong barrier along its perimeter. The final barrier of interest is that which divides the region Dorset (including Bournemouth) into the more urban and coastal South East (including towns such as Weymouth, Poole and Bournemouth) and the more rural, inland North West of the county. The northern edge of this barrier closely follows the Dorset/ Somerset border.

8. Discussion

The results presented in this paper are, to our knowledge, the first attempt to create a regional classification of Great Britain based on two complete population registers.

The results demonstrate that surname regions do clearly exist. These regions are the outcome of inductive generalisation on the geography of surnames in Britain in two time periods and, by extension, can be used as a basis for further hypothesis generation and more in depth analyses to tease out the interplay between historical as well as contemporary processes of cultural interaction in accounting for contemporary distributions, as a contribution to our understanding of domestic and international migration.

The classification methods evaluated here produce broadly similar regionalizations for each time period, although there are subtle differences in the detail of the results. This discussion will begin by addressing some of the methodological considerations before highlighting some common patterns in the results, placing them in the context of previous work. Intended future research will be outlined

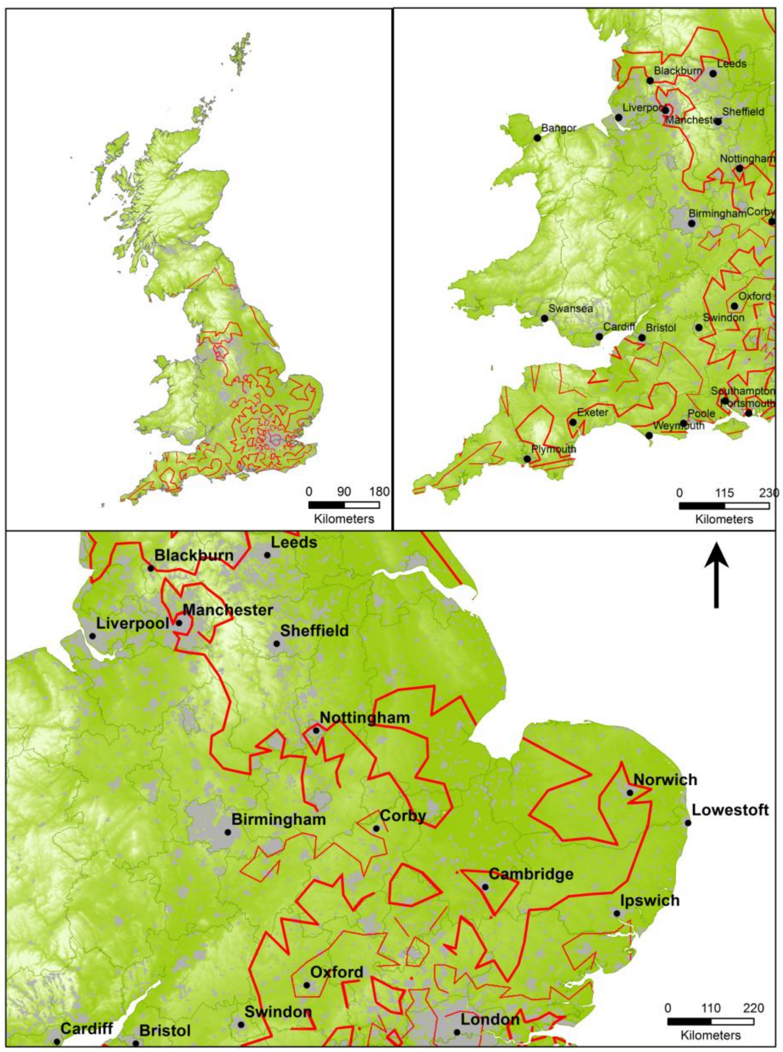


FIGURE 15: 2001 SURNAME BARRIERS CREATED FROM THE MONMONIER'S ALGORITHM MAPPED WITHOUT THE UNDERLYING DELAUNAY TRIANGULATION AND OVERLAIN ON SRTM DATA. IN ADDITION LARGE SETTLEMENT FOOTPRINTS ARE MAPPED IN GREY AND COUNTY BOUNDARIES IN DARK GREEN TO ADD ADDITIONAL CONTEXT. BOUNDARY AND SRTM DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

before concluding.

8.1. METHODOLOGICAL CONSIDERATIONS

8.1.2. SPATIAL UNITS

The choice of spatial unit chosen for data input into the Lasker Distance calculation and subsequent mapping will have had an important impact on the results. Splitting Britain into 379 units for 2001 and 662 units for 1881 along administrative boundaries could be misleading, especially when attempting to gauge the effect of natural barriers to surname interaction as many of the boundaries were drawn along such barriers in the first place. In this case the use of smaller spatial units may clarify the effect of pre-defined boundaries on the results.

In addition, a balance is needed between sufficient numbers of surnames and sufficient detail to depict their diversity. Calculating the Lasker distance using smaller scale spatial units than the District in the case for 2001 will reduce the initial level of generalization in this study but will also increase the noise if results are presented at a national level. The smallest available spatial unit for the 1881 Census is Parish Level rendering this data more limited in finer scale studies than the 2001 Electoral Roll with its geocoding to Postcode level.

In the case of Monmonier's Algorithm , this study may have benefited from the use of coarser resolution data. Figure 18 shows that by utilising relatively few data points on the European scale, clear barriers can be discerned. The large number of short barriers produced by this study are likely to be more representative of small scale variation between districts. By demonstrating that surname regionality is present within Britain, this study may provide justification for the input of the centroids from larger spatial units, as Rosser et al. (2000) have done. The alternative would be to take a more localised approach, such as that implemented by Manni and Barrai (2001).

When comparing the 1881 maps to those produced from 2001 data the smaller populations within each district may be important. The fact that the 1881 data have been partitioned into 50% more spatial units, one can expect a greater degree of small scale variation. This does not appear to have been the case with the Ward's clustering results; it is harder to quantify with Monmonier's algorithm results, but more evident with the K-means and MDS maps. Aggregating the 1881 data to larger spatial units or reducing the size of the 2001 spatial units may serve to clarify the extent to which the noise is an artefact of the spatial units as opposed to data quality, or the result of genuine differences between populations' surnames in Great Britain.

8.1.3. Lasker Distance

The underlying assumptions of the Coefficient of Isonymy as discussed in the introductory section are seen by some to undermine the validity of the Lasker distance. Whilst Roger's (1991) concerns are acknowledged, there can be no doubt that the results produced in this study are plausible and externally verifiable, making the measure a compelling one in this context. It should be emphasised that the intended application of these results is for hypothesis generation and to be used as a basis for further work regarding the clustering of surnames. Many of the limitations of using Lasker distance measures are levelled at those drawing conclusions about the genetic similarity of a population as maintained by the degree of inbreeding that occurs. The measure remains one of the most widely used and can be relatively straightforwardly applied to large datasets.

8.1.4. Including Space

The regionalization methods outlined above do not require boundary data. The spatial aspects of the data are only utilised in the visualization (mapping) and interpretation phases of the analysis. The acceptance of such methods became a key debate during the Quantitative Revolution in Geography hinging on whether, when areas are being grouped to form regions, location should be: (1) used as one of the discriminant variables, (2) the dominant variable, or (3) considered at all (Johnston, 1970). Subscribing to (1) and (2) entails the acceptance of contiguity constraints and requires the development of classification methods specific to geography. Whilst the authors acknowledge that in certain contexts contiguity is important (for example, when partitioning space for administrative purposes (Monmonier, 1973)), the authors of this work share Johnston's (1970) view that "regionalizing with contiguity constraints over simplifies and operates against efficient hypothesis testing. There is no basis in geographical theory...for the adjacency requirement" (1970: 295).

Applying a contiguity constraint prevents the creation of multiple geographically separated regions that share a common class (Johnston, 1970). In the context of surnames this is unsatisfactory as it masks the existence of similar regions that have developed as a result of migratory processes between areas, such as the example of Cornish economic migrants moving to Middlesbrough in the 19th Century as uncovered through surnames by Longley et al. (2007) and the Scottish migrants in Corby (see below).

8.1.5. Regionalization Methods

Based on their consistency with the regions delimited by other methods, historical information and the simplicity of interpretation, it would appear that Ward's Hierarchical Clustering and MDS are the most promising methods for delimiting the surname regions of Britain.

A key limitation for all clustering methods is the underlying assumption that the optimal number of clusters in the data is known beforehand, something that is not necessarily true in the real world (Peńa et al., 1999). A subjective decision is required about the number of clusters to be created. This work has sought to select optimal cluster solutions based on ease of interpretation and a priori substantive knowledge.

Creating MDS maps by assigning colour values to the coordinates of the values in 3D MDS space appears more elegant than straightforward clustering because it depicts both gradual and abrupt change. In addition, a subjective number of clusters are not required thus minimising the chance of misleading distributions created by too few/ many clusters. The limitation of this approach, however, is the subjective nature of the interpretation caused by different perceptions of colour.

A central aim of this work is to identify broad regions that share a similarity in surname composition. Ward's clustering was the method that best achieved this. Whilst *K*-means has demonstrated its effectiveness in other geodemographic classifications, such as the Output Area Classification (OAC) (Vickers and Rees, 2007), it results in a noisier picture of British surname geography. The information contained within this noise should not be discarded; *K*-means, for example, appeared to highlight the differing groups within Wales that have been supported by historical evidence. The technique, however, may have been more effective at partitioning into larger numbers of clusters, as is the case with the OAC. Ward's clustering not only identified the broad regions, but was also sufficiently sensitive anomalies, such as Corby. The number of clusters represented can be easily varied without having to re-cluster the data; this enables outputs such as Figures 12 and 13 to be produced, something not possible with *K*-means due to the algorithm's stochastic nature. For these reasons alone, Ward's can be considered the strongest of the methods used here for creating generalised regions in Britain.

Less spatially extensive studies, such as those concerned with a particular Government Office Region, or those seeking a greater number of regions, may be better suited for *K*-means or Monmonier's Algorithm. Both methods have demonstrated merits and have provided a useful addition to Ward's Clustering and MDS.

8.2. Common Patterns

8.2.2. Wales

The strongest regions identified in Great Britain are England, Scotland and Wales. The Welsh surname border consistently appears to extend beyond its administrative border into parts of England. This is especially evident from the Ward's clustering results (Figure 10) for 1881 where the first division places the Welsh cluster as far east as Birmingham. The reduction in influence of Welsh names along the border regions between 1881 and 2001 is unsurprising, as the relatively low diversity of Welsh surnames (Hey, 2000) makes relatively minor increases in surname diversity likely to cause relatively homogenous regions to "retreat" to their heartlands. Additionally, the *K*-means results for 1881 suggest that the Welsh border region extends west within the contemporary Welsh administrative border and along the South coast of the region. There appears increased similarity between the Welsh districts and this region as the cluster containing the border and south Wales reduced in extent when compared to 2001.

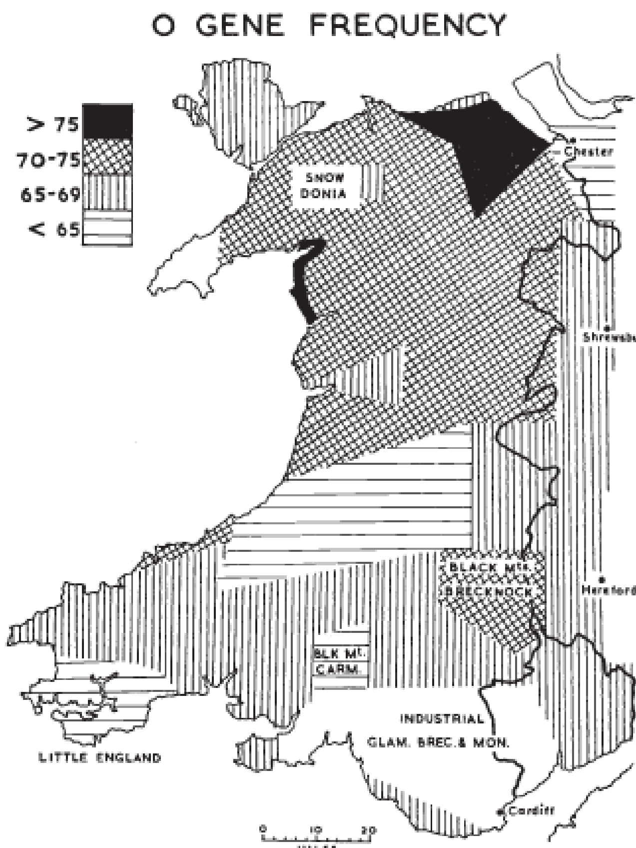


FIGURE 16: AN ILLUSTRATION OF THE DIFFERENT GENETIC CHARACTERISTICS OF THE WELSH POPULATION, MANIFESTED THROUGH BLOOD TYPE O GENE PERCENTAGES. NOTE THE DISTINCTIVENESS OF THE 'LITTLE ENGLAND' AREA RELATIVE TO ITS SURROUNDINGS. SOURCE: WATKIN, M., 1956. P. 66

A much-vaunted cultural border within Wales is "Little England beyond Wales". A Norman invasion, around 1100 created an outpost in Pembrokeshire (Laws, 1888) that has been described as "a peninsula which is physically and psychologically semi-detached and somewhat independent-minded" (Heath, 1997). The area is known for its opposition to Welsh devolution, for example. There is also evidence of linguistic and genetic differences between "Little England" and the rest of Wales (see Figure 16) (Watkin, 1956). One would, therefore, expect a clear difference in the surname structure of this region in comparison to the rest of Wales. The *K*-means clustering appeared most sensitive to this difference with Pembrokeshire being highlighted for both centuries; the MDS maps also show a slight colour difference between Pembrokeshire and the rest of Wales, especially in 1881. Monmonier's algorithm produced a small barrier in the region for 1881 but created nothing for 2001. Ward's clustering appeared least sensitive with no differentiation of Pembrokeshire, even when partitioning Great Britain into 20 clusters. The methods that recognize "Little England" invariably cluster it with the southern urban areas of Wales. These are the most connected to England and may therefore have attracted a relatively large number of migrants from the outside the region as far back as the 19th Century. The *K*-means results and Figure 17 suggest that the Welsh Capital, and location of the Welsh Parliament, may have a population that is as much English as it is Welsh; a fact that appears to have been the case as far back as 1881.

8.2.3. Cornwall and the South West

The South West of England (and more specifically the approximate area of Cornwall) is a distinctive British region. Its extent differs between 1881 and 2001. As with Pembrokeshire, there are present day political manifestations of Cornwall's historical difference from the rest of Britain; for example 50, 000 people signed a petition in 2001 calling for a Cornish Assembly (BBC News, 2001). For the 1881 Census, MDS, Monmonier's Algorithms and Ward's clustering all present convincing evidence of a distinctive region centered upon Cornwall. This remains the case in 2001.

Cornwall provides a good illustration of the links between surnames and genetics. Figure 18, produced by Rosser et al (2000), demonstrates a clear genetic barrier between Cornwall and the rest of Europe arising from significant differences in Y-Chromosomal diversity. Differences between the Cornish and the rest of the British have been noted since the Middle Ages given that the natural defenses of the sea on three sides and the River Tamar on the fourth side allowed the population to maintain independence from the Anglo Saxons until 937 AD (Stoyle, 1999). According to Stoyle (1999) there were those who believed that Cornwall possessed

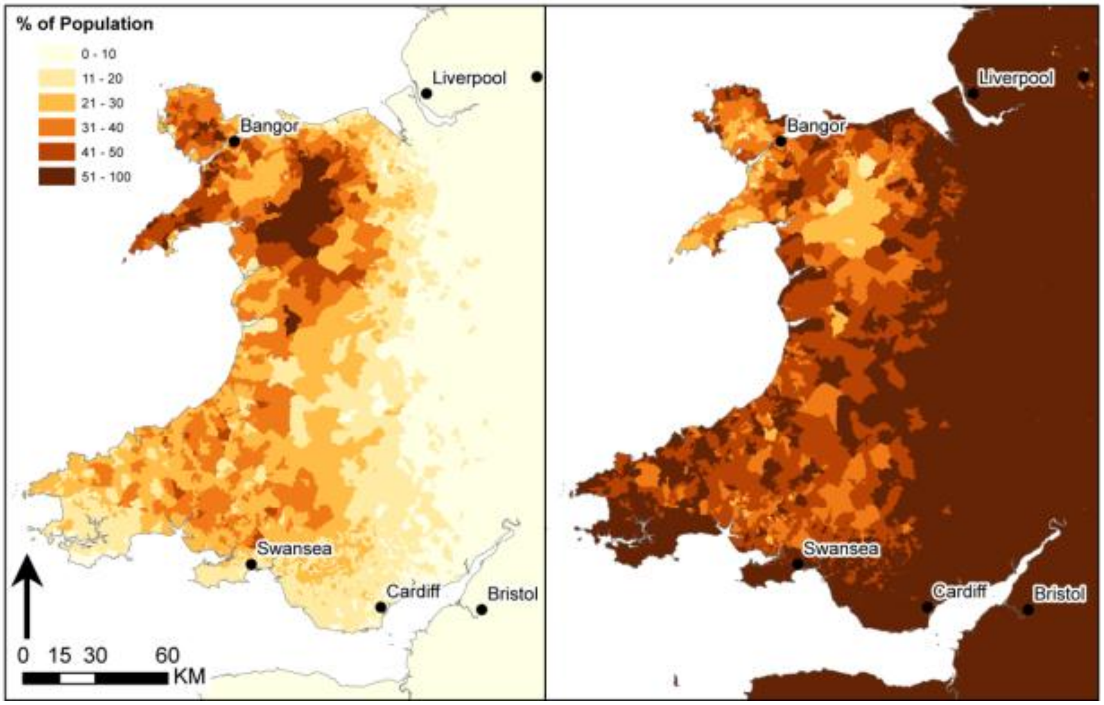


FIGURE 17: MAPS ILLUSTRATING THE PERCENTAGE OF THE POPULATION WITH WELSH (LEFT) AND ENGLISH (RIGHT) NAMES AT CENSUS OUTPUT AREA LEVEL. THE DATA ARE TAKEN FROM 2001 AND CLASSIFIED USING THE ONOMAP CLASSIFICATION (MATEOS, 2008). BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

a separate identity even before the creation of Scotland and Wales. It is nonetheless remarkable that many of these sentiments still exist, and that an area as small as Cornwall has maintained a unique surname structure over the last century despite the unprecedented connectedness of contemporary society.

8.2.4. CORBY: A SCOTTISH TOWN IN ENGLAND?

The primary motivation for comparison of the 2001 Electoral Roll with the 1881 Census is to highlight areas that may have been especially affected by migration during the past century. The town of Corby in Northamptonshire presents one such illustration of domestic migration. When mapping the Ward's result for K= 2 (Figure 10, also clear from Figure 20) it is evident that Corby is clustered with the Scottish districts in 2001, but not 1881. The town is also highlighted in the 2001 Monmonier's barrier and MDS maps. One could infer that a migration event from Scotland has occurred since 1881 to produce a surname composition so similar to that of Scotland. This proves to be the case. In 1932 a company called Stewarts and Lloyds announced a project for a new iron and steel works in Corby. The development transformed Corby from a village of 1,500 people to a new town of 34,000 with 10,000 employed at the works (Pocock, 1960). Labour was sourced from the contracting or closing Scottish steel works; where workers had the choice

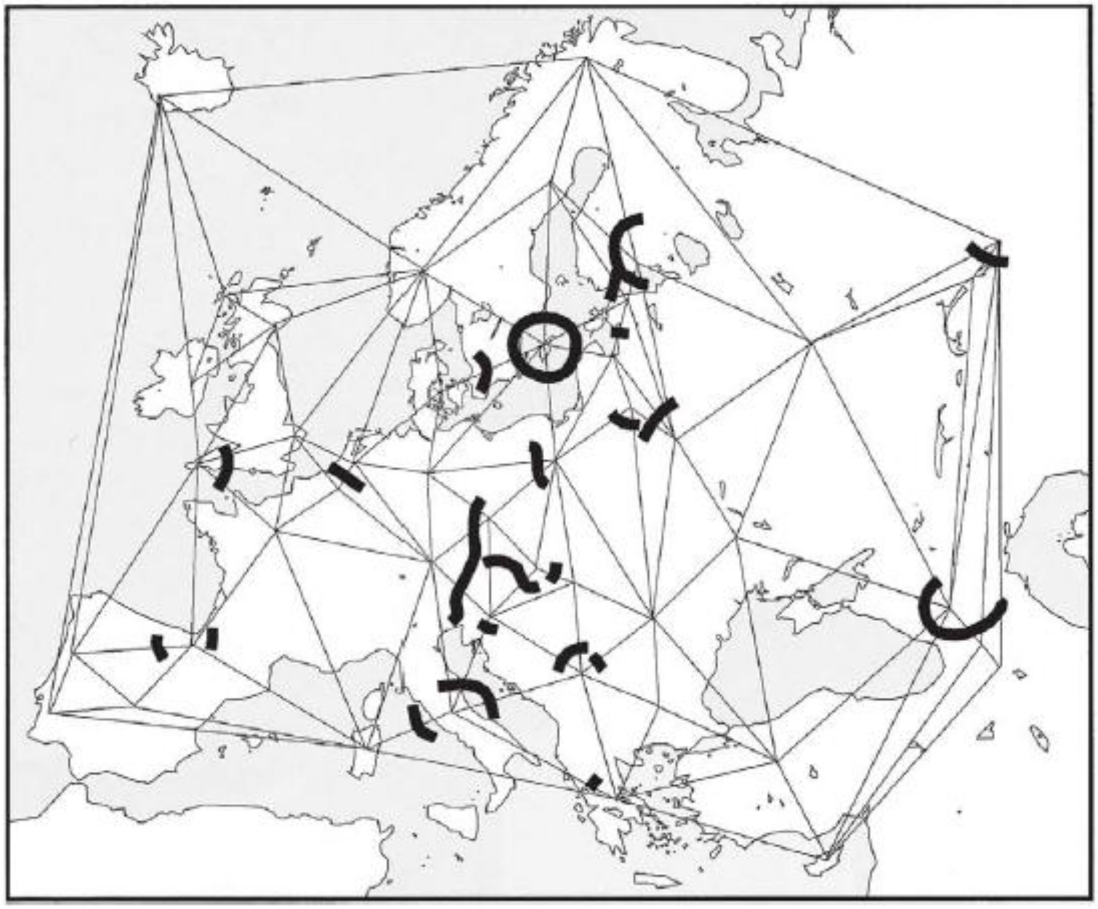


FIGURE 18: AN INTERPRETATION OF GENETIC BOUNDARIES PRODUCED FROM THE ORINICO PROGRAM (SEE PAPER FOR DETAILS). SOURCE: ROSSER ET AL., 2000, P. 1538.

of redundancy or moving south to Corby (Grieco, 1985). Recruitment continued into the 1970s, with Scottish migrants accounting for up to 50% of the incoming population and up to 57% of inhabitants reporting Scottish origin in some areas (Grieco, 1985). Grieco (1985) reports the maintenance of strong links between Corby and Scotland with an annual Highland Games and 55% of all visitors registered in the 1981 census reportedly from Scotland. The steel industry collapsed in the 1980s; the departure of British Steel left the town "with a severely imbalanced social composition, a labor force with skills inappropriate to the economic activity of the surrounding area [and] poorly placed to attract employers into the area" (Grieco, 1985, P16). With such bleak prospects and strong links to Scotland, it is surprising that significant out-migration of the Scottish community in the past two decades has not occurred. That this is not the case presents interesting research questions. For example, how many of the present-day inhabitants of Corby were actually born in Scotland? Here surname geography also shows its value to identify second and subsequent generations of migrant descendants. Having ancestors that were economic migrants in difficult times can have a lasting effect

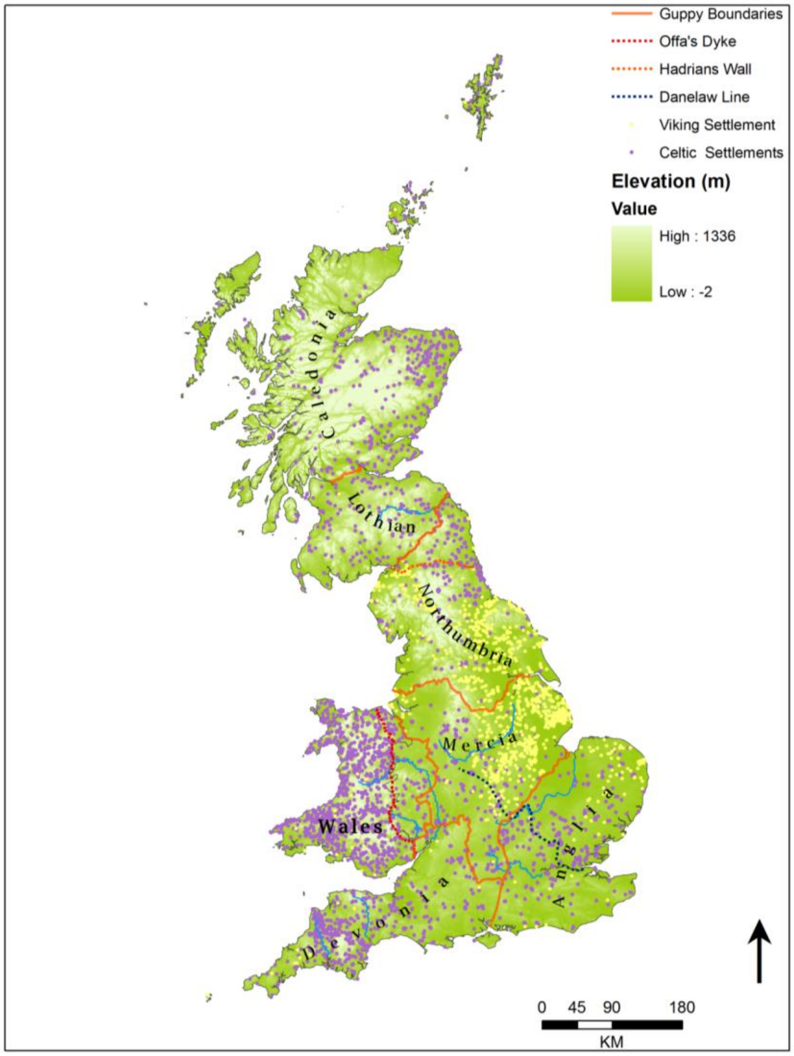


FIGURE 19: BRITAIN'S HISTORICAL BOUNDARIES, GUPPY'S SUGGESTED REGIONS WITH THEIR ASSOCIATED NAMES, TOPOGRAPHIC INFORMATION AND MAJOR RIVERS. IN ADDITION THE MAP ALSO SHOWS SETTLEMENTS THAT FOLLOW VIKING (YELLOW) AND CELTIC (PURPLE) NAMING CONVENTIONS. THESE CONVENTIONS ARE LISTED IN APPENDIX 5. BOUNDARY, SETTLEMENT AND SRTM DATA CROWN COPYRIGHT ORDNANCE SURVEY 2009.

through generations. This can be disclosed by surname geography, as already demonstrated by Longley et al. (2005) in the study of Cornish miners to Middlesbrough and the socioeconomic characteristics of their descendants.

8.2.5. Similarities with Historical Boundaries: the Danelaw line

Anglo-Saxon Britain may provide an explanation for the North/ South split in the surname composition of England evident in Figures 12 and 13. As Figure 19 shows there is also a division in settlement naming conventions along the Danelaw line which marks the southern extent of Danish rule in England during the 9th and 10th centuries (Darby, 1973). Whilst it is unwise to "read too much between the dots" (Keynes, 1997) to infer the population characteristics of the area (in relation to

Celtic/ Saxon/ Viking origins) they do provide useful context. To the north of this line it is likely that there was some integration of place naming practices between the Danish and native populations within Danelaw. Evidence suggests that the spread of Danish names south of Danelaw took place through the land owning elite and would therefore have had a minor influence on the broader population (Keynes, 1997). Figure 20 shows the path of the Danelaw line against the results from each of the classification algorithms used here, for both 1881 and 2001. It illustrates the clear division in surname structure along the Danelaw, in a similar fashion to place names in Figure 19. Unsurprisingly, the correspondence with the classification is most evident in 1881, but remains apparent in 2001. Until 1881, at least, it appears that a thousand years of population change had left the underlying surname geography fundamentally unchanged in Britain.

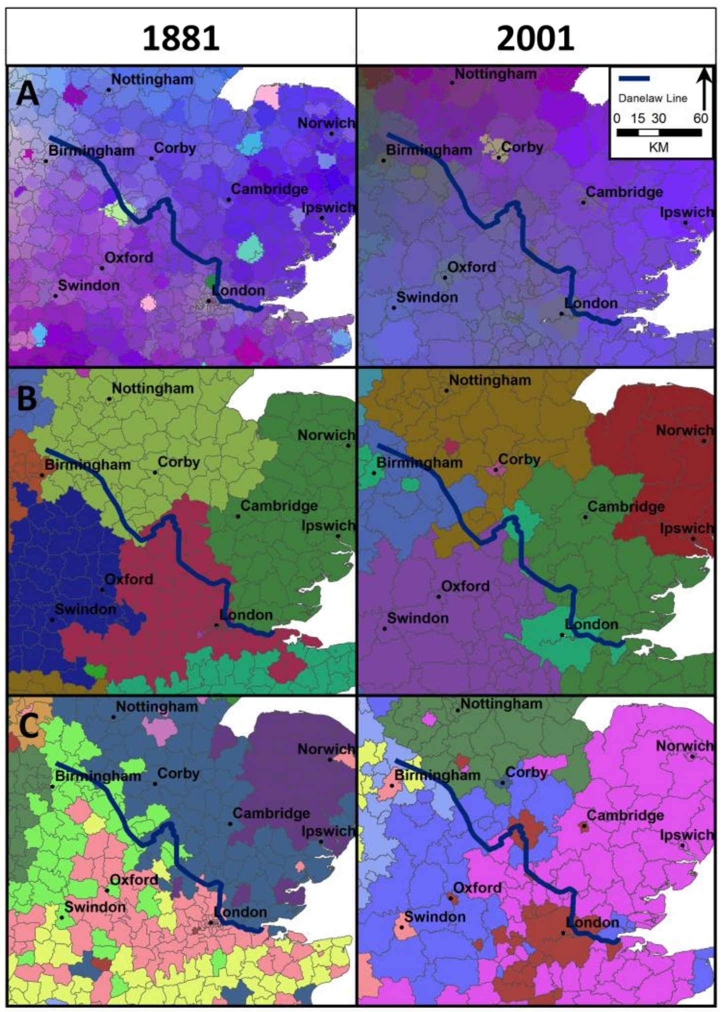


FIGURE 20: MAPS DEMONSTRATING THE CORRESPONDENCE BETWEEN THE PATH OF THE DANELAW LINE AND BOUNDARIES BETWEEN SURNAME REGIONS AS IDENTIFIED BY MDS (A), WARD'S CLUSTERING (B) AND K-MEANS CLUSTERING (C). THE MAPS ALSO DEMONSTRATE THE SCOTTISH CLUSTER ALLOCATIONS ASSIGNED TO CORBY IN 2001. BOUNDARY DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

8.3. COMPARISONS WITH PREVIOUS WORK

Few studies have sought to identify surname regions within Great Britain, but those that are available nevertheless provide a useful comparison with the results obtained here. Guppy's 1890 description of surname regions (Figure 19) closely resembles the Ward's clustering results with transitions in surname structure occurring along most of the specified borders (See Figure 21). This result is interesting because Yeoman were the only group studied in Guppy's work on account of their "stationary habits and purity of extraction". The results presented above suggest that Guppy was overcautious in his work and that the wider population exhibited the Yeoman characteristics of being "but little affected by the wars and political fractions of their times...not troubled with ambition, and few cared to wander far from the vicinity of their birthplace" (Guppy, 1890, P2). Unsurprisingly, there is less correspondence between these borders and the 2001 data, suggesting that the migration associated with a more mobile society is having an effect on the traditional surname regions, at least in England.

Although the surname-frequency boundaries, shown in Figure 1, from Sokal et al.'s (1992) "Spatial Analysis of 100 Surnames in England and Wales" suggest a different distribution of surname regions to those produced here, their general observations concur with this study. They produce strong evidence for isolation by distance; that is populations further apart are less likely to mix and share genetic characteristics. The MDS maps represent this phenomenon well where those regions furthest apart, such as Northern Scotland and Cornwall, have very different colour assignments with a gradual transition of colours between them. The notable exceptions to this are the Welsh border region in both centuries and isolated districts, such as Corby, in 1881 and 2001 and the Scottish border in 2001. In these instances other phenomena, for instance large-scale migration, generate anomalously large differences in surname structure between districts.

Sokal et al. (1990) find historical influences and traditions to be the primary influences on surname distribution. These findings are supported here by the effect of the Danelaw line on surname regions and also the lack of influence that topographic barriers such as large rivers or mountains have had on the surname regions in this study. The three migration patterns - North/South and East/West diffusion combined with local dispersal- discussed by Sokal et al. (1993) may be sufficient to explain some of the patterns found, such as the advance of Welsh names or the changing division between North and South Scotland. Sokal et al. (1999) sampled only 100 unique surnames from the 2001 total of 1,597, 805 to obtain their findings; this may go some way to explaining the differences in geographical boundaries between surname regions. Further analysis is also required to establish

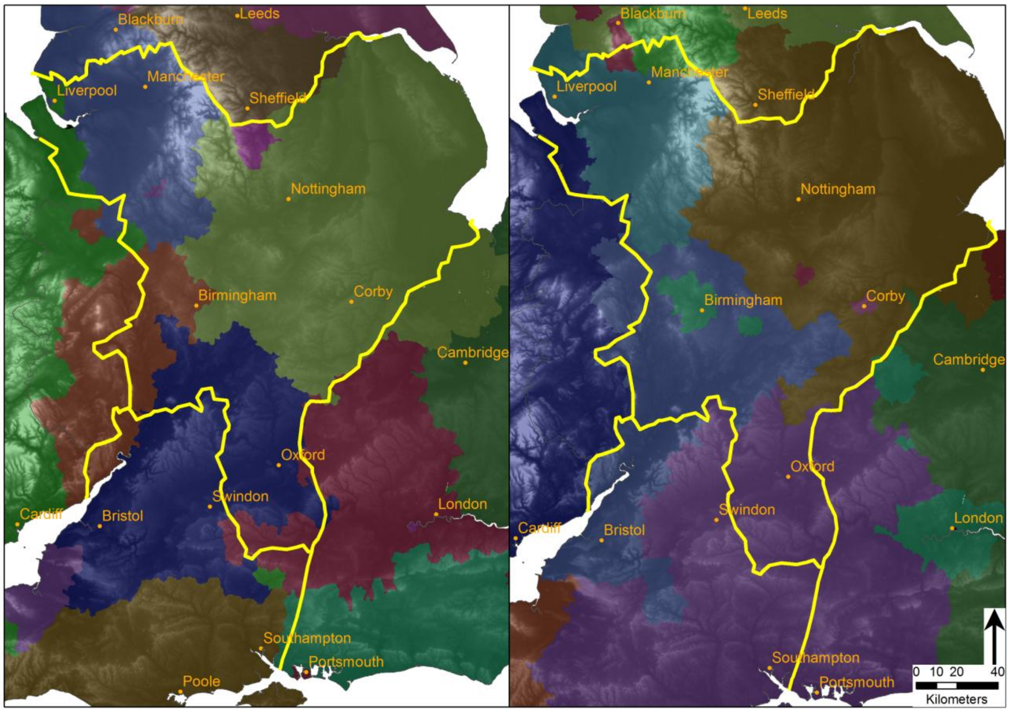


FIGURE 21: A DEMONSTRATION OF THE CORRESPONDENCE BETWEEN GUPPY'S SUGGESTED BOUNDARIES FOR CENTRAL ENGLAND AND THOSE CREATED FROM WARD'S CLUSTERING OF 1881 (LEFT) AND 2001 (RIGHT) USING LASKER DISTANCES. BOUNDARY AND SRTM DATA: CROWN COPYRIGHT ORDNANCE SURVEY 2009.

the degree to which the regional changes presented here are a characteristic of the varying spatial units used in 1881 and 2001. Of more interest is the fact that the conclusions they make regarding phenomena such as isolation by distance and the importance of historical factors have been found to hold for the broader population in Great Britain, and not just the 100 surnames sampled.

The results not only reinforce past research but provide an empirical basis to thinking about new regions that have hitherto attracted little attention. On a cautionary note, the significance of these new regions may need to be weighted by their population size- something that is a possible extension to the presented methodology. The confidence in the results should encourage their intended use as a basis for hypothesis generation.

8.4. FUTURE WORK

The results presented here provide a firm basis for continued research into generalised patterns and hypothesis generation. The former should include continued refinement of the methods employed here whilst undertaking a critical analysis of the resulting surname regions / trends. Alternatively, these results could provide a basis for hypothesis generation. There are many possibilities; studies could focus on the specific by investigating the local patterns such as the endurance of a 'Little England' or the closeness of the relationship between Corby and Scotland. Smaller geographic units, such as Middle Level Super Output Areas could present a more appropriate scale for these studies. Establishing whether the processes behind the creation of the observed discontinuities in surname structure are continuing, stable, or in decline presents a further interesting avenue of research. Although one should be cautious about the analogous treatment of surnames and genetic regions, the regions created here would provide an interesting foundation on which to base a sample design for genetics research. As has been demonstrated, the elapse of many generations has failed to homogenise the distributions of surnames across Britain. This contradicts the genetics of the Great Britain that can be characterised by much greater homogeneity in the distribution of genes (Kaplan and Lasker, 1983). Regions that demonstrate the greatest discontinuity of surnames in relation to their neighbours may be of particular interest when developing a sampling strategy. Close work with geneticists through intelligent sample design would begin to unravel the explanation for the apparent contradiction between Britain's heterogeneous pattern of surname distributions and the relative genetic homogeneity of its population.

9. Conclusions

"*It might appear...that the family of nomenclature of Englishmen was for the most part in a confused jumble, and that on account of the rapid means of inter-communication, which we enjoy in the present Century, most of the distinctions that existed in the past would have been lost in the whirl and bustle of the industrial era in which we live. It might have seemed...that chance had played such as part in the intermingling of inhabitants of different counties and districts, that it would seem a hopeless task to unravel the entangled skein...I found it was yet possible to pick up the threads. By this means I have found order where I expected disorder and method where I only looked for chance. "* Henry Guppy, 1890.

By unearthing the broad regional geography of the British population a basis is established for future work on its population dynamics. The hereditary, and therefore genetic, nature of surnames provides additional context to inform more local studies into the strength of association between settlements in Britain. Regional identity can also be explored in those areas on periphery or between major regions such as "Little England", the Welsh and Scottish borders, and Cornwall as these populations become mixed with national and international migrants.

The contemporary relevance of the extract above suggests little has changed since the 19th Century. Guppy (1890) outlines the importance of recording present surname distributions before “present peculiarities and distinctions are lost”. He is unlikely to have anticipated the technological innovation that has facilitated communication and migration on an unprecedented scale. But, he would be surprised to hear that despite over a century of unprecedented migration this work has shown that many historical distributions still remain in addition to the creation of new “peculiarities”, such as the town of Corby. There is no doubt that areas have become more similar and that current population trends suggest a continuation of the homogenisation of the population characteristics of Britain, but compelling evidence has been presented for the persistence of underlying trends that, to this day, have not been lost in the “whirl and bustle” of the post-industrial era. In addition, Henry Guppy would be pleased to hear that 120 years after his request there has been an attempt to continue his pioneering work of recording Britain’s surname distributions in the late 19th Century.

10. References

Adnan, M., Singleton, A.D., Brunsdon, C., Longley, P.A. 2009. Moving to Real-Time Segmentation: Efficient Computation of Geodemographic Classification. *Proceedings of GISRUK 2009.* 35:41.

Bação F, Lobo V, Painho M. 2004. Clustering Census Data: Comparing the Performance of Self-Organising Maps and K-means Algorithms. KDNet Symposium, Bonn, Germany.

Bação F, Lobo V, Painho M. 2005. Self-Organizing Maps as Substitutes for K-Means Clustering. *Lecture Notes in Computer Science* 3416: 476-483.

Barbujani, G., Sokal, R. 1990. Zones of Sharp Genetic Change in Europe are Also Linguistic Boundaries. *Proceedings of the National Academy of Science, USA.* 87: 1816-1819.

Barbujani, G., Oden, N., Sokal, R. 1989. Detect Regions of Abrupt Change in Maps of Biological Variables. *Systematic Zoology.* 38: 376-389.

Barker, S., Spoerlein, S., Vetter, T., Viereck, W. 2007. *An Atlas of English Surnames*. Peter Lang, Frankfurt.

Barnes, T. 2009. “Not Only...But Also“: Quantitative and Critical Geography. *The Professional Geographer.* 61, 3: 292-300.

Barrai, I., Barbujani, G., Beretta, A., Maestri, I., Russo, A. 1987. Surnames in Ferrara: Distribution, Isonymy and Levels of Inbreeding. *Annals of Human Biology.* 14, 5: 415-423.

Barrai, I., Rodriguez-Larralde, A., Manni, F., Rugggiero, V., Tartari, D., Scapoli, C. 2003. *Annals of Human Genetics.* 68: 1-16.

Barrai, I., Scapoli, C., Beretta., Nesti, C., Mamolini, E., Rodriguez-Larralde, A. 1996. Isonymy and Genetic Structure of Switzerland I. The Distribution of Surnames. *Annals of Human Biology.* 23, 6:431-455.

Batty, M., Longley, P. 1996. Analytical GIS: The Future, in Batty, M., Longley, P (eds) *Spatial Analysis: Modeling in a GIS Environment.* Geoinformation International, Cambridge.

BBC News. 2001. "Blair Gets Cornish Assembly Call". http://news.bbc.co.uk/1/hi/england/1704112.stm (accessed 15/06/09).

Branco, C., Mota-Vieira, L. 2003. Population Structure of São Miguel Island, Azores: A Surname Study. *Human Biology* 75, 6: 929-939.

Branco, C., Mota-Vieira, L. 2005. Surnames in the Azores: Analysis of the Isonymy Structure. *Human Biology*. *77, 1: 37-44.*

Brown, L., Holmes, J. 1971. The Delimination of Functional Regions, Nodal Regions, and Hierachies by Functional Distance Approaches. *Journal of Regional Science*. 11, 1: 57-72.

Bunge, W. 1966. "Gerrymandering, geography and grouping." *Geographical Review* 56, 2: 256-263.

Claval, P. 1998. *An Introduction to Regional Geography*. Blackwell, Oxford.

Cloke, P., Philo, C., Sadler, D. 1991. *Approaching Human Geography.* Sage, London.

Colantonio, S., Lasker, G., Kaplan, B., Fuster, V. 2003. Use of Surname models in Human Population Biology: A Review of Recent Developments. *Human Biology.* 75, 6: 785-787.

Crow, J., Mange, P. 1965. Measurement of Inbreeding from the Frequency of Marriages Between Persons of the Same Surname. *Eugenics Quaterly.* 12:199-203.

Crow, J. 1980. The Estimation of Inbreeding from Isonymy. *Human Biology.* 52:1-14.

Crow, J. 1989. The Estimation of Inbreeding from Isonymy With Update. *Human Biology,* 61, 5/6: 935-948.

Darby, H., 1973. *A New Historical Geography of England.* Cambridge University Press, Cambridge.

Darwin, G. 1875. Marriages Between First Cousins in England and Their Effects. *Journal of the Statistical Society.* 38, 2: 153-184.

De Smith, M., Goodchild, M., Longley, P. 2007 *Geospatial Analysis (2nd ed.).* Matador, Leicester.

Dipierri, J., Alfaro, E., Scapoli, C., Mamolini, E., Rodriguez-Larralde, A., Barrai, I. 2005. Surnames in Argentina: A Population Study Through Isonymy. *American Journal of Physical Anthropology.* 128: 199-209.

Duranton, G., Monastiriotis, V. 2001. Mind the Gap: How Much does Britain's North South Divide Matter? *CentrePiece Spring 2001*. http://cep.lse.ac.uk/centrepiece/v06i1/duranton_monastiriotis.pdf (accessed 10/07/09)

Everitt, B., 1972. Cluster Analysis: A Brief Discussion of Some of the Problems. *British Journal of Psychiatry.* 120: 143-145.

Everitt, B., Landau, S., Leese, M. 2001. *Cluster Analysis 4th Edition.* Hodder, London.

Fotheringham, S., Brunsdon, C., Charlton, M. 2007. *Quantitative Geography Perspectives on Spatial Data Analysis.* Sage, London.

Gordon, A. 1987. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General).* 150, 2: 119-137.

Gordon, A. 1999. *Classification (2nd Edition).* Chapman and Hall, London.

Gilbert, A. 1988. The New Regional Geography in English and French Speaking Countries. *Progress in Human Geography.* 12, 208: 208-228.

Guppy, H., 1890. *Homes of Family Names in Britain.* Harrison and Sons, London.

Grigg, D. 1965. The Logic of Regional Systems. *Annals of the Association of American Geographers.* 55, 3: 465-491.

Grigg, D. 1965. Regions, Models and Classes. In Chorely, R., Haggett., P. (eds). 1965. *Models in Geography.* Methuen and Co, London.

Hall, P. 1975. *Urban and Regional Planning.* Allen and Unwin, London.

Harris R, Sleight P, Webber R. 2005. *Geodemographics: neighbourhood targeting and GIS*. John Wiley and Sons. Chichester, UK:

Hartigan, J. A. and Wong, M. A. 1979. A K-means clustering algorithm. *Applied Statistics* 28**:** 100–108.

Heath, T. 1997. In Wales: Little England holds the key. *The Independent.* Friday 15th August 1997. http://www.independent.co.uk/news/in-wales-little-england-holds-the-key-1245488.html (accessed 15/06/09).

Hey, D. 2000. *Family Names and Family History.* Hambledon Continuum, London

Jobling, M. 2001. In the Name of the Father: Surnames and Genetics. *Trends in Genetics.* 17, 6: 353-357.

Johnston, R. 1968. Choice in Classification: The Subjectivity of Objective Methods. *Annals of the Association of American Geographers.* 58, 3: 579-589.

Johnston, R. 1970. Grouping and Regionalizing: Some Methodological and Technical Observations. *Economic Geography.* 46: 293-305.

Johnston, R., Gregory, D., Pratt, G., Watts, M. 2000. *The Dictionary of Human Geography* (4th Edition). Blackwell, Oxford.

Johnston, R. 1991, *A Question of Place.* Blackwell, Oxford.

Jombart, T. 2008. adegenet: A R Package for the Multivariate Analysis of Genetic Markers. *Bioinformatics* 24: 1403-1405.

Kaplan, B., Lasker, G. 1983. The Present Distribution of Some English Surnames Derived From Place Names. *Human Biology* 55, 2: 243-250.

Keynes, S. 1997. 'The Vikings in England c.790-1016'. In Sawyer, P.(ed). *The Oxford Illustrated History of the Vikings.* 48-82.

Kleiweg, P. 2006. iL04: An interface to RuG/L04, software for dialectometrics and cartography. R package version 1.13. http://www.let.rug.nl/~kleiweg/L04/

Kleiweg, P., Nerbonne, J., Bosveld, L. 2004. Geographic Projection of Cluster Composites. In Blackwell, A., Marriott, K., Shimojima, A. *Diagrams 2004, Lecture Notes in Computer Science.* Springer, New York.

Kwan, M., Schwanen, T. 2009. Quantitative Revolution 2: The Critical (Re) Turn. *The Professional Geographer.* 61,3: 283-291.

Lankford, P. 1969. Regionalization: Theory and Alternative Algorithms. *Geographical Analysis.* 1**:** 196–212.

Lasker, G. 1968. The occurrence of identical (isonymous) surnames in various relationships in pedigrees: A preliminary analysis of the relation of surname combinations to inbreeding. *American Journal of Human Genetics.* 20:250–257.

Lasker, G. 1977. A Coefficient of Relationship By Isonymy: A Method for Estimating the Genetic
Relationship Between Populations. *Human Biology.* 49, 3: 489-493.

Lasker, G. 1985. *Surnames and Genetic Structure.* Cambridge University Press, Cambridge.

Lasker, G., 2002. Using Surnames to Analyse Population Structure. In Postle, D. (ed) *Naming, Society and Regional Identity.* Leopard's Head Press, Oxford.

Lasker, G., Mascie-Taylor, C., 1985. The Geographical Distribution of Selected Surnames in Britain. Model Gene Frequency Clines. *Journal of Human Evolution,* 14: 385-292.

Lasker, G., Mascie-Taylor, C. 1990. *Atlas of British Surnames*. Wayne State University Press.

Laws, E., 1888. *The History of Little England Beyond Wales.* 1995 reprint published by Cedric Chivers Ltd, Bristol.

Longley, P., Webber, R., Lloyd, D. 2007. The Quantitative Analysis of Family Names: Historic Migration and the Present Day Neighborhood Structure of Middlesbrough, United Kingdom. *Annals of the Association of American Geographers.* 97, 1: 31-48.

MacLeod, G., Jones, M. 2001. Renewing the Geography of Regions. *Environment and Planning D: Society and Space.* 19: 669-695.

MacQueen J. 1967. Some Methods for classification and analysis of multivariate observations. *Proceedings from the 5th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley 281-297.

Manel, S., Schwartz, M., Luikart, G., Taberlet, P. 2003. Landscape Genetics: Combining Landscape Ecology and Population Genetics. *Trends in Ecology and Evolution.* 18,4: 189-197.

Manni, F., Guérard, E., Heyer, E. 2004. "Geographic patterns of (Genetic, Morphologic, Linguistic) variation: How barriers can be detected by using Monmonier's Algorithm." *Human Biology* 76, 2: 173-190.

Manni, F. Barrai, I. 2001. Genetic Structures and Linguistic Boundaries in Italy: A Microregional Approach. *Human Biology.* 73, 3: 335-347.

Manni, F. Heeringa, W. Toupance, B. Nerbonne, J. 2008. "Do Surname Differences Mirror Dialect Variation?" *Human Biology* 80, 1: 41-64.

Mascie-Taylor, C., Boyce, A., Brush, G. 1985 in Lasker, G. 1985. *Surnames and Genetic Structure.* Cambridge University Press, Cambridge.

Mascie-Taylor, C., Lasker, G. 1984. Geographical Distribution of Common Surnames in England and Wales. *Annals of Human Biology*. 12, 5: 397-401.

Mascie-Taylor, C., Lasker, G. 1990. The Distribution of Surnames in England and Wales: A Model for Genetic Distribution. *Man.* New Series, 25: 521-530.

Massey, D. 1995. *Spatial Divisions of Labour, Social Structures and the Geography of Production (2nd Ed.).* Palgrave Macmillan, London.

McClure, P. 1979. Patterns of Migration in the Late Middle Ages: The Evidence of English Place-Name Surnames. *The Economic History Review.* 32, 2:167-182.

McElduff, F., Mateos, P., Wade, A., Cortina Borja, M. 2008 What's in a name? The frequency and geographic distributions of UK surnames. *Significance*, 5(4) 189-192

McEvoy, B., Montgomery, G., McRae, A., et al. 2009. Geographical Structure and Natural Selection Amongst North European Populations. *Genome Research*. Published Online 5th March, 2009. http://genome.cshlp.org/content/early/2009/03/05/gr.083394.108.abstract

McQuitty, L. 1957. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. *Educational and Psychological Measurement.* 17:207-229.

Milligan, G., 1980. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika.* 45: 325-342.

Mourrieras, B., Darlu, P., Hochez., J., Hazout., S. 1995. Surname Distribution in France: a Distance Analysis by a Distorted Geographical Map. *Annals of Human Biology* 22, 3: 183-198.

Monmonier, M. 1973. "Maximum-Difference Barriers: An Alternative Numerical Regionalization Method." *Geographical Analysis* 5, 3: 245-261.

Murphy, A. 1991. Regions as Social Constructs: the Gap Between Theory and Practice. *Progress in Human Geography.* 15, 1:22-35.

Nef (New Economics Foundation). 2005. *Clone Town Britain.* http://www.neweconomics.org/gen/uploads/t3zly355dpog3w55ctaiuu4506062005082504.pdf (accessed 08/07/09).

Patten, J. 1976. Patterns of Migration and Movement of Labour to Three Pre-Industrial East Anglian Towns. *Journal of Historical Geography.* 2, 2: 111-129.

Peńa, J., Lozan, J., Larrańaga, P. 1999. An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm. *Pattern Recognition Letters*. 20, 10:1027-1040.

Porteus, J. 1982. Surname Geography: a Study of the Mell Family Name c. 1538-1980. *Transactions of the Institute of British Geographers.* 7, 4: 395-418.

Pooley, C., Turnbull., J. 1998. *Migration and Mobility in Britain Since the 18th Century.* UCL Press, London.

Pudup, M. 1988. Arguments Within Regional Geography. *Progress in Human Geography* 12: 369-390.

R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rodriguez-Larralde, A., Gonzales-Martin, A., Scapoli, C., Barrai, I. 2003. The names of Spain: a study of the isonymy structure of Spain. *American Journal of Physical Anthropology.* 121: 280-292.

Rodriguez-Larralde, A., Pavesi, A., Siri, G., Barrai., I. 1994. Isonymy and the Genetic Structure of Sicily. *Journal of Biosocial Science.* 26: 9-24.

Rodriguez-Larralde, A., Scapoli, C., Beretta, M., Nesti, C., Mamolini, E., Barrai, I. 1998. "Isonymy and the genetic structure of Switzerland. II. Isolation by distance." *Annals of Human Biology* 6: 533-540.

Rogers, A. 1991. Doubts about Isonymy. *Human Biology*. 63, 5: 663-668.

Rosser, Z., Zerjal., T., Hurles, M., et al. 2000. Y-Chromosonal Diversity in Europe Is Clinal and Influenced Primarily By Geography, Rather than by Language. *American Journal of Human Genetics.* 67: 1526-1543.

Scapoli, C., Goebl, H., Sobota, S., Mamolini, E., Rodriguez-Larralde, A., Barrai. I. 2005. Surnames and Dialects in France: Population Structure and Cultural Evolution. *Journal of Theoretical Biology*. 237: 75-86.

Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez-Larralde, A., Barrai, I. 2006. Surnames in Western Europe:A Comparison of the Subcontinental Populations through Isonymy. *Theoretical Population Biology* 71, 37-48.

Schürer, K. 2004. Surnames and the search for regions. Paper presented at *Surnames as a Quantitative Resource,* Centre of Advanced Spatial Analysis. http://www.casa.ucl.ac.uk/surnames/papers.htm (accessed 23/04/2009).

Singleton, A., Longley, P. 2008. Creating Open Source Geodemographic Classifications for Higher Education Applications. *CASA Working Paper 134: http://www.casa.ucl.ac.uk/working_papers/paper134.pdf*

Sokal, R., Harding, R., Lasker. G., Mascie-Taylor, C. 1992. A Spatial Analysis of 100 Surnames in England and Wales. *Annals of Human Biology* 19, 5: 445-476.

Spence, N., Taylor, P. 1970. Quantitative Methods for Regional Taxonomy. *Progress in Geography* 2: 1-64.

Spruit, M., Heeringa, W., Nerbonne, J. 2009. Associations Among Linguistic Levels. *Journal Linguistics* (In Press), doi:10.1016/j.lingua.2009.02.001.

Stoyle, M. 1999. The Dissidence of Despair: Rebellion and Identity in Early Modern Cornwall. *The Journal of British Studies*. 38, 4: 423- 444.

Sykes, B., Irven, C. 2000. Surnames and the Y Chromosome. *American Journal of Human Genetics*. 66: 1417- 1419.

Székely, G., Rizzo, M. 2005. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification.* 22: 151-183.

Tarskaia, L., El'chinova, G., Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez-Larralde, A., Barrai, I. 2009. Surnames in Siberia: A Study of the Population of Yakutia Through Isonymy. *American Journal of Physical Anthropology.* 138:190-198.

Thrift, N. 1994. Taking Aim at the Heart of the Region. In Gregory, D., Martin, R., Smith, G. *Human Geography Society, Space and Social Science.* Palgrave Macmillan, London.

Tobler, W. 1970. "A computer movie simulating urban growth in the Detroit region." *Economic geography* 46, 2: 234-241.

UK Data Archive. *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version)*. Available from: http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=4177

UK Data Archive. *1881 Census for Scotland,* available from*.:* http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=4178

Vickers, D., Rees, P. 2007. Creating the UK National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* 170, 2: 379-403.

Zelinksy, W. 1970. Cultural Variation in Personal Name Patterns in the Eastern United States. *Annals of the Association of American Geographers.* 60, 4:743-769.

Zelinsky, W. 1997. Along the Frontiers of Name Geography. *Professional Geographer.* 49, 4: 465-466.

Ward, J. 1963. "Hierachical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58, 301:236-244

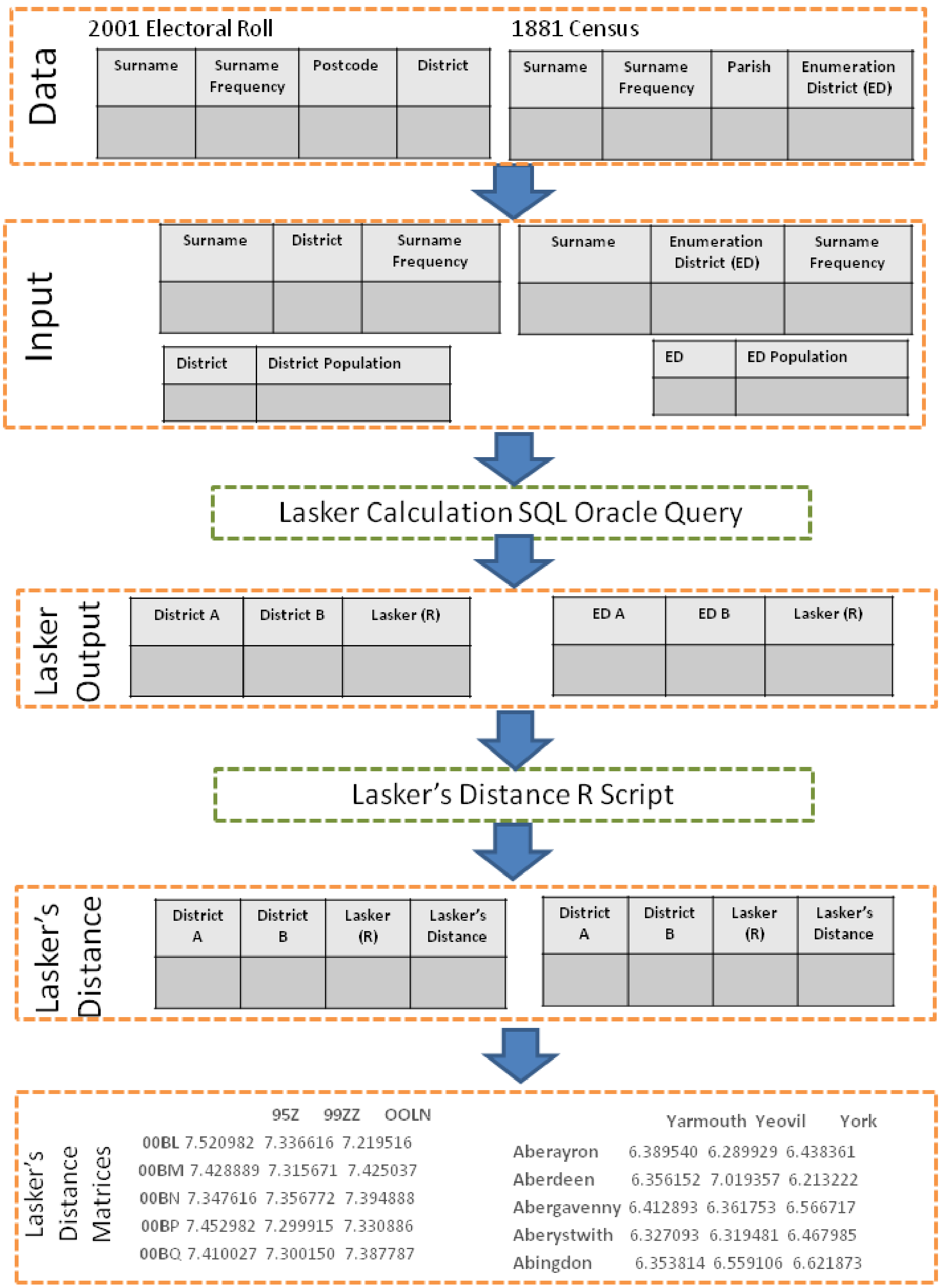
Watkin, M. 1956. ABO Blood groups and racial characteristics in rural Wales. *Heredity* 10: 161- 193.

Wikipedia http://en.wikipedia.org/wiki/Little_England_beyond_Wales. Accessed 26/06/09.

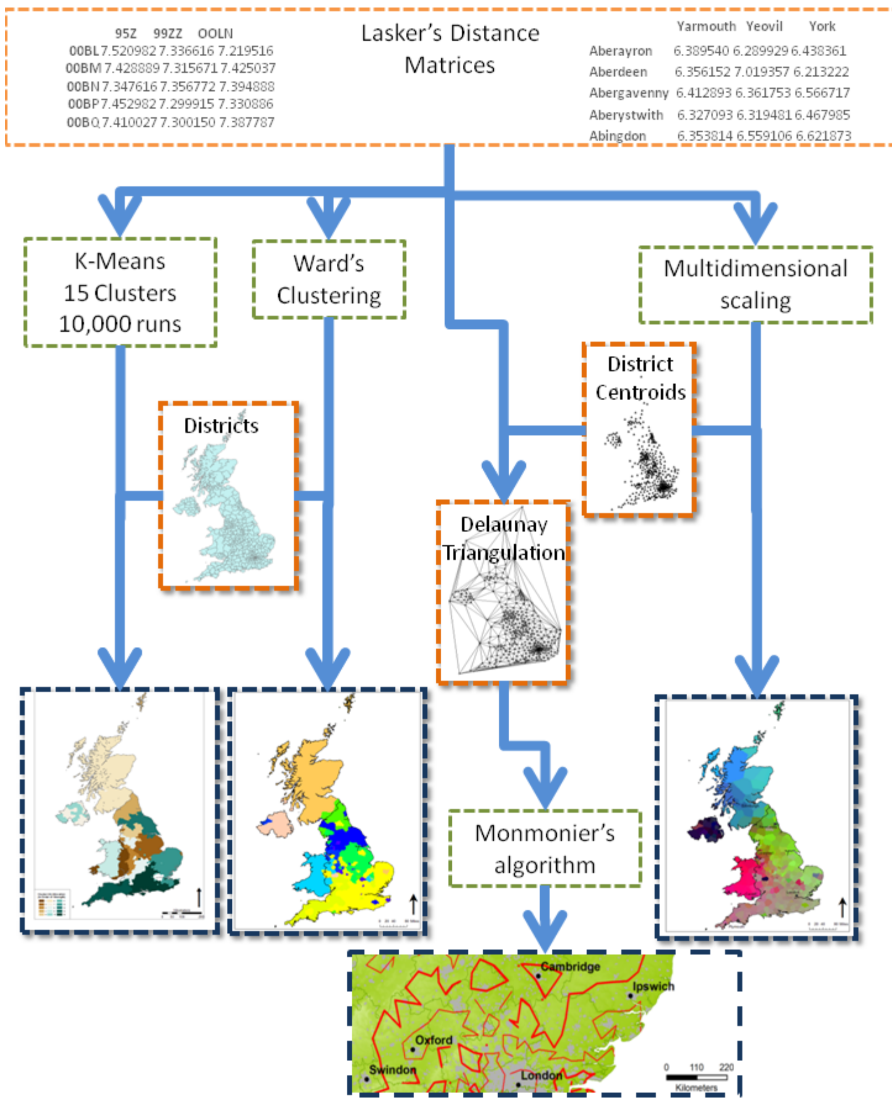
Woollard, M., Allen, M. 1999. *1881 Census for England and Wales, the Channel Islands and the Isle of Man: Introductory User Guide V.04*. Distributed by History Data Service, Data Archive, University of Essex, Colchester.

11. APPENDIX

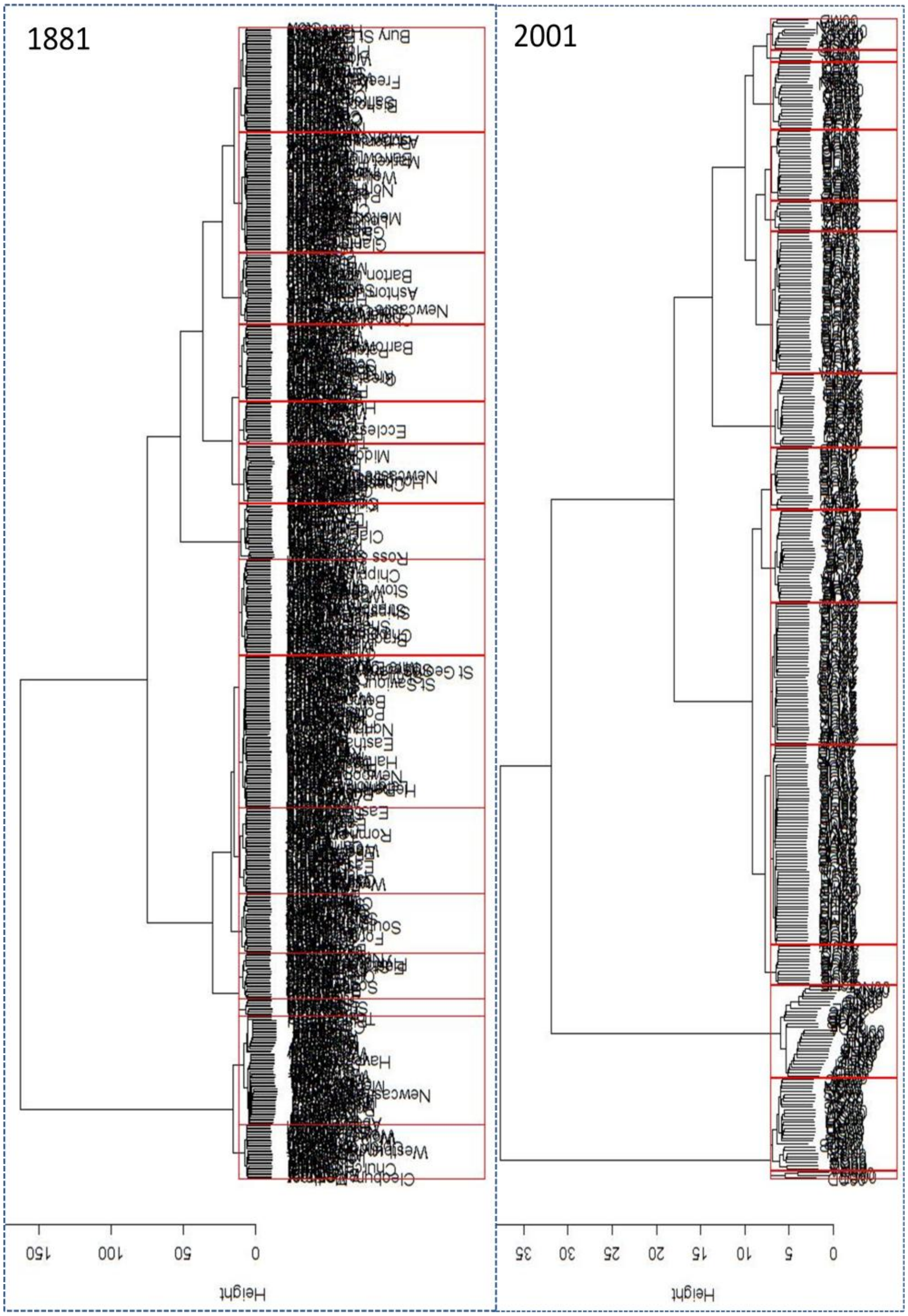
1. A FLOW CHART TO ILLUSTRATE THE LASKER DISTANCE CALCULATION PHASE OF THE METHODOLOGY.



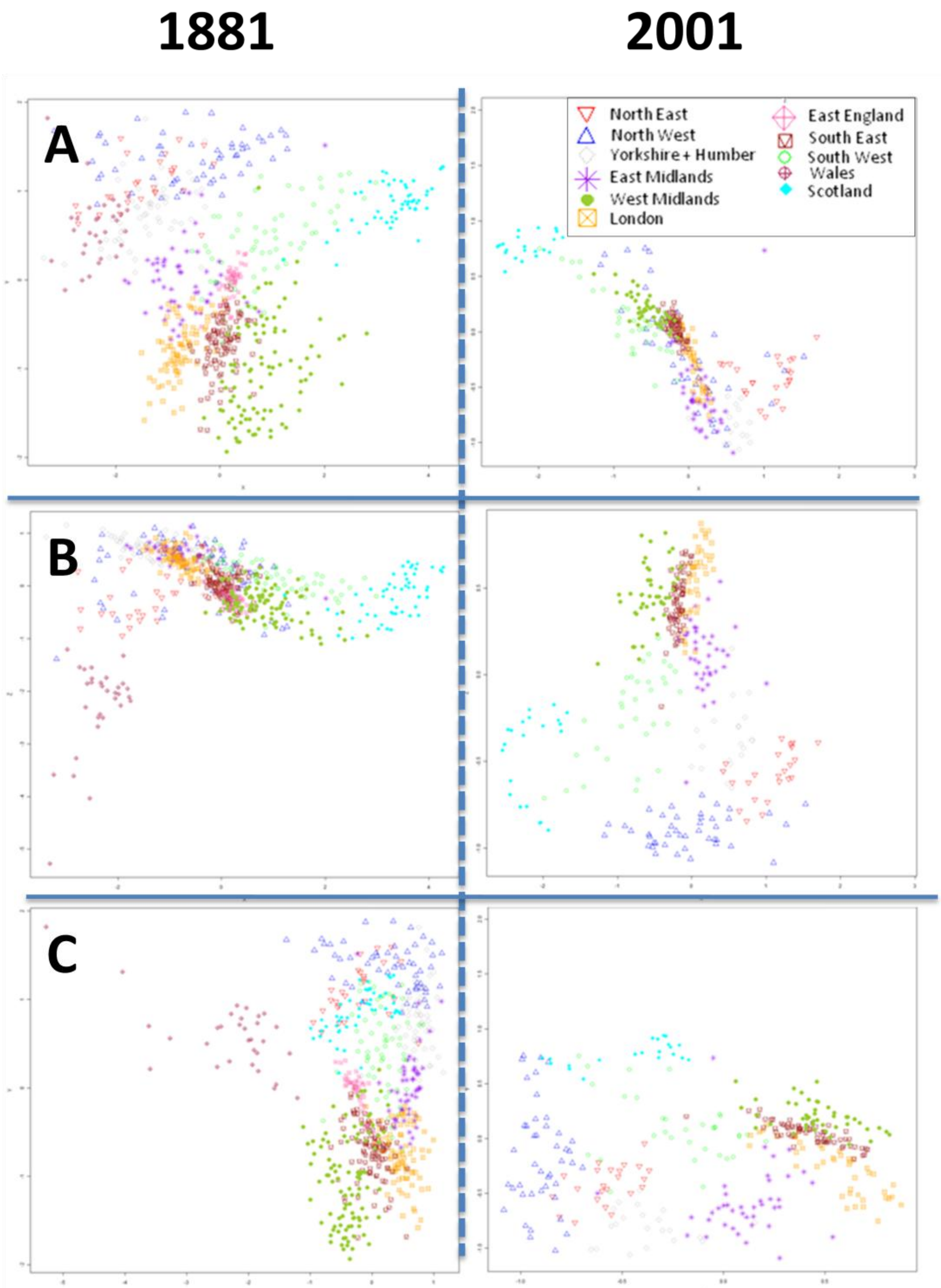
2. A FLOW CHART OUTLINING THE REGIONALIZATION AND VISUALIZATIONS PHASES OF THE METHODOLOGY.



3. DENDROGRAMS ILLUSTRATING THE COPHONETIC DISTANCES BETWEEN CLUSTERS following Ward's clustering of the 1881 (left) and 2001 (right) surnames. The red boxes represent the 15 clusters used to produce Figure 8. The first split of the tree distinguishes England and Scotland from Wales in 1881 and England and Wales from Scotland in 2002. much shorter cophonetic distances in 2001 suggest a move towards more similar regions that are less distinguishable



4. MDS RESULTS PLOTTED ON THE YX (A), ZX (B) AND YZ (C) AXES. The colour and symbol of each point represents the Government Office Region (GOR) that the District falls within. The plots demonstrate the clustering of districts that share a GOR. Districts that are closer together on these plots will be allocated more similar colours in the MDS maps.



5. CATEGORIES USED TO MAP CELTIC AND VIKING SETTLEMENTS.

Celtic Naming Conventions:

'aber'
'afon'
'allt'
'don'
'drum'
'brae'
'caer'
'capel'
‘coed'
'cwm'
'dinas'
'pont'
'bont'
'porth'
'treath'
'ynys'

Viking/ Danish Conventions:

‘thorpe’
‘toft’
‘holme’
‘kirk’
‘kir’
‘thwaite’
‘wick’
‘borough’
‘ness’