The Surname Regions of Great Britain

JAMES A. CHESHIRE, PAUL A. LONGLEY and ALEX D. SINGLETON

Department of Geography, University College London, Gower Street, London, WC1E 6BT, UK; James.Cheshire@ucl.ac.uk.

Abstract

The British Population retains a strong sense of regional identity, epitomized by periodic campaigns for Scottish and Welsh devolution, or for Cornish self-government. There have been few studies into the regionalization of British surnames and none that utilize any register that can claim to be nationally representative. The National Social Map presented in this paper is the first comprehensive attempt to create a regional geography of Great Britain based upon the clustering of surnames. The resulting map illustrates a strong relationship between the populations surnames and geographic location. The homogeneity within each of the surname regions identified is striking given that spatial contiguity constraints were not included within the clustering process. The map will hopefully set a bench-mark for future work by geographers in the field of surname research.

(Received 12th September 2009; Revised 25th January 2010; Accepted 10th February 2010)

ISSN 1744-5647
doi:10.4113/jom.2010.1103

1. Introduction

The British Population retains a strong sense of regional identity, epitomized by periodic campaigns for Scottish and Welsh devolution, or for Cornish self-government. At the same time, recent years have seen a phenomenal increase in amateur genealogy, providing testimony to the individual quest to maintain a sense of heritage and place in our increasingly globalised world (MacLeod and Jones, 2001). The emergence of geo-genealogy is also manifest in interest in web mapping sites: for example, the UCL WorldNames profiler (http://www.publicprofiler.org/worldnames) attracted over 1.7 million unique visitors within 12 months of being launched, providing testimony to the significant worldwide public interest in the cultural and geographical significance of surnames. In Great Britain most surnames have followed the male line and therefore can be used as a proxy for genetic information (King and Jobling, 2009).

To date, with the exception of the works of Mascie-Taylor and Lasker (for example, 1984), there have been few studies into the regionalization of British surnames and none that utilize any register that can claim to be nationally representative. The National Social Map presented in this paper is the first comprehensive attempt to create a regional geography of Great Britain based upon the clustering of surnames.

2. The Origins of Surnames and their Significance

It is unclear precisely when hereditary surnames became formalised in Great Britain, and their adoption appears to have been later in some areas of the country. There is evidence of the regular recording of surnames by the onset of the 13^{th} Century, although no suggestion that all surnames were hereditary (McClure, 1979). The strongest evidence of adoption of hereditary surnames exists from the 15^{th} Century onwards in England (Lasker and Mascie-Taylor, 1985) and the 16^{th} Century in Scotland (Barker et al., 2007).

Surnames can be categorised into local surnames (toponyms), occupational surnames (metonyms), surnames of relationship (diminutives), or nicknames (Barker et al., 2007). The existence of regionalized concentrations of surnames is well-known, most obviously in the case of specific instances of toponyms, but also with respect to the general categories diminutives are particularly common in Wales, for example, and nicknames pertaining to roles in local pageants remain common in East Anglia. The majority of Anglo Saxon surnames remain spatially concentrated in the areas where they first came into popular usage (Sokal et al., 1992; Hey, 2000). For this reason the spatial characteristics of surnames have become of increasing interest to a range of academics in fields ranging from Genetics to Epidemiology, Linguistics and Geography (Colantonio, 2003; Mateos and Tucker, 2008; Zelinsky, 1997).

Guppy (1890) was perhaps the pioneer of regional surnames analysis in Britain, and viewed the geographical distribution of surnames as important primarily in devising a regional, rather than political taxonomy of the British population. There are a number of interesting applications that immediately suggest themselves. Using surnames one could, for example, establish whether the Welsh political boundary matches the extent of the Welsh population. Surnames can also provide an indication as to the mobility and diversity within a population. A high proportion of a relatively small number of surnames suggests an isolated population with low mobility and little inter-mixing with the surrounding populations (McElduff et al., 2008). Mapping provides insight into the cause of the isolation, such as topographic or cultural barriers. The concept of barriers to population flow is widely explored within genetics; surnames provide a useful, albeit crude, method of inferring information about the genetic diversity and origins of a population (King and Jobling, 2009).

Surname maps can therefore provide useful insights into important spatial characteristics of the British population. The Social Map presented in this paper represents a tentative look at creating a new regional geography of Great Britain and can be used as a basis for further investigation into how closely surnames match other population characteristics, such as genetic groups.

3. The Great Britain Surnames Database

The map was produced using data from a 2001 Enhanced Electoral Register (EER) supplied by CACI (London, UK). The core data comprise the names and addresses of UK residents aged 17 or over who were (or were about to become) eligible to vote in UK or European Elections and who did not opt out of inclusion in the publicly available version of the register. In addition, the EER includes further data that are sourced from commercial surveys and credit scoring databases, supplementing information on those adults who opted out of the public register, or who were not registered to vote. It should be noted that the EER is likely to under-enumerate non-UK citizens, particularly those originating outside the EU because of their ineligibility to vote in any UK elections. The sources and operation of resulting biases represents a wide ranging issue that lies beyond the scope of this paper. However, in the context of producing a map highlighting regions formed by the enduring stability of British social structure, the largely international names that are omitted are disproportionately likely to create noise that disrupts the observed patterns. Despite the limitations, these data represent the 45.6 million people resident in the UK during October 2001. The surname frequency plot (see map) shows that there are very few names possessed by more than 100,000 people. In fact the majority of names (1.45 million) have less than 10 occurrences; creating a distribution with a high negative skew.

The map presents the distinctive regional clustering within the national population. Each individual within the surname database is assigned to an administrative district, thus providing the building blocks of a national map. There are 410 local authority or administrative districts in Great Britain (pre-April 2009), each containing an average population of approximately 105,000 people. Those districts that lie within Greater London together comprise Great Britains only world city (Taylor, 2001), and this focus of national and international immigration presents a profoundly un-British surname mix with, for example, the highest incidence of unique surnames in Britain (McElduff et al., 2008). This can be seen in the right hand inset of the presented map. The 32 London Boroughs have therefore been aggregated into a single district; leaving 378 districts in the final analysis.

4. Measuring Surname Isonymy

The Lasker Distance, derived from a coefficient of isonymy, is a measure that determines the similarity in surname structure between two geographical areas (Colantonio, 2003). Isonymy refers to the possession of the same surname (Lasker, 1977) and forms a basic premise in genetics that individuals with the same surname are more likely to share the same family lineage, and as such, isonymy also indicates a biological relatedness (Lasker, 1985). The coefficient of isonymy therefore compares the probability isonymy occurring between members of two geographic areas (Fox and Lasker, 1983). It is calculated as:

$$R_i = \frac{\sum S_{i1} S_{i2}}{2 \sum S_{i1} \sum S_{i2}} \tag{1}$$

where S_{i1} is the number of occurrences of the ith surname from area 1 and S_{i2} is the number of occurrences from the same surname from area 2 (Lasker, 1985). In order to compare the surname characteristics of each location relative to all the others in a country, the Lasker coefficient (R_i) values are converted to a distance measure. The Lasker Distance measures the isonymic distance between each pair of locations. This metric produces a distance matrix between all pairs of locations in the study area, in which distances are symmetric with a zero principal diagonal. The Lasker Distance is calculated as:

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

where L is the Lasker distance and i and j are two separate areas. One can think of the Lasker Distance as a measure of similarity, or difference, between two areas in surname space (Rodrguez-Larralde et al., 2008). The greater the Lasker Distance the less similar the composition of surnames in the two geographic areas.

The formation of larger regions from multiple districts was achieved by grouping districts based on their Lasker distance from the surrounding districts using the K-means clustering algorithm. K-means is an iterative clustering algorithm that assigns each data point into one of K clusters until convergence to a local minimum of its objective function (Everitt, 1972; Singleton and Longley, 2008). Here the objective function is the sum of squared Euclidean distance (within sum of squares) between each data point and its nearest cluster centroid. K observations, or seeds, are selected from the data at random then each of the remaining observations are provisionally assigned to their nearest seed. The centroids of the resulting clusters become the new seeds and the process is repeated. If the composition of subsequent clusters changes, the centroid is recalculated. This process continues until the total within sum of squares across all clusters is minimized or a specified number of iterations are reached (Singleton and Longley, 2008).

To ensure the best practical outcome, 10,000 clustering runs were completed and the cluster allocations with the lowest total within sum of squares, and therefore tightest clusters, across all the groupings was selected. The resulting output of the clustering procedure comprised a table listing each district and its resulting cluster allocation. A subjective decision was made in relation to K number of clusters used to partition the data. The clustering process was implemented for K= 5 to K=20 clusters. Increasing the values of K resulted in further partitioning of the broad regions shown in the final map. In the context of representing broad regions and informed by the preliminary research documented in Cheshire et al. (2009), K=10 represented the optimum in terms of parsimony in representation and making the maps readily intelligible to users.

5. Cartography and Map Production

The inherently spatial nature of surnames makes mapping the most effective method of communicating their distributions. The little interest in surnames shown by geographers has left investigation into, and subsequent mapping of, the spatial distributions of surnames neglected. Therefore, as a demonstration of the important contribution geographers can make to contemporary surname research, and in response to Smith's (2005) call for increased emphasis on visualization, sound cartographic principles were central to the production of the map presented here.

Each district was assigned a color based on its cluster allocation and was then mapped. Given the focus of (Guppy, 1890) early surnames analysis upon physical geography factors, contextual information relating to relief has also been included, sourced from NASA Shuttle Radar Topography Mission (http://www2.jpl.nasa.gov/srtm/) data. This context is important as physical barriers, such as rivers, may, in the past, have limited migration and therefore the mixing of surnames and the effects of these might still be expected to be evident in contemporary surname distributions. Present day Government Office Region (GOR) boundaries have been included in addition to those of districts, in order to aid interpretation and description of the underlying surname patterns.

6. Discussion

The map illustrates a strong relationship between district surname structure and geographic location. The homogeneity within each of the surname regions identified is particularly interesting given that spatial contiguity constraints were not included within the clustering process. As one would expect, Scotland and Wales are identified as clear surname regions in Britain. The close correspondence between Scotland and the contemporary administrative border suggests an abrupt transition in surname types. The Welsh surname region (in pink) shows a close match to the contemporary border, however there appears greater interaction between English and Welsh surnames. In blue is a peripheral Welsh region that enters the West Midlands and North West, it also includes southern Wales and Pembrokeshire (areas known to have high proportions of English surnames). A further region, similar in shape and area, in brown, suggests a extended Welsh influence into the North West, West Midlands and South West. The core Welsh surname border fails to reach the town of Audlem, a town that famously voted to become Welsh (BBC News, 2008). The map therefore suggests that, based on the populations surnames, Audlem's allegiances are misguided.

South West England is identified as a separate region. With the exception of larger settlements the districts of the South East and the East of England appear to have similar surname compositions. London has been clustered with Birmingham and number of larger settlements, such as Leicester, across England, something that may be by the large numbers of ”non-British” names found in these places. Highlighting London and other large cities as uncharacteristic of the rest of Great Britain concurs with geodemographic classifications such as the Output Area Classification (OAC) (Vickers and Rees, 2007).

It is beyond the scope of this work to provide explanations for the patterns in surname distributions revealed by this analysis. However, the results presented here provide a firm basis for continued research into generalising patterns of surname distribution and hypothesis generation. There are many possibilities for future study, such as: a focus on more local patterns in surname regions; examination of apparent anomalous results such as the linked clusters appearing within London and Birmingham; or screening out the records that are added through enhancement of the Electoral Roll (since these will be disproportionately non-voters with non-traditional British names). Furthermore, investigation is required to examine scaling effects within the data, such as a consideration of the most appropriate geographic units to base these studies upon. Finally, by examining temporal surname records it could be established whether the processes behind the creation of the observed clusters in surname structure have accelerated or decelerated over time: this would, of course, require a number of new population datasets from different time periods.

7. Conclusions

This Surname Map of Great Britain provides compelling evidence that there is a regional geography underpinning the contemporary geography of surnames in Britain. The facts that contiguity constraints were not included in the K-means clustering, but that a clear spatial pattern has emerged is a strong indication of inherent regionality. On the basis that many surnames occur most frequently at their place of origin, the regions highlighted in the map reflect historical patterns and processes.

The map is the first systematic attempt to produce a comprehensive and high quality cartographic representation of British surname geography, and will hopefully set a bench-mark for future work by geographers in the field of surname research.

Software

The Great Britain Surname Regions social map was produced using a number of software packages. The dataset required the use of Oracle Database software for storage of the Electoral Roll and calculation of the Coefficient of Isonymy. The resulting table was small enough to be processed as a single object using the **R** software platform (http://www.r-project.org). **R** was used to calculate the Lasker Distance matrix and complete the K-means clustering using the *stats* package (available in the basic **R** installation). The resulting table of administrative districts and their cluster allocations was exported as a `.csv` file and joined in ArcGIS 9.3 to a shapefile of the district boundaries. The final map was produced in ArcGIS 9.3 with the ColorBrewer (http://colorbrewer2.org/) colour palettes and Maplex Labeling Engine.

Acknowledgements

This work was completed as part of James Cheshires ESRC CASE funded PhD in association with ESRI (UK). Paul Longley is a co-investigator on the GENESIS Node of the ESRC National Centre for E-Social Science. Alex Singleton is principal investigator of an ESRC funded First Grant (RES-061-25-0303).

References

BARKER, S., SPOERLEIN, S., VETTER, T. and VIERECK, W. (2007) An Atlas of English Surnames, Peter Lang, Oxford.

BBC NEWS (2008) English Village Votes to be Welsh [Online]. Available from: http://news.bbc.co.uk/1/hi/wales/7364464.stm, [Last accessed: 25 August, 2009].

CHESHIRE, J., MATEOS, P. and LONGLEY, P. A. (2009) Family Names as Indicators of Britain's Regional Geography [Online]. Available from: http://www.casa.ucl.ac.uk/working_papers/paper149.pdf, CASA Working Paper Series, 149.

COLANTONIO, S. E. (2003) Use of surname models in human population biology: a review of recent developments, Human Biology, 75, 785–807.

EVERITT, B. S. (1972) Cluster Analysis: A Brief Discussion of Some of the Problems, The British Journal of Psychiatry, 120, 143–145.

FOX, W. R. and LASKER, G. W. (1983) The Distribution of Surname Frequencies, International Statistical Review / Revue Internationale de Statistique, 51, 81–87.

GUPPY, H. (1890) The Homes of Familly Names in Britain, .

HEY, D. (2000) Family names and family history, Hambledon and London, London.

KING, T. E. and JOBLING, M. A. (2009) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution, 25, 351–360.

LASKER, G. and MASCIE-TAYLOR, C. (1985) The geographical distribution of selected surnames in Britain. Model gene frequency clines, Journal of Human Evolution, 14, 385–392, doi:10.1016/S0047-2484(85)80046-4.

LASKER, G. W. (1977) A coefficient of relationship by isonymy: a method for estimating the genetic relationship between populations, Human Biology; an International Record of Research, 49, 489–493.

LASKER, G. W. (1985) Surnames and Genetic Structure, Cambridge University Press, Cambridge.

MACLEOD, G. and JONES, M. (2001) Renewing the geography of regions, Environment and Planning D: Society and Space, 19, 669–695, doi:10.1068/d217t.

MASCIE-TAYLOR, C. and LASKER, G. (1985) Geographical distribution of common surnames in England and Wales, Annals of Human Biology, 12, 397–401, doi:10.1080/03014468500007951.

MATEOS, P. and TUCKER, K. (2008) Forenames and Surnames in Spain in 2004, Names: A Journal of Onomastics, 56, 165–184, doi:10.1179/175622708X332860.

MCCLURE, P. (1979) Patterns of Migration in the Late Middle Ages: The Evidence of English Place-Name Surnames, The Economic History Review, 32, 167–182.

MCELDUFF, F., MATEOS, P., WADE, A. and BORJA, M. C. (2008) What's in a name? The frequency and geographic distributions of UK surnames, Significance, 5, 189–192, doi:10.1111/j.1740-9713.2008.00332.x.

RODRGUEZ-LARRALDE, A., PAVESI, A., SCAPOLI, C., CONTERIO, F., SIRI, G. and BARRAI, I. (2008) Isonymy and the genetic structure of Sicily, Journal of Biosocial Science, 26, doi:10.1017/S0021932000021027.

SINGLETON, A. and LONGLEY, P. A. (2008) Creating open source geodemographic classifications for Higher Education applications [Online]. Available from: http://www.casa.ucl.ac.uk/working_papers/paper134.pdf, CASA Working Paper Series, 134.

SMITH, M. (2005) The Journal of Maps: an electronic journal for the presentation and dissemination of map based data, Journal of Maps, v2005, 1–6, doi:10.4113/jom.2005.39.

SOKAL, R., HARDING, R., LASKER, G. and MASCIE-TAYLOR, C. (1992) A Spatial Analysis of 100 Surnames in England and Wales, Annals of Human Biology, 19, 445–476.

TAYLOR, P. (2001) Specification of the World City Network, Geographical Analysis, 33, 181–194.

VICKERS, D. and REES, P. (2007) Creating the UK National Statistics 2001 output area classification, Journal of the Royal Statistical Society: Series A (Statistics in Society), 170, 379–403, doi:10.1111/j.1467-985X.2007.00466.x.

ZELINSKY, W. (1997) Along the Frontiers of Name Geography, Professional Geographer, 49, 465–466.