Regionalisation and Clustering of Large Spatially-Referenced Population Datasets: the Case of Surnames

J. A. Cheshire[1], P. A. Longley [2], Pablo Mateos[1]

[1]University College London Dept. of Geography, Gower Street, London, UK, WC1E 6BT.
Email: james.cheshire@ucl.ac.uk

[2]Centre for Advanced Spatial Analysis, University College London, 1-19 Torrington Place, London, UK, WC1E 7HB.
Email: plongley@ucl.ac.uk

1. Introduction

Family names (surnames) have very distinctive geographies, many of which are only now being investigated following developments the processing of large georeferenced datasets. This paper considers the use of surnames as a basis to regionalise Europe. The analysis is both data rich and computationally intensive, entailing as it does the aggregation, clustering and mapping of close to 6 million surnames. The resulting regionalisation can be used to infer cultural, linguistic and genealogical information about the European Population.

2. Data

The 5.95 million unique surnames, taken from a list of in excess of 400 million people from 16 countries, are drawn from the UCL World Names database (www.publicprofiler.org/worldnames/). The georeferenced surname data were originally derived from publicly available population registers and telephone directories from the 2000-2005 period. Two levels of Nomenclature of Territorial Units for Statistics (NUTS) geography, NUTS 1 and NUTS 2, were used in this study in an attempt to standardize the population size in each spatial unit. A list of the 16 countries and their respective administrative geographies is provided in Table 1.

3. Methods

Surnames are commonly mapped individually or in grouped according to a shared characteristic (for example, according to presence of a patronymic ‘s’ suffix). Such maps are inadequate for large scale, generalized regionalization. It would, for example, be impossible for an individual to process the 5.95 million maps required to analyse the spatial patterns of every European surname. This paper sets out to develop a generalised portrayal of European surname geography.

Geneticists have long been interested in surnames as they provide an exploitable link to genetics (King and Jobling, 2009). From this research the Coefficient of Isonymy has been developed and provides a method of aggregating the information contained within the spatial locations of millions of surnames into a similarity matrix (Lasker, 1977). The Coefficient of Isonymy establishes the extent to which the same name (isonymy) occurs between the populations of two or more spatial units.

It can be defined as:

$$Ri = \sum_{i} piA\ piB$$

(1)

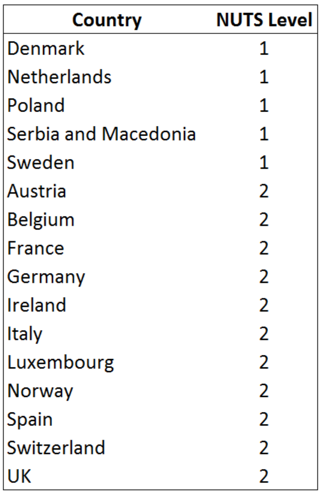
where p_{iA} is the relative frequency of the i^{th} surname in population A and p_{iB} is the relative frequency of the i^{th} surname in population B. Little similarity between very diverse populations will result in very small Coefficients of Isonymy that are hard to interpret. The Lasker Distance (Rodriguez-Larralde et al. 1994) attempts to remedy this and is defined as:

$$L_{iAB} = -\ln(R_{iAB})$$

(2)

The Lasker Distance is simply a (dis)similarity index where the values between pairs of spatial units can be thought of as distance in “surname space” such that larger values between spatial unit pairs represent greater difference in surname composition. This matrix provides a convenient input for subsequent analysis.

Table 1. A list of the countries with the level of granularity used in this study.



Country	NUTS Level
Denmark	1
Netherlands	1
Poland	1
Serbia and Macedonia	1
Sweden	1
Austria	2
Belgium	2
France	2
Germany	2
Ireland	2
Italy	2
Luxembourg	2
Norway	2
Spain	2
Switzerland	2
UK	2

Ward’s (1963) grouping algorithm is a popular method of hierarchical agglomeration and used here to summarise the Lasker Distance matrix into clusters, or regions. The procedure forms hierarchical groups of mutually exclusive subsets in attribute space. It does this by minimising the increase (which is proportional to the squared Euclidean distance between cluster centres) in total within-cluster variance (Székely and Rizzo, 2005). The algorithm begins by assigning the *n* initial number of observations to (*n* – 1) exclusive sets by considering the union of all possible *n*(*n* – 1)/2 pairs and selects the combination that minimises within-cluster variance, before repeating the process in subsequent iterations (Ward, 1963). The resulting clusters are not necessarily optimal because the "best route" between the clusters has the priority and may only be achieved at the expense of a minor reduction in the individual clusters' homogeneity. As with other hierarchical classifications (see Gordon, 1999), a dendrogram can be used to illustrate the relationships between observations. All of the observations are joined together at the “trunk of the tree” with nodes joining branches that lead to the observations (in this case the NUTS regions). The length of these branches (cophenetic distances) indicates the strength of the relationship between the observations. Joining the clustering outcome to the spatial boundary data enables the allocations to be shown as a choropleth map. Justification for the use of Ward's clustering in this context is provided by Cheshire et al. (2009).

Selecting an appropriate number of clusters, 18 in this study, remains a subjective process (Johnston (1970) and Gordon (1999)). It was given careful consideration and informed by the configuration of the dendrogram, the cophenetic distance between NUTS spatial units and the plausibility of the mapped allocations. The "plausibility" of the cluster is the most subjective criterion but is arguably the most important. Too many clusters could create a regionalisation outcome that suggests a more diverse surname geography than reality, equally too few clusters would suggest homogeneity where there is diversity.

In addition to hierarchical clustering, multidimensional scaling (MDS) was used to provide an effective summary of the degree to which surnames registered in the same country are clustered in relative space. Following Golledge and Rushton's (1972) pioneering work, MDS has found many spatial analysis applications (Gatrell, 1981). MDS reduces the dimensionality of a data set from an *m* x *n* (dis)similarity matrix with a large value of *n* to a matrix with few values of *n* that can be treated as coordinates in relative rather than absolute space. Our application here maps the NUTS areas of the dissimilarity matrix into a space of minimum dimensionality such that the distances in that space closely match the observed dissimilarities (Gatrell, 1981). MDS can either be metric or non-metric; both seek a regression of the distances on the dissimilarities with the former utilising the numerical values of the dissimilarities and the latter their rank-order. There are a number of criteria available around which to base the optimal reduction in dimensionality. In this study, guided by the visual interpretability of the results, we use MDS in 2 dimensions. MDS undertaken for greater than 2 dimensions had little impact on the positioning of the NUTS regions in relative space and becomes increasingly hard to visualize on paper.

4. Results and Discussion

Figures 1 and 2 demonstrate the regionality of European surnames in subtly different ways. The result of mapping the Ward's cluster allocations, Figure 1 illustrates the spatial extent of the regions and their similarity to linguistic and national areas. Figure 2 illustrates the degree of similarity within each country. Tighter the distributions of points in the plot indicate shorter closer Lasker Distances between each country's NUTS regions. The unsurprising nature of many of the surname regions highlighted (judged by their conformity to well-known national and linguistic boundaries) provides strong evidence that the inductive approach of this study, as demonstrated through its data and methods, is appropriate when attempting to establish the existence of regional patterns in Europe's surname distributions.

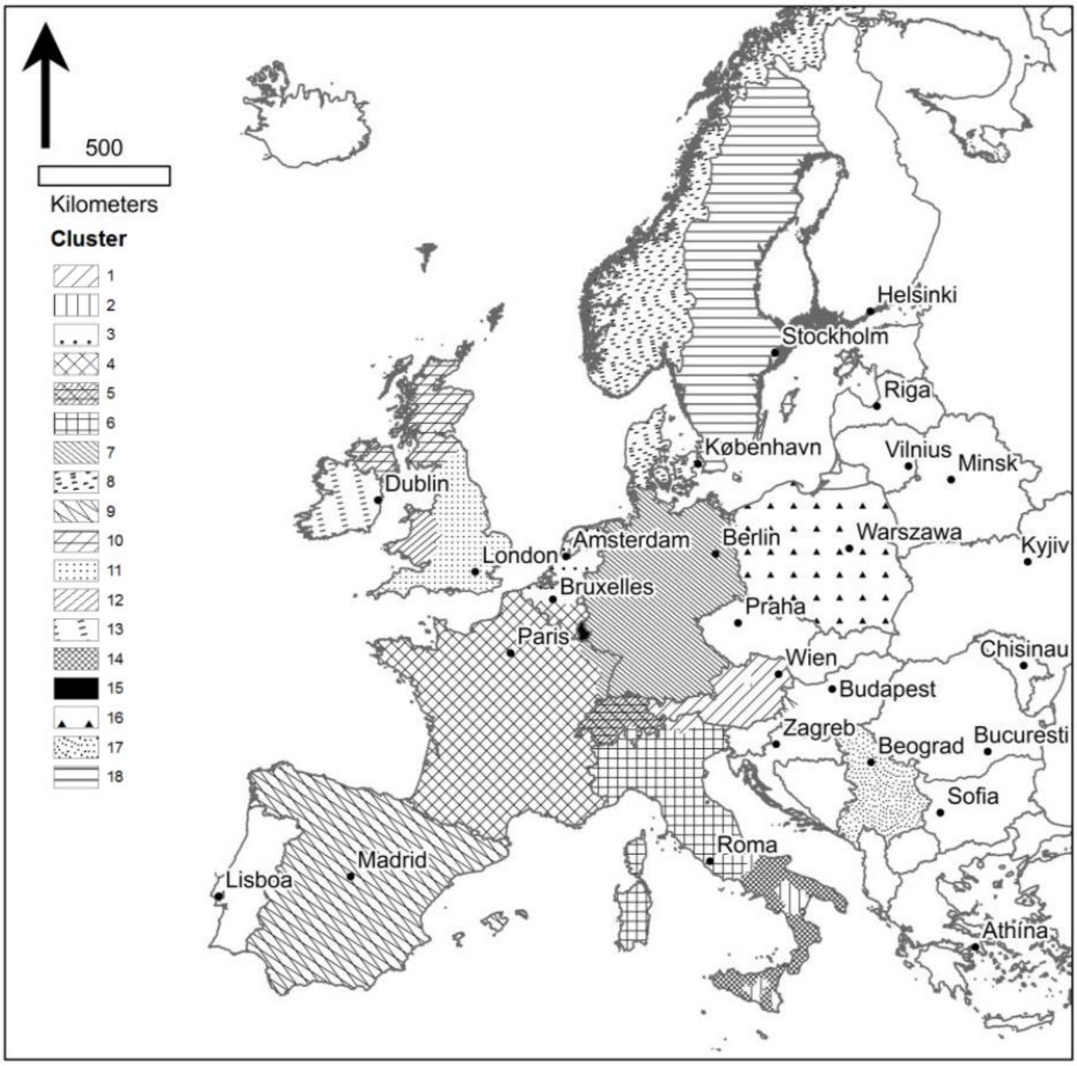


Figure 1. A map of the 18 cluster allocations produced from the Ward's Hierarchical Clustering of Lasker Distances. Each allocation is represented as a unique pattern. Cophenetic distances between adjacent clusters can be large, as is the case between Poland and Germany, or relatively small such as between England and Wales. Areas of no-data are white.

It is acknowledged that Figures 1 and 2 merit more detailed discussion, however the focus in this paper is primarily methodological. While the methods used in this study are already proven and established, this paper is particularly innovative in their deployment by using such a large and novel geographic dataset. Confidence in the methods, and their parameterisation, is increased by the close resemblance of Figure 1 to well-established linguistic regions. A number of potential limitations, however, have been identified and will be the focus of future research. For example, the Lasker Distance and its subsequent clustering outcomes are sensitive to the number of spatial units into which a population is segmented. Each spatial unit is assigned an equal weighting in the analysis regardless of its contributing population. Thus two areas with equal contributing populations will not have equal influence if one has been spatially partitioned more than the other. Implications include the allocation of more clusters to areas of relatively uniform surname compositions but relatively large numbers of spatial units in addition to the distortion of the MDS space such that the other spatial units have less space to occupy. This effect may explain separate cluster allocations for

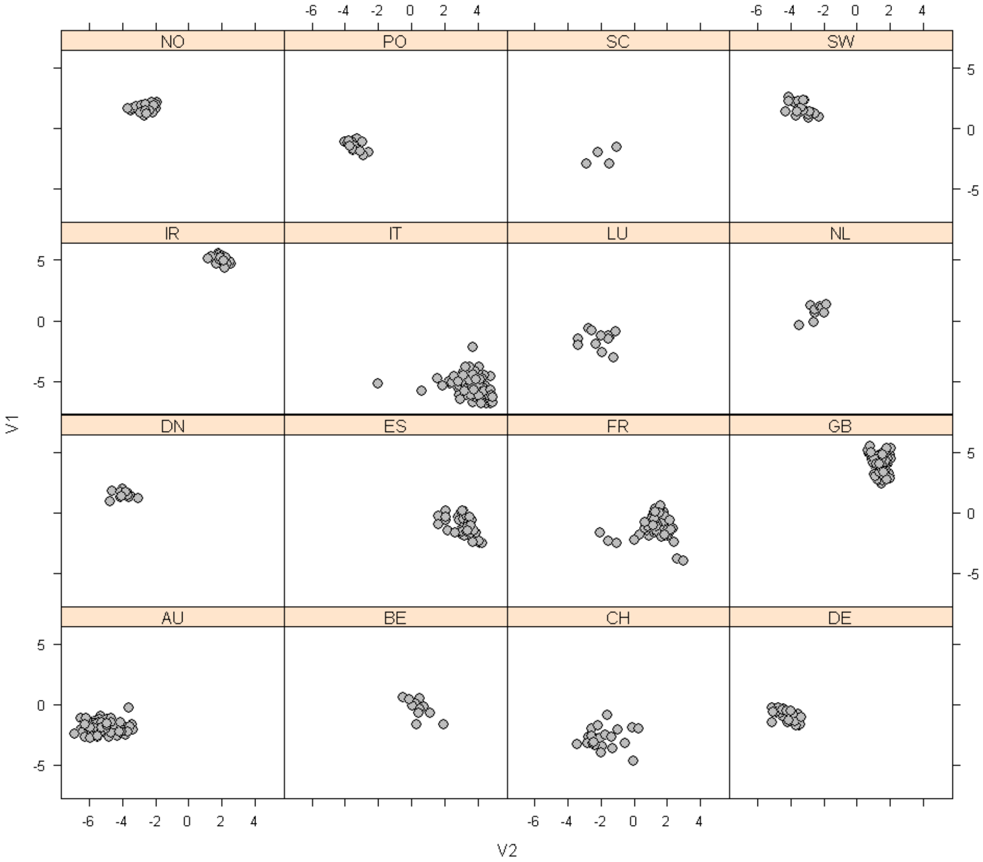


Figure 2. Plots produced from the 2-dimensional MDS for each of the 16 countries. From top left the countries are: Norway (NO), Poland (PO), Serbia and Macedonia (SC), Sweden (SW), Ireland (IR), Italy (IT), Luxembourg (LU), Netherlands (NL), Denmark (DN), Spain (ES), France (FR), United Kingdom (GB), Austria (AU), Belgium (BE), Switzerland (CH), Germany (DE).

England, Scotland and Wales in Figure 1 but apparent uniformity Spain, a country well known for linguistic differences. Future work will seek to apply a weighting to insure that smaller populations are assigned a lower weighting in the classification than larger populations. In addition, a sensitivity analysis should be undertaken to establish the extent to which individual spatial units can alter the cluster outcome.

In conclusion, this paper has sought to demonstrate the utility of an inductive approach to summarising and analysing large population datasets, the outcomes of which can provide the basis to hypothesis generation about social and cultural patterning and the dynamics of migration and residential mobility.

References

Cheshire, J., Mateos, P., Longley, P. 2009, Family Names as Indicators of Britain's Changing Regional Geography. *CASA Working Paper 149.* Available from http://www.casa.ucl.ac.uk/publications/workingpapers.asp

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

Gatrell, A. C. 1981. Multidimensional Scaling. In Wrigley, N., and Bennett, R. J., *Quantitative Geography.* Routledge, Oxon.

Golledge, R. G., Rushton G. 1972. Multidimensional Scaling: Review and Geographical Applications. *Association of American Geographers Commission on College Geography, Technical Paper No. 10.*

Gordon, A. 1999. *Classification.* CRC Press, Florida.

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

King, T., Jobling, M. 2009. What's in a name. Y chromosomes, surnames and the genetic genealogy revolution. *Trends in Genetics.* 25, 8: 351-360.

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

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

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