Delineating Europe's Cultural Regions: Population Structure and Surname Clustering

Authors:

James Cheshire*

Pablo Mateos

Paul A. Longley

Affiliation

Department of Geography, University College London, United Kingdom

*james.cheshire@ucl.ac.uk

Keywords: Surnames; Consensus clustering; Lasker Distance; Europe

Abstract:

Surnames (family names) show distinctive geographical patterning and remain an underutilised source of information about population origins, migration and identity. In this paper we investigate the geographical structure of surnames in 16 European countries through the use of the Lasker Distance, consensus clustering and multidimensional scaling. Our analysis is both data rich and computationally intensive, entailing as it does the aggregation, clustering and mapping of 8 million surnames collected from 152 million individuals. The resulting regionalisation demonstrates the utility of an innovative inductive approach to summarising and analysing large population datasets across cultural and

geographic space, the outcomes of which can provide the basis to hypothesis generation about social and cultural patterning and the dynamics of migration and residential mobility in Europe. The research also contributes a range of methodological insights for future studies concerning spatial clustering of surnames.

1. Introduction

Family names, also known as surnames, are widely understood to provide good indicators of the geographic, ethnic, cultural and genetic structure of human populations. This is mainly because surnames were 'fixed' in most populations several centuries ago, and their transmission over generations (mostly patrilinearly) typically conforms to socio-economic, religious and cultural characteristics (Smith, 2002) as well as geographical constraints (Manni *et al.*, 2004). The outcome is a variety of spatial patterns that manifest processes of biological inheritance (Lasker, 1985) and intergenerational inheritance of culture (Cavalli-Sforza and Feldman, 1981). The vast literature in this area is principally concerned with analysing population structure in surname frequency distributions at national or sub-national levels (for a review see Colantonio *et al.*, 2003). Here we are solely concerned with how such population structure is manifest across space, rather than between religious, ethno-cultural or social groups *per se*. One of the primary methodological concerns of these studies is the development of: a) adequate measures of surname relatedness - or surname distance - between localities or regions and b) areal classification algorithms to partition space according to such distances. In this paper we seek to make two contributions to this line of research; first we investigate the geographical structure of surnames at a continental level in 16 European countries, and second we consider a relatively new regional clustering technique at this pan-European scale. In so doing we draw upon expertise developed population genetics and geography. The result is a cultural regionalisation of Europe based purely on the geography of surname frequencies that is key to the search for Europe's cultural regions. We use techniques derived from population genetics to devise and cluster measures of surname distance between populations, and use regionalisation concepts and

spatial database skills from geography to structure millions of address records and map the results.

Our analysis is both data rich and computationally intensive, entailing as it does the aggregation, clustering and mapping of 8 million surnames collected from 152 million individuals. The resulting regionalisation can be used to infer cultural, linguistic and genealogical information about the European population over the preceding centuries, for example with a view to design a genetic sampling framework.

2. Cultural and surname distance between areas

Surnames first appeared in Europe during the Middle Ages (Hanks, 2003) and can be characterised by frequency distributions within a population that are driven by initial population size, rate of endogamy between populations and socio-cultural preferences within a group's reproduction patterns. Such processes are in turn a product of demographic, geographic, ethno-cultural and migration factors. One of the most striking and recurrent findings of surname research is that, in spite of the relative mobility of modern populations, surnames usually remain highly concentrated in or around the localities in which they were first coined many centuries ago (e.g. Longley *et al.*, 2011). The size of the databases available for the study of surnames have been increasing in line with the computational resources required to process them (see Scapoli *et al.*, 2007, Cheshire *et al.*, 2010, Longley *et al.* 2011). Such advances enable the continued progression of surname research in the context of the many exemplary studies outlined below.

Following the early work of Cavalli-Sforza and colleagues using Italian telephone directories in magnetic tape form in the 1970s (see Piazza et al. (1987) and Cavalli-Sforza *et al.,* (2004)),

the increasingly wide availability of digitally encoded names registers has led to a host of studies of the surname structure of populations of individual countries. Throughout, one group has dominated this research through the publication of a succession of national-level surname analyses: their studies include Austria (Barrai *et al.*, 2000); Switzerland (Rodriguez-Larralde *et al.*, 1998); Germany (Rodriguez-Larralde *et al.*, 1998); Italy (Manni and Barrai, 2001); Belgium (Barrai *et al.*, 2004); the Netherlands (Manni *et al.*, 2005); and France (Mourrieras *et al.,* 1995, Scapoli *et al.*, 2005). More recently they have amalgamated these findings for eight countries in Western Europe, analysing a sample of 2094 towns and cities, grouped into 125 regions (Scapoli *et al.*, 2007). This study found clear regionalisation patterns in surname frequency distributions, closely matching the national borders for eight countries, but also highlights anomalies arising from the historical geography of languages.

Whilst being wide-ranging, both geographically and in terms of the number of surnames sampled, the work by Scapoli *et al.* (2007) is still limited to the partial sampling of "representative" locations. The motivation for the work reported here is to expand this work in methodological terms by including more European countries (16 in this paper), to use data representing complete populations (i.e. without sampling), and to use new classification algorithms in the form of consensus clustering to delineate cultural surname regions and barriers to population interactions over space.

The remainder of this paper is organised as follows: first, we outline the choice of surname distance metric used in this analysis; second, we review the most commonly used regional classification algorithms and suggest a new methodological approach; third, we present the materials and methods used in the analysis; and fourth, we present and discuss the resulting

surname regionalisation of Europe and the benefits and challenges of the proposed methodology.

2.1. Measuring surname distance

Interest in the relationship between surnames and genetic characteristics first emerged in the late 19th century when George Darwin (1875) – son of Charles and himself offspring of first cousins – used surnames to calculate the probability of first cousin marriages in Britain. Little further research was undertaken until In the 1960s when Crow and Mange (1965) proposed a probability of relatedness defined as the frequency of repetition of the same surname, known as isonymy (Lasker, 2002). In addition to applications to the study of inbreeding between marital partners or social groups, isonymy can be also be used to establish the degree of relatedness between two or more population groups at different geographic locations (Smith, 2002). It is this latter, regional, interpretation of isonymy that has gained greater currency over the last decades and is the one used here. The coefficient of isonymy extends the idea of monophyly (sharing a single common ancestor) between two populations and is defined by Lasker (1985) 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). This coefficient is based on the similarity of the surname frequency distribution between two populations. In the two region case, isonymy is calculated as:

$$R_{AB} = \sum_{i} \frac{p_{iA}\, p_{iB}}{2} \qquad (1)$$

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*. In many cases, especially when comparing international populations, the similarity between population groups is very small and this creates very small values of R_{AB}. Therefore, a more meaningful transformation of this measure, termed the *Lasker Distance* (Rodriguez-Larralde *et al.*, 1994) is used here. It is defined as:

$$L_{AB} = -\ln(2R_{AB}) \quad (2)$$

where $R_{AB} = (p_{iA} \times p_{iB})/2$. The inverse natural logarithm creates a more intuitive measure that can be thought of as distance in surname space such that larger values between populations suggest greater differences between them (that is, less commonality in their surnames). Scapoli *et al.* (2007) suggest that this measure can be used to isolate differences in cultural inheritance because two populations that are genetically homogenous, yet distant in Lasker distance terms are likely to exhibit subtle differences in cultural behaviour.

Doubts about the value of isonymy studies are founded upon the fundamental assumptions that they entail. An implicit assumption is that at some previous generation each male had a unique (monophyletic) surname, and that all surnames were first coined synchronically in the same generation (Rogers, 1991). We know this not to be the case in several countries, for example in Great Britain, where for a multitude of reasons permanent surnames were acquired gradually in a number of distinctive and separate sub- populations. The name 'Smith', for example, describes an occupation found within every community across the country and hence resulted in a heterophyletic surname. However, it is also the case that

even if two populations with very similar surname distributions do not share unique common ancestors, they are nevertheless much more likely to be genetically related to one another, in comparison with a population that has a very different surname makeup (Lasker, 1985).

One important alternative to the Lasker Distance was proposed by Nei (1973). His measure of genetic distance, originally intended for the study of allele similarities between populations (Nei, 1978), has been applied to surnames as *Nei's distance* of isonymy in a number of studies (such as Scapoli *et al.*, 2007). Others have also successfully used the measure (see Manni *et al.,* 2008 and Manni *et al.*, 2006) and found it less sensitive to heterophyletic surnames and also likely to be more correlated with geographic distance. The purpose of this paper is to propose an innovative set of clustering techniques across a large number of countries. Therefore, it was thought best to avoid comparisons of multiple established distance measures and focus our clustering efforts on a single surname (dis)similarity measure so as to keep this aspect of the analysis fixed and concentrate on clustering and representational issues. On the basis that the work presented here is an extension of previous national level studies with the Lasker Distance (see Longley *et al.,* 2011 and Cheshire, 2011) the authors felt most comfortable using the measure here. The intention is to conduct further research into the utility of dissimilarity measures from both population genetics and demographics more widely and the Nei's distance will form part of this.

2.2. Delineating surname regions: consensus clustering and MDS

As the result of the studies outlined above, the similarities between frequency distributions of surnames and the genetic structure of populations across space are now quite well known. However, there continues to be an important research gap with respect to the most appropriate spatial analysis techniques to automatically detect the geographical patterns of surname distributions at various scales. In population genetics, most studies posit clinal

transitions in genetic characteristics punctuated by abrupt barriers (Lasker and Mascie-Taylor, 1985). In contrast, and with a few exceptions, surname geography research is usually founded upon discrete administrative areal building blocks, and as such produces valid generalisations for only a presepecified range of scales. We are not the first to apply clustering and data reduction techniques to surnames (in addition to the studies listed above, see Chen and Cavalli-Sforza (1983)) but we hope to improve on previous research by suggesting a good compromise between the continuous and discrete representations of space by using two areal classification methods: consensus clustering and multidimensional scaling (MDS). The former creates discrete groupings of prespecified areal units whilst the latter, when used to inform areal colour values on a map, can produce a more continuous representation of population change over space.

Indicating the certainty of a clustering outcome is an important aspect of population geography research, especially in regionalisation. Readers should refer to Kaufman and Rousseeuw (1990) and also Gordon (1999) for a review of these. Of direct relevance here is Nerbonne *et al.*'s (2008) use of the aggregate data matrices produced in dialectometrics as a basis for identifying linguistic regions. The certainty of such regions were determined through bootstrapping and composite clustering techniques and visualised both as a dendrogram and composite cluster maps. In the former, each branch has information about the number of times a particular grouping between its sub-branches occurred, while in the latter lines between geographic regions were drawn with increasingly dark shading, corresponding to the number of times contiguous spatial units on both sides of the line were assigned to different clusters. Using a different approach but with a similar cartographic effect, Manni *et al.* (2004) and Manni *et al.* (2006) implement Monmonier's (1973) boundary

algorithm to detect dissimilarities between contiguous regions. The mapped results of both methodologies do not require the assignment of all spatial units to a particular cluster, but the objective is to identify only the most abrupt boundaries.

In this paper consensus clustering (Monti *et al.* 2003) was chosen as a promising method of creating a robust cluster outcome, consistent with providing a number of metrics to indicate the optimal number of clusters and the certainty associated with each cluster assignment. Such metrics are useful because they give context to the final clustering outcome: in particular they address the issue that, contrary to what many surname regionalisation maps suggest, not all resulting clusters are equally probable to occur within the data.

Consensus clustering, first proposed by Monti *et al.* (2003), is a relatively new method for class discovery. It has become increasingly popular in the genetics literature - Monti *et al.* (2003) is highly cited - and there are a number of papers, such as Grotkjær *et al.* (2005), that compare its effectiveness to other more established clustering methods. The underlying hypothesis states that items consistently grouped together are more likely to be similar than those appearing in the same group less frequently (Simpson *et al.* 2010). The method is designed to increase the stability of the final cluster outcomes by taking the consensus of multiple runs of a single cluster algorithm. Simpson *et al.* (2010) have provided an extension to this approach, called merged consensus clustering, by enabling the cluster assignments to be the product of multiple runs of multiple algorithms or kinds of data. By merging the results from different algorithms it is thought that the confidence in the result will increase because the limitations of one clustering algorithm will be offset by the strengths of another. For example Ward's hierarchical clustering is sensitive to outliers in the data, but offers a stable solution over-all in terms of consistency of cluster outcome; by contrast, the over-all

arrangement of K-means clusters is relatively unstable, but the solutions are less sensitive to outliers. In addition to the increased stability of the results, consensus clustering can provide a range of metrics to help inform the optimum number of clusters as well as the robustness of the resulting cluster outcome in terms of its structure and the membership of individual clusters.

Before undertaking the merged consensus clustering procedure, the user has to select the clustering algorithms to be used. Theoretically there is no limit on the number of algorithms that contribute to the final result aside from the practical constraints related to computation time and the degree to which the result will actually improve. Some of the most popular data classification methods are Ward's hierarchical clustering (Ward, 1963), K-means (Hartigan and Wong, 1979), partitioning around medoids (PAM) (see Kaufman and Rousseeuw (1990)) and self-organising maps (SOM) (Kohonen, 1990). The algorithms selected for this study are listed under the analysis section below. Table 1 shows the definitions of the variables and the algorithms used – the latter are adapted from Monti *et al.* (2003) to make them more applicable in this context.

<- Table 1 about here ->

Consensus clustering first samples the complete dataset D to create a new subset $D^{(h)}$ before clustering using the specified algorithm(s). The sampling (using methods such as bootstrapping) and clustering are repeated multiple times in order to gauge sensitivity to repeat sampling from the total number N of randomly selected geographic units e_i. The results from each iteration are stored in a consensus matrix $\mathcal{M}$, which records for each possible pair of e_i the proportion of the clustering runs in which they are both clustered together. The consensus matrix is derived by taking the average over the connectivity

matrices of every perturbed dataset (Monti *et al*., 2003). The entries to the matrix are defined in the following way:

$$M^{(h)}(i,j) = \begin{cases} 1 \text{ if } i \text{ and } j \text{ belong to the same cluster} \\ 0 \text{ otherwise} \end{cases} \quad (3).$$

$D^{(h)}$ is the (*N x N*) connectivity matrix, required to keep track of the number of iterations in which both geographic units are selected by resampling, such that its (*i,j*)th entry is equal to 1 if both *i* and *j* are present in $D^{(h)}$, and 0 otherwise. According to Monti *et al.* (2003) the consensus matrix $\mathcal{M}$ is the normalised sum of the connectivity matrices of all the perturbed datasets $\{D^{(h)}: h = 1,2, \dots, H\}$:

$$\mathcal{M}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)} \quad (4).$$

The *i,j*th entry in the consensus matrix records the number of times the two items have been assigned to the same cluster divided by the number of times both items have been selected (sampled). It therefore follows that a perfect consensus result would produce a matrix containing only 0s and 1s. $\mathcal{M}$ in essence provides a similarity measure to be used in further clustering or agglomerative hierarchical tree construction (Simpson *et al.* 2010).

To create a merged result a *merge matrix* provides a way of combining the outcomes multiple methods by weighted averaging their respective consensus matrices (Simpson *et al.* 2010). The weighting can be adjusted to increase or decrease the influence of certain

clustering methods. The advantage of this approach is that it mitigates the issues associated with the different classification properties in each of the algorithms discussed above.

Two types of clustering robustness measures can be calculated. The first relates to the cluster structure (called *cluster consensus* $m(k)$) and the second to the cluster membership (called *item consensus* $m_k(i)$). In regionalisation problems, the latter is especially useful because it enables the comparative visualisation of the geographic unit's cluster allocations alongside their summary measures of cluster robustness. As is often the case, a geographic unit may only be assigned to a particular cluster on the basis that all units have to be assigned to one of the set of clusters. Where allocations are marginal, there will be low confidence in the allocation and it can therefore be interpreted accordingly. Monti *et al.* (2003) first define I_k as the set of indices of items (geographic units in this case) belonging to cluster k. This can then be used to define the cluster's consensus as the average consensus index between all pairs of items belonging to the same cluster.

$$m(k) = \frac{1}{N_k(N_k - 1)/2} \sum_{\substack{i,j \in I_k \\ i<j}} \mathcal{M}(i,j) \tag{5}$$

The corresponding item consensus for each item e_i and each cluster k is defined as:

$$m_i(k) = \frac{1}{N_k - 1\{e_i \in I_k\}} \sum_{\substack{j \in i_k \\ j \neq i}} \mathcal{M}(i,j) \tag{6}$$

where 1 {cond} is the indicator function that is equal to 1 when cond is true and 0 when false. Item consensus $m_i(k)$ measures the average consensus index between e_i and all other items (geographic units) in cluster *k.* In the case of perfect consensus across all runs, the cluster consensus would be 1 for each cluster. As is demonstrated in the results section, this measure provides the level of confidence in the final result, expressed as a function of the number of times a geographic unit has been assigned to a particular cluster.

The use of multiple classification methods across a range of cluster values enables consensus clustering to provide a number of metrics to help inform the selection of the optimal number of clusters. Monti *et al.* (2003) state that the true number of clusters (*k*) can be estimated by finding the value of *k* at which there is the greatest change in the *empirical cumulative density function* (CDF) calculated from the consensus matrix $\mathcal{M}$ across a range of possible values of *k.* If the unique elements of $\mathcal{M}$ are placed in descending order, it is possible to define the CDF*(c)* over a range c=[0,1] using the following equation.

$$CDF(c) = \frac{\sum_{i<j} 1\{\mathcal{M}(i,j) \leq c\}}{N(N-1)/2} \quad (7).$$

It is then possible to calculate the area under the curve (*AUC*) of CDF as follows:

$$AUC = \sum_{i=2}^{m} [x_i - x_{i-1}] CDF(x_i) \quad (8)$$

where x_i is the current element of the *CDF* and m is the number of elements. If every iteration from the consensus clustering identifies the same groups then the $\mathcal{M}$ elements will

be either 0 or 1, and thus *AUC*= 1. This provides the benchmark against which to compare the clustering results. One can experiment with the number of clusters into which to group the data, ranging from values between K=2 to K_{max} and compare their results with the benchmark *AUC*=1 result. The result with the number of groups that comes closest to this can therefore be considered the optimum number of clusters. To establish the best outcome the quantity ΔK is calculated, which is the change in *AUC* as k varies. The optimal k value is broadly considered to coincide with the peak in ΔK. Using Simpson *et al.'s* (2010) merged method the resulting consensus matrices (one from each cluster method used) from the optimal k are combined through weighted averaging. The merged matrix maintains the same properties as a consensus matrix and can therefore be used as a dissimilarity matrix for re-clustering.

In addition to the identification of discrete surname regions we also use multidimensional scaling (MDS) to show more subtle and continuous differences that depict trends or surfaces of closeness or dissimilarity between populations. MDS provides an effective summary of the degree to which regions are related to each other in 'surname space'. Following Golledge and Rushton's (1972) pioneering work, MDS has found many spatial analysis applications (Gatrell, 1981). MDS reduces the dimensionality of a (dis)similarity matrix of *m* rows by *n* columns with a large value of *n,* to one with very few values of *n.* In geographic applications, the dissimilarity matrix between areas can be converted through MDS into a space of minimum dimensionality (typically two or three dimensions or number of *n*) closely matching the observed (dis)similarities in the data (Gatrell, 1981). MDS can either be metric or non-metric; both seek a regression of the distances on the (dis)similarity matrix with the former utilising the numerical values of the (dis)similarities and the latter their rank-order. For its application in this paper, we acknowledge Manni *et al.*'s (2004) concerns that MDS

(like principal components analysis) does not provide a statistical analysis of the pattern of change, instead portraying an interpolated landscape in geographic space. This, of course, differs little from the maps produced by Lao *et al.* (2008), or Cavalli-Sforza (2000), which rely on spatial interpolation techniques to infer genetic characteristics in areas where samples have actually not been taken. This, in part, is the reason why we adopt a mixed approach here by combining MDS with cluster analysis in order that one set of results can provide context to the other.

3. Materials and methods

3.1. Data and geography

The UCL Worldnames database (see worldnames.publicprofiler.org) contains the names and addresses of more than 400 million people in 26 countries, derived from a range of publicly available population registers and telephone directories collected since 2000. For purposes of this paper, surname data for 16 European countries in Worldnames were extracted – more than 8 million unique surnames – along with their geographical locations and frequencies of occurrence. A list of countries, name frequencies and geographical characteristics is shown in Table 2. The countries used in this study reflect those available in the Worldnames database, and thus omissions reflect an inability to source the requisite data, rather than a deliberate exclusion of particular countries.

<- Table 2 about here ->

The ongoing assembly of this database is a major ongoing enterprise, involving the acquisition, normalisation, cleaning and maintenance of publicly available telephone directories and commercial versions of public registers of electors. The extract used in this

paper comprises a commercially enhanced version of the 2001 Electoral Register for the UK and landline telephone directories from the remaining countries identified as current during the period 2001-2006. There are many potential sources of bias in these sources, and some are likely to be systematic in their operation. Non-electors (of different types) are likely to be under-represented in the UK data, for example, and such individuals are more likely than average to bear names recently imported from abroad. Landline rental is likely to introduce some socioeconomic and geographic bias in some European countries, while the bearers of some names may be more likely to withhold their telephone numbers from public directories than others. These are all complicated issues to address and thus, in order to expedite analysis, we have taken the decision to accept the data in the form in which they were supplied to us. We view the time period as helpful in sustaining this decision, in that it predated the period of mass mobile phone ownership, which may have reduced the penetration of land line services amongst some groups, and the heightened privacy concerns that are leading to attrition in the size of the public version of the UK Electoral Register.

Selection and calibration of appropriate spatial units is a key problem in geographical research (Openshaw, 1984) and one that requires much more thorough consideration in the population genetics literature. In order to analyse Europe's surname regions we first had to adopt a geographical unit of analysis that was as consistent as possible throughout the study area. The international nature of the Worldnames database means that it contains data at geographic scales ranging from an individual's address through to name frequencies within administrative areas. Individual addresses have been carefully geocoded to a set of geographical coordinates (latitude and longitude) at levels of resolution ranging from the

building level to the street, postcode, city, metropolitan area and administrative region. Since this study is concerned with general patterns at the pan European level we are interested in aggregating detailed locations onto a set of standard geographical regions of similar size and population. European Union (EU) NUTS regions (*Nomenclature d'Unités Territoriales Statistiques*) provide a convenient set of geographic units that broadly conform to these aims. NUTS are a standard referencing system for the hierarchy of five levels of administrative sub-divisions of EU countries for statistical purposes, ranging from broad country regions (NUTS 1) to municipalities (NUTS 5). Initially all surname data were aggregated to NUTS 3 level (the province or department), but it subsequently became apparent that some countries with relatively large numbers of NUTS 3 units relative to their population sizes (such as Germany where these correspond to 429 *Kreise* or Districts) were having an undue influence on the results. This was especially evident at the clustering and MDS stages of the analysis. Therefore, for this study we have opted for a combination of NUTS 2 and NUTS 3 regions in an attempt to address this problem and to produce a set of homogeneous areas in terms of population size and geographical extent. In so doing, we follow common practice in geographical analysis of NUTS data in Europe. Table 2 identifies the NUTS level selected for each country and the number of areal units. This resulted in a total number of 685 geographic units across the 16 countries.

3.2. Analysis

Our analysis consisted of applying consensus clustering and MDS to the 685 geographic units. The analysis was implemented using the statistical software *R* (R Development Core Team, 2010); in particular the consensus clustering required the *clusterCons* package, developed by Simpson *et al.* (2010). The package is a new release and designed primarily for

gene expression microarray data and we provide its first documented use in the context of population genetics/ geography.

A matrix of the Lasker Distances between all pairs of NUTS geographic units provided the input for the *clusterCons* package. Consensus clustering was performed using three different algorithms: K-Means, partitioning around medoids (PAM) and Ward's hierarchical clustering. These were chosen for their success in previous studies (see Cheshire *et al.* 2010, Longley *et al.* 2011). In order to select the most appropriate number of clusters (*K*) in which to group the geographic units, each of these algorithms was run using *K* values ranging between 5 and 45. For each value of K, subsampling was used to provide 200 selections for each algorithm in the consensus clustering. The results of this process produced a merged consensus matrix – an average of the three consensus matrices (one for each clustering methodology) – for each value of *k* (resulting in the creation of 40 matrices). The merged consensus matrices provided the basis for the ΔK calculations, the results of which are shown in Figure 1.

Figure 1 shows a dramatic decrease in ΔK values between K=5 and K=12, fluctuating between 12 and 20 before stabilising after K=21. Solely on the basis of Monti *et al.*'s (2003) number of clusters criterion (outlined in Section 1.2.3.) 10 should have provided the best outcome. It was however decided to relax this criterion and select 14 clusters for a number of reasons.Firstly, this does not exceed a practical number of clusters for visualising regions in a choropleth map and secondly it makes intuitive sense as it approximates the number of countries used in this analysis and hence it is likely to capture the most significant interactions between countries. We did trial a number of results with more clusters but we found, as predicted by Monti et al. (2003), that random clusters can be created with the

consensus clustering methodology if the stopping criterion moves beyond the highest ΔK values. The results with *K*> 14 thus contained some questionable regional groupings within countries. The picture at *K*> 9 but < 14 appeared too generalised when mapped (although was more stable) for the purposes here.

<- Figure 1 about here ->

Figure 2 shows a box plot with the robustness values associated with the final cluster structures at 14 clusters (as outlined in Equation 5). In addition to the results from clustering the final merge matrix, those from the non-merged consensus clustering are also included for comparison. In agreement with preliminary research using different data (Cheshire and Adnan, 2011), the merge matrix result produced higher median robustness values across all algorithms when compared with the non-merged results. Overall, based on Figure 2, it was thought that PAM on the merge matrix produced the most robust cluster structure. Although, the PAM inter-quartile range was greater than that for Ward's algorithm, six of the 'Ward clusters' (nearly half) were classified as outliers. The membership robustness values were also highest, on average, for the PAM clustering result: these have been mapped alongside the final cluster outcome in Figure 3.

<- Figure 2 about here ->

In this study, guided by the visual interpretability of the results, we also use MDS in two and three dimensions. MDS undertaken for greater than three dimensions had little impact (see stress values in Figures 4 and 5) on the positioning of the NUTS regions in relative space and becomes increasingly hard to visualise effectively in print. Results from the MDS are shown in two ways. Figure 4 shows a conventional plot of the results from two-dimensional MDS

for each country, where each dot represents a NUTS region and each axis each of the two MDS dimensions. Figure 5 is a more novel representation, previously used in linguistics (see, Nerbonne (2010)) and shows the three-dimensional MDS values on a 2-D map. In this figure the raw MDS coordinates have been rescaled to values between 0 and 255 in order that they can be substituted for a value in the Red, Green, Blue (RGB) colour model. Each separate component is mapped onto one of these colours (Dim. 1= red, Dim. 2=green, Dim. 3= blue) before all three are combined into a single map to produce the colour map in Figure 5. We consider this to be a particularly effective, although not perfect, method of visualising MDS results as it demonstrates both continuous and abrupt changes in structure.

<- Figure 3 about here ->

Finally, in order to measure the effect of 'isolation by distance', Figure 6 plots for each of 234,270 possible pairs of spatial units their geographic distance (measured as Euclidean distance in kilometres from the NUTs centroids) against their *Lasker Distance* in surname space (Equation 2). The same type of plots is also separately repeated for each country and shown in Figure 7.

4. Results

This section presents the key results of the analysis presented above with the general objective of describing the geographical patterns found and offering some insights into the performance of the classification and visualisation methods used. The specific methodological aspects derived from these results will be discussed in the next section.

4.1. Isolation by distance

The scatterplot in Figure 6 hints at a relationship between Lasker Distance and geographic distance across Europe, although the strength of this relationship may be less forceful than could have been expected from general knowledge. This can be attributed to the fact that Euclidean distance fails to reflect well-known physical barriers to movement, such as coastlines and mountain ranges that facilitate or impede movement. The mean Lasker Distance across Europe is 10.45 with the maximum value (19.68) occurring between Northern Ireland and southern Italy, hinting at a measure of isonymy with a low dispersion across Europe compared to geographic distances.

<- Figure 4 about here ->

At the country level, the relationship between surname and geographical distance presents some interesting and particular national trends, as shown in Figure 7. Multilingual countries, such as Belgium and Switzerland, unsurprisingly show the strongest relationship between geographic distance and differences in the surname composition of its regions. Counter-intuitively perhaps, the plot for Norway suggests that surname diversity increases with proximity. This is most probably because of the greater surname diversity (resulting from domestic and international migration) in urban areas that are close to one other in the southwest of the country. This diversity appears to be sufficiently strong and in close proximity, managing to offset the more distant but more homogenous rural areas. In countries such as Denmark, a de-facto archipelago, Euclidean distance does not reflect actual population interaction. Moreover, the plots in Figure 7 provide an important indication of the sub-national interactions between distance and surname diversity.

<- Figure 5 about here ->

4.2. Consensus Clustering

The clustering results shown in Figure 3A conform to many well-known national and linguistic divisions across Europe, and most notably, follow linguistic or historical political boundaries, in some cases reflecting the effects of contemporary global migration to large urban areas.

The clusters generally follow national borders, with some interesting exceptions: multilingual countries and those with unique regional patterns. Large parts of Switzerland have been allocated to the same cluster as the Alsace region in France, Southern Luxembourg and the Bolzano region in Northern Italy, denoting similar surname characteristics shared by these multilingual areas with German language heritage. The analysis has also split Belgium along linguistic lines, assigning Flanders to the same cluster as the Netherlands and Wallonia to the French cluster, with part of Brussels appearing as a French enclave within Wallonia.

<- Figure 6 about here ->

Denmark, Norway and Sweden have been assigned to the same cluster except for one sparsely populated area of northern Sweden that is well known to have commonalities with its Finnish neighbour. This particular area has been grouped together with more “peripheral” countries such as Poland and Serbia, Montenegro and Kosovo. The robustness values associated with this area in Sweden are low, suggesting the region shares relatively little in common with the countries included in this cluster, which is truly a Polish cluster, with the ex-Yugoslavia region being associated with it because of its small size in relative terms (in effect an outlier as the aforementioned Northern Swedish region).

Beyond contemporary national political boundaries there are some interesting within country regionalisations that derive from the analysis. In the UK, historical linguistic regions such as Wales, and the Scottish Islands are clearly distinguishable from the rest of the UK. It is also interesting to see the urban corridor around London suggesting that the surname composition of these areas is much more diverse and hence disconnected from the national picture. This demonstrates the uniqueness in the surname composition of contemporary global migrants to the London area (see also Longley *et al.,* 2011). In the rest of the British Isles, Ireland (Eire) is grouped under a single cluster, that includes most of Northern Ireland, except for the Eastern coast reflecting the close migration and trade flows with Great Britain.

<- Figure 7 about here ->

In France, the mainland except for the Alsace-Lorraine has been allocated to a single cluster that includes the island of Corsica and the Geneva region in Switzerland, as well as the Wallonia region in Belgium. This is hence a 'tight French surnames cluster' automatically identified by the clustering algorithm. Italy has been split in two clusters, with a northern and western cluster separated from the rest of the country. Spain solidly belongs to a single cluster, despite its strong multilingual cleavages (Mateos and Tucker 2008), perhaps because of its overall low surname diversity (Scapoli *et al*, 2007). Most of Germany is allocated to a single cluster, while most of Austria belongs to a separate cluster, with some spillover regions between the two.

4.3. Multidimensional Scaling (MDS)

The results from the multidimensional scaling largely support the consensus clustering outcome. The 2-D MDS plots for individual countries shown in Figure 4 provide an indication

of the location of each of the spatial units in their multidimensional surname space. Those countries that have largely homogenous surname distributions form very tight clusters, such as Germany, Ireland or Denmark. Others such as Switzerland, Luxembourg, France or Spain, show a greater degree of scatter, reflecting present or historic multilinguism. Of most interest are the outlier points for each of the countries. For example, the three highlighted points in Italy's distribution are spatial units on the island of Sardinia, and those highlighted in France represent the border region of Alsace-Lorraine.

Figure 5 provides the geographic context to the results of the MDS analysis and is, in many ways, much more informative as a result. The maps (best viewed electronically at www.spatialanalysis.co.uk/surnames) create a similar impression to those in Figure 3 in addition to some more subtle distinctions. For example MDS Dimension 3 suggests a rather strong north-south split within Germany that is not noticeable in the consensus clustering results or the three-colour map in the same figure. Multi-lingual countries are also clearly identified in this figure, as well as some of the diversity within the Netherlands identified by Barrai *et al.* (2002). It is clear from Figure 5 that the European map has a number of abrupt transitions in its surname compositions. There are clear splits between the British Isles and the Continent, between Romance and Germanic languages, between Scandinavia and the rest of Europe, and between Poland and Germany. The latter abrupt transition is especially striking since the current Polish-German border only dates to 1945. This probably reflects rapid population movement during World War II and the practice of surname change or forced migration on the Soviet side during the Cold War. Such distinctions are perhaps unsurprising but these maps show, for the first time, how abrupt boundaries across Europe

can simply be captured by surname frequencies derived from data assembled from telephone and other directories.

5. Discussion

5.1. Regionalisation methods

The fact that the outcomes from the two separate regionalisation techniques used in this paper, consensus clustering and MDS, are in broad agreement with previous research in this area is encouraging and serves to endorse their use in geographic analysis of population structure. Clustering the merged matrix provided a more consistent outcome than consensus clustering, which in turn was more reliable than clustering areas using a single algorithm. The method does not obviate the need for the selection of a single algorithm to produce the final result, but it does provide some useful metrics upon which to base this decision. As Figure 3 demonstrates, the ability to map the cluster membership robustness of each spatial unit to its respective final cluster provides a powerful way of assessing the appropriateness of the outcome for each specific area. A key flaw with conventional clustering routines is the requirement to assign every item to one of a limited set of clusters, since this may result in questionable cluster allocations. Using robustness measures, such 'weak' allocations can be identified and interpreted with an appropriate degree of caution. In addition the ΔK measure is useful for indicating the optimal number of clusters that should be used as an input to the algorithm. It should be noted that "optimal" in the quantitative sense, might not be optimal in the practical sense. If the outcomes were to be mapped, for example, there would be a limit on the number of cluster outcomes that can be readily discriminated by the map user. A substantial advantage of the methods presented

here is in the visual outputs that they provide so this limitation should not be underestimated.

A final consideration relates to the opposite scenario where the ΔK measure indicates that a very low cluster number is optimal but the researchers may wish to identify a greater number of clusters to highlight diversity. In this case the desired clustering result can be shown alongside that which is optimal. Merged consensus clustering cannot therefore entirely remove the need for subjective guidance of cluster analysis, but it does provide measures upon which researchers can base their decisions. We do not claim that our use of consensus clustering has provided a panacea to the many issues surrounding the clustering of surname data. We do hope, however, to have made a substantial empirical contribution to the debates surrounding such issues through the application of the method to such a large dataset.

The maps shown in Figure 5 demonstrate the power of mapping MDS values in this context. The resulting impression of regionalisation is similar to that produced by the computationally more intensive consensus clustering with the additional advantage that less discrete phenomena such as isolation by distance is also shown. Assigning discrete groupings to the visual impressions created by the maps is best left to the sorts of clustering methodologies shown here, but the relative simplicity (using most widely available statistical software packages) and speed of the MDS classification makes it a powerful one in this context.

5.2. Issues of geographical scale and size

The datasets used here contain information at the level of the individual for most countries, and therefore, they offer the potential for much finer-scale analysis than has been presented

here for the 685 NUTS2/3 areas. Very fine spatial units will create different regionalisation outcomes out of the same input dataset as those based at a coarser scale. This effect is clearly seen if Figure 7 in Scapoli *et al.* (2007) is contrasted with Figure 3A above. For example, Scapoli *et al.* (2007) have clustered the entire region of Lorraine as part of the Franco-German border area using NUTS 2 regions, while the smaller geographical units presented in Figure 3 (NUTS 3) suggest that it is only those departments contiguous with the German border (and not with Belgium or in the interior) that fall into this category.

The issue of scale is partially resolved through the application and context of the surname research being undertaken. If, for example, surname analysis is used as a proxy for genetic information at the European level then fine scale analysis may be unnecessary since most traits are only noticeable at coarse granularity (Cavalli-Sforza, 2000). That said, as Longley *et al.* (2011) demonstrate using similar methods for Great Britain, the use of fine granularity units of analysis will still preserve the large-scale trends if these are legitimate and not just artefacts of the units used. A major advantage of smaller spatial units is their ability to highlight detail, such as that arising out of more recent migration events. This may be especially useful in the context of understanding segregation in global cities such as London, Paris and other large European cities. Whilst such fine-scale analysis would not be practical at a European level, it could nevertheless be undertaken within each of the 14 or so groupings created in this study in order to identify the dynamics *within* each of these surname sub-regions.

An issue to be considered alongside the size of spatial unit selected is the size, concentration and distribution of the populations contained within them. The (dis)similarity between the surname compositions of populations has been established between areas with the Lasker

Distance. The subsequent clustering of the measure is sensitive to the different levels of aggregation and sampling associated with the inconsistent population sizes represented by each spatial unit. Dissimilarity measures, such as the Lasker Distance, rely on comparisons between aggregate population groups that are often equally weighted for the analysis. A spatial unit representing 100 people is therefore treated in the same way as one with 1,000 or even 10,000. A country's influence on the analysis is in part based on the number of spatial units it has rather than the size of its population. The likely result is an apparent increase in diversity for countries partitioned into large numbers of regions, despite relatively uniform surname compositions. It is therefore the case that the resulting classification is dependent on the size of the spatial units, the population size per spatial unit and the surname heterogeneity within and between the spatial units. The use of merged consensus clustering has helped to accommodate some of these effects, in addition to minimising the impact of outliers in the cluster analysis. Future work will seek to establish a number of heuristics around which to base a suitable weighting methodology to account for the varying populations in each spatial unit across Europe.

A number of approaches could be used to mitigate the drawbacks associated with inconsistent levels of aggregation within distance measures. The obvious solution would be the greater standardisation of spatial units across Europe, in order that they better reflect population density. This, however, leads to complications such as whether the size of the resulting units should reflect the target population density or the sampled population density. In addition, more sparsely populated areas are going to require larger units (in terms of geographic extent) in order to meet a population threshold and this is likely to risk amalgamating culturally distinct groups as potential surname boundaries are crossed. This

solution would present a major undertaking at the European level and may not produce significantly improved results. More practical options could therefore include weighting the dissimilarity calculation or its subsequent clustering. One possible approach, in this context, would simply be to multiply the elements of the Lasker Distance matrix by a suitably normalised population weight. Such an approach may also require some nationally varying "alpha" value to alter the influence of the population weighting on the cluster outcome.

We consider the disparities in sample size for each population a lesser issue because, as Fox and Lasker (1983) demonstrate, the relative proportions of each surname tend to be consistent whatever the percentage of the population is sampled so long as the sample is representative. We believe that our data sources are broadly representative of their target populations (with the caveats below) and therefore will have adequate proportions of each surname to calculate realistic pairwise distances. Finally, an element of uncertainty has also been introduced in this analysis by the different provenance of the surname frequency data for each country. While the ultimate data source for most of the countries is the national telephone directory (except the UK where an enhanced electoral register was used), these obviously do not present identical characteristics across the 16 countries. These include national variations in the gender bias towards male registration in telephone directories, variable penetration of land line rental in the population, different conventions for subscribers removing their entries from directories, different customs in registering names to telephone lines and different procedures and conventions by the companies that commercialised the data. Following from the previous discussion on geographical scale, this can also be applied to geographical extent. If we had clustered surname distances

individually within each country the results would have been somewhat different to doing so for the whole of Europe in a single step.

6. Conclusions

This research has offered a number of important contributions to our understanding of the spatial distributions of surnames. It has combined a commonly used method of establishing the similarities in the surnames composition between different populations or areas (isonymy) with novel forms of data clustering and geographic visualisation (consensus clustering and MDS). It has created the most comprehensive surname regionalisation of Europe to date by examining the 8 million surnames of over 150 million people who can reasonably be deemed representative of the entire populations of each of the 16 countries included here. The unprecedented size and comprehensiveness of the dataset used has provided new insights into the problem of identifying the regionalisation of European populations using surname distributions as a proxy for cultural and genetic structure. The introduction of a new method – merged consensus clustering – in this context has greatly increased the stability and consistency of traditional clustering algorithms. In addition the mapping of a measure of cluster robustness alongside the final results provides important context about the strength of the resulting regions. This information is augmented by the results of MDS analysis that, as shown in Figure 5, capture both the abrupt transitions in surname composition as well as more gradual trends. This goes some way towards combining the traditionally continuous models of genetic diversity with the discrete transitions commonly established in surname analysis.

In conclusion, this paper has sought to demonstrate the utility of an inductive approach to summarising and analysing large population datasets across cultural and geographic space, 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:

Barrai, I., Rodriguez-Larralde, A., Mamolini, E., Manni, F., Scapoli, C. 2000. Elements of the Surname Structure of Austria. *Annals of Human Biology.* 27, 6: 607-622.

Barrai, I., Rodriguez-Larralde, A., Manni, F., Scapoli, C. 2002.Isonymy and Isolation by Distance in the Netherlands.*Human Biology*. 74, 2: 263-283.

Barrai, I., Rodriguez-Larralde, A., Manni, F., Ruggerio, V., Tartari, D., Scapoli, C. 2004. Isolation by Language and Distance in Belgium. *Annals of Human Genetics.* 68, 1: 1-16.

Cavalli-Sforza L. L., Moroni A, Zei G. 2004. *Consanguinity, inbreeding, and genetic drift in Italy*. Princeton, NJ: Princeton Univ Press.

Cavalli-Sforza, L. L. 2000. *Genes, Peoples and Languages*. Penguin Books, London.

Cavalli-Sforza, L. L. and M. Feldman. 1981. Cultural Transmission and Evolution. Princeton University Press, Princeton.

Cheshire, J. Adnan, M. Gale, C. 2011. The Use of Consensus Clustering in Geodemographics. *Proceedings of GIS Research UK, Portsmouth.*

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

Cheshire, J. A., Longley, P.A., SIngleton, A. The Surname Regions of Great Britain. *Journal of Maps.* 401-409.

Colantonio S. E., Lasker G. W., Kaplan B. A., Fuster V. 2003. Use of surname models in human population biology: a review of recent developments. *Human Biology* 75, 6: 785-807

Crow, J.F., and Mange, A. 1965. Measurements of Inbreeding from the Frequency of Marriages Between Persons of the Same Surnames. *Eugenics Quarterly*. 12, 199-203.

Darwin G. H. 1875. Marriages between first cousins in England and their effects. *Journal of the Statistical Society of London* 38: 153-184.

Fox, W. and Lasker, G. 1983. The Distribution of Surname Frequencies. *International Statistical Review.* 51: 81-87.

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

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 (2nd Edition).* Chapman and Hall, London.

Hanks, P. 2003. *Dictionary of American Family Names.* New York: Oxford University Press.

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

Grotkjær, T., Winther, O., Regenberg, B., Nielsen J. and Hansen, L. 2005. Robust Multi-Scale Clustering of Large DNA Microarray Datasets with the Consensus Algorithm. *Bioinformatics* 22, 1: 58-67.

Kaufman , L. and Rousseeuw, P. 1990. *Finding Groups in Data*. Wiley: New York.

Kohonen, T. 1990. The Self-Organizing Map. *Proceedings of the IEEE*. 9: 1464-1480.

Lao, O. et al., 2008. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16), 1241-1248.

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

Lasker, G. 2002. Using surnames to analyse population structure. In Postles D (eds.), *Naming, Society and Regional Identity*, Leopard's Head Press: Oxford: 3-24

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.

Longley, P. A., Cheshire, J. A., Mateos, P. 2011. Creating a Regional Geography of Britain Through the Spatial Analysis of Surnames. *Geoforum* (In Press). DOI: 10.1016/j.geoforum.2011.02.001

Longley, P. A., Webber, R., Lloyd, D. 2006. The Quantitative Analysis of Family Names: Historic Migration and the Present Day Neighbourhood Structure of Middlesborough, United Kingdom. *Annals of the Association of American Geographers.* 97, 1: 31-48.

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.

Manni, F., Guerard, 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., Heeringa, W. and Nerbonne, J. 2006. To What Extent are Surnames Words? Comparing Geographic Patterns of Surname and Dialect Variation in the Netherlands. *Literary and Linguistic Computing*. 21, 4: 507-528.

Manni, F., Toupance, B., Sabbagh, A., and Heyer, E. 2005.New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling.American Journal of Physical Anthropology. 126: 214- 228.

Mateos, P. And Tucker, D.K. 2008. Forenames and Surnames in Spain in 2004. Names, a Journal of Onomastics, 56, 3: 165-184.

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

Monti, S., Tamayo, P., Mesirov, J., Golub, T. 2003. Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning.* 52: 91-118.

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, 183–198.

Nei, M., 1973. The Theory and Estimation of Genetic Distance. In *Genetic Structure of Populations*. Edited by Morton, N. E. 45-64.

Nei, M., 1978. Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics*. 583-590.

Nerbonne, J., Kleiweg, P., Heeringa, W. and Manni, F. 2008.*Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering. Data Analysis, Machine Learning and Applications*. 647-654.

Nerbonne, J. 2010. Mapping Aggregate Variation. In Lameli, M., Kehrein, R. and Rabanus, S. (eds.) *An International Handbook of Linguistic Variation*. Vol. 2. Berlin: Mouton De Gruyter. Chap. 24. pp. 476-495.

Openshaw, S. 1984. *The Modifiable Areal Unit Problem.* Norwich: Geo Books.

Piazza, A., Rendine, S., Zei, G., Moroni, A., Cavalli-Sforza, L. 1987. Migration Rates of Human Populations from Surname Distributions. *Nature.* 329: 714-716.

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., Berretta, M., Nesti, C., Mamolini, E., Barrai, I. 1998. Isonomy and Genetic Structure of Switzerland. II. Isolation by Distance. *Annals of Human Biology.* 25, 6: 533-540.

Rodriguez-Larralde, A., Barrai, I. Nesti, C., Mamolini, E., Scapoli, C. 1998. Isonymy and Isolation by Distance in Germany. *Human Biology.* 70, 6: 1041-1056.

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

Scapoli, C., Goebl, H., Mamolini, E., Rodriguez-Larralde, A., Barrai, I. 2005. Surnames and Dialects in France: Population Structure and Cultural Evolution. *Journal of Theoretical Biology.* 237, 2: 75-86.

Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez- Larralde, R., Barrai., I. 2007. Surnames in Western Europe : A Comparison of the Subcontinental Populations Through Isonymy. *Theoretical Population Biology*. 71 : 1, 37-48.

Simpson, I., Armstrong, D., Jarman, A. 2010. Merged Consensus Clustering to Assess and Improve Class Discovery with Microarray Data. *BMC Bioinformatics*. 11: 590.

Smith, M. T. 2002. Isonymy analysis. The potential for application of quantitative analysis of surname distributions to problems in historical research. In Smith, M. (ed.). *Human Biology and History*. Taylor and Francis: London: 112-133

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

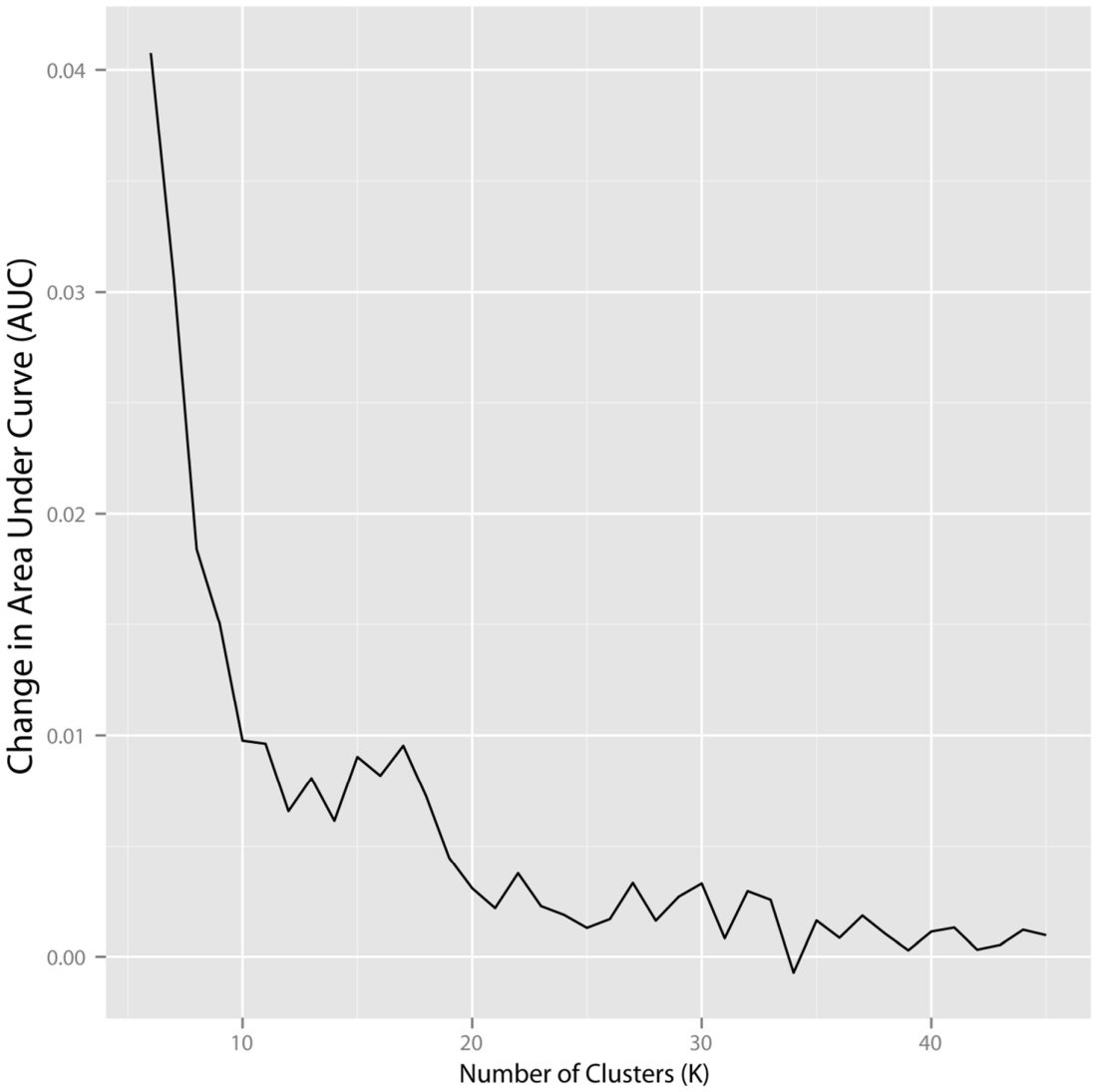

Figure 1: The delta *K* plot used to inform the decision to cluster the Lasker Distance matrix into 14 groups. It shows the change in AUC values as calculated in Equation 8.

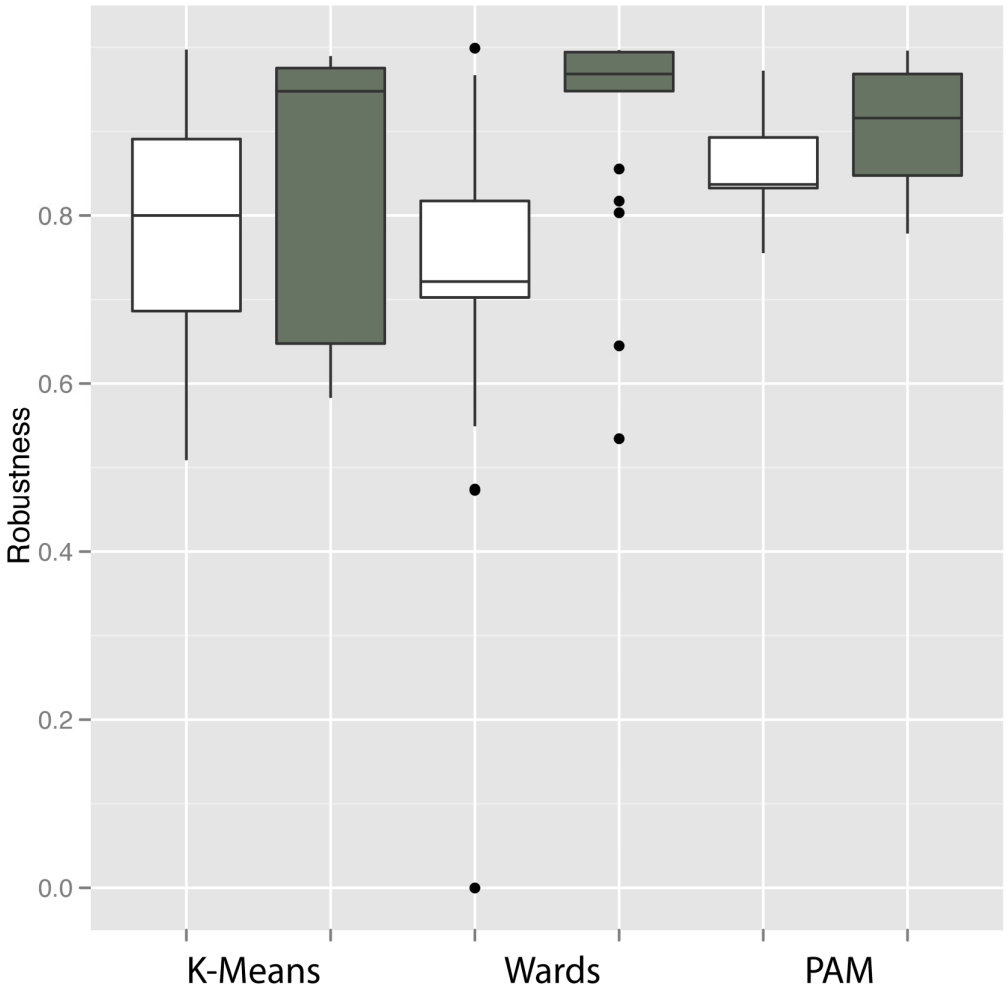

Figure 2: Box-plots showing the robustness values associated with the structures of each of the cluster outcomes. White boxes are produced from direct clustering of the distance matrix and grey boxes are produced from clustering the merged consensus matrix. For reasons outlined in the text, PAM provides the best solution in this instance.

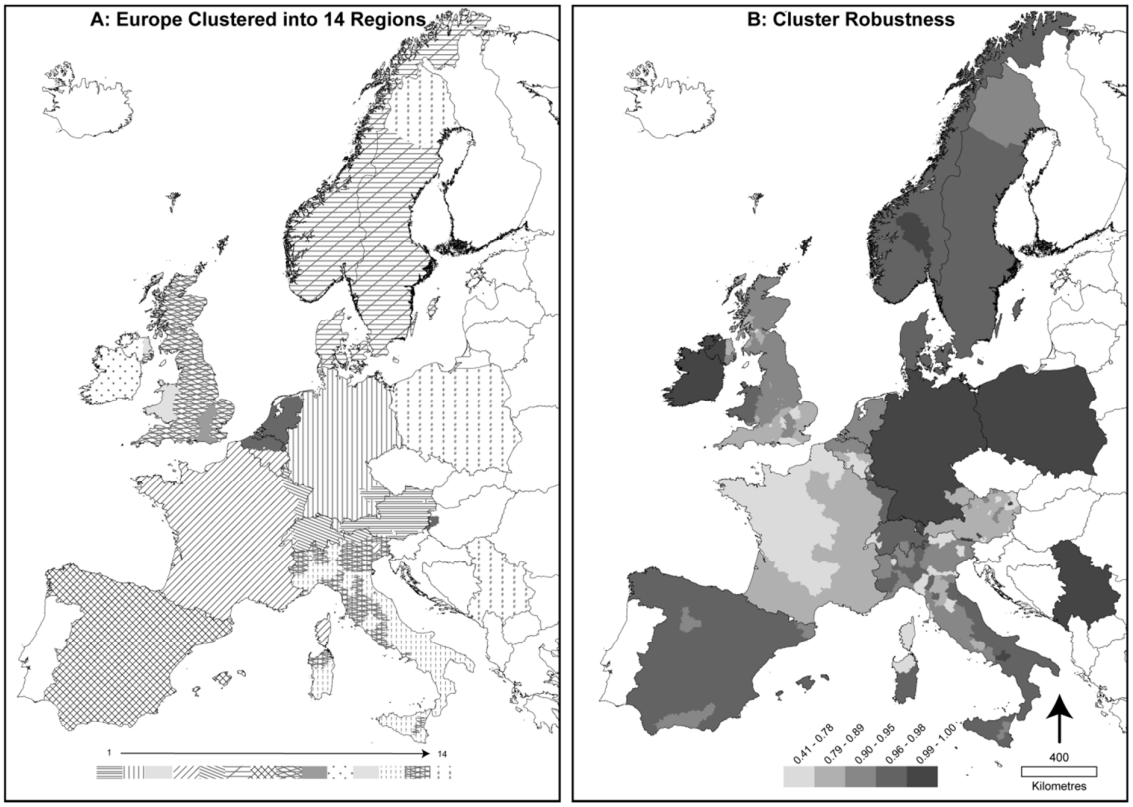

Figure 3: Maps showing the spatial distributions of each of the 14 cluster allocations (left) and their respective robustness values (right). Higher robustness values represent a better result. On the left hand plot each cluster has been assigned a unique pattern. A full colour version can be found at spatialanalysis.co.uk/surnames.

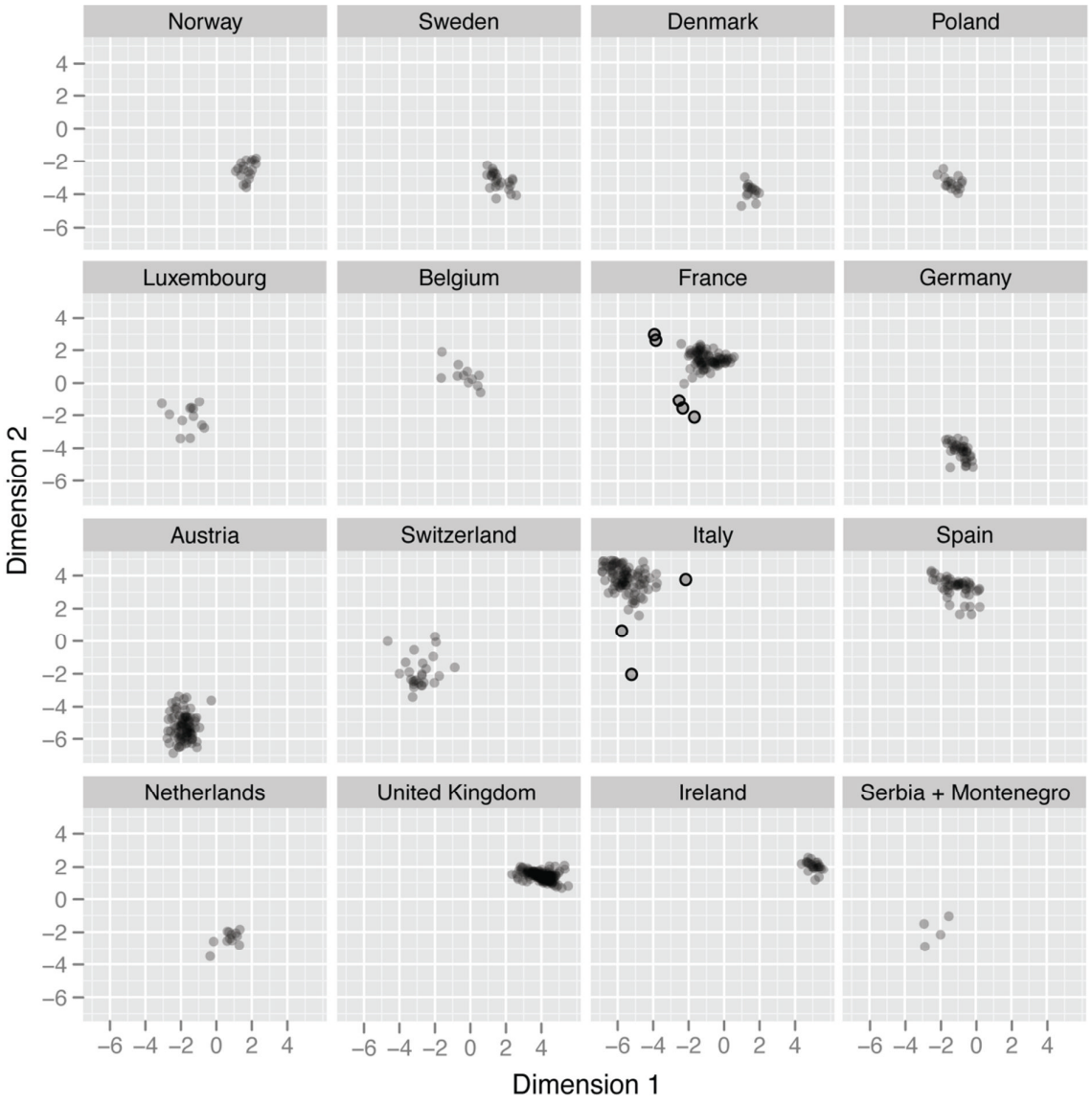

Figure 4: Plots illustrating the results of the 2-dimensional MDS analysis on the Lasker Distance matrix. Each country has been separated for ease of comparison and each point represents a NUTS region. Stress value= 17.089.

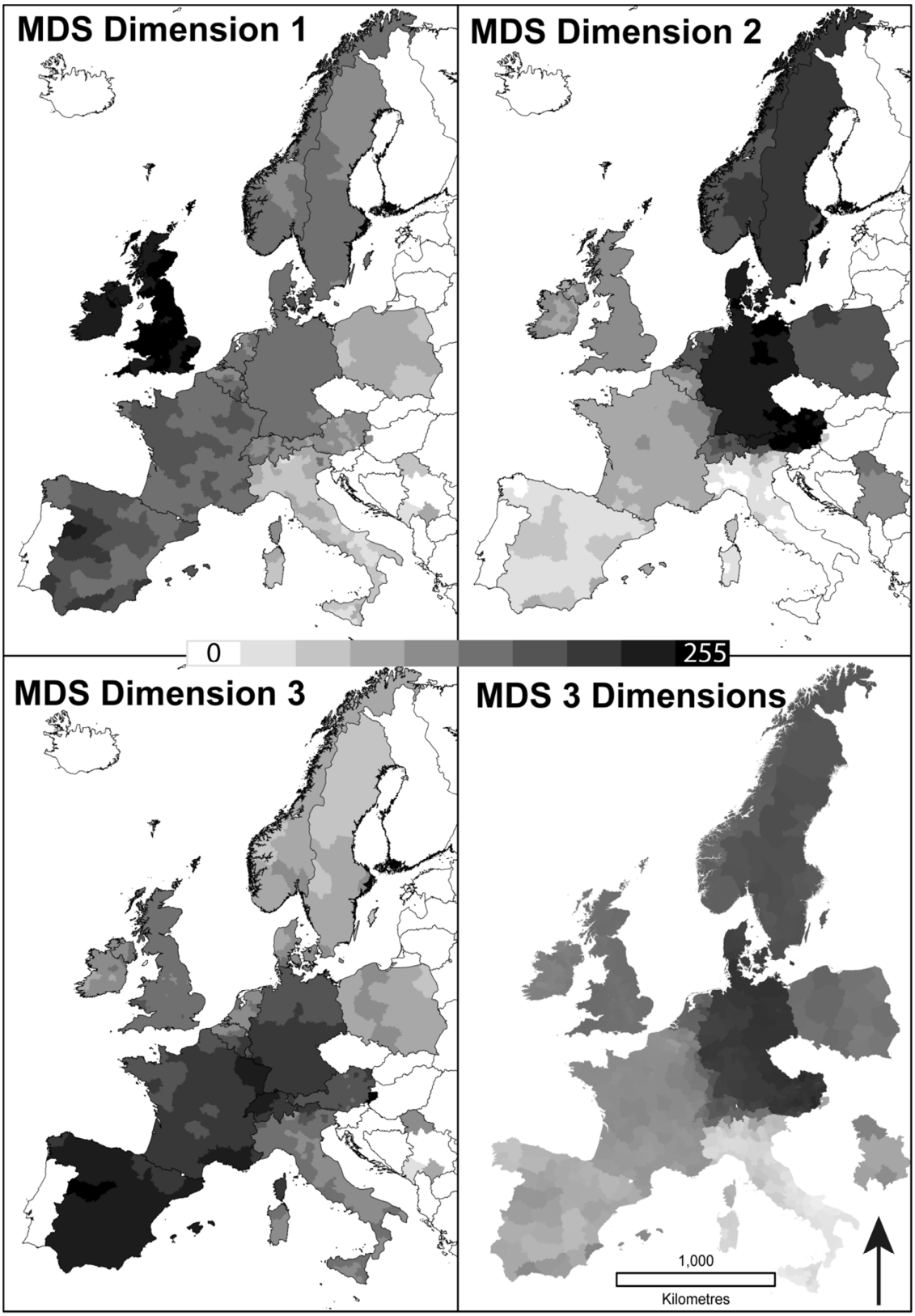

Figure 5: Maps showing the spatial distributions of each dimension produced from the 3 dimensional MDS. Each dimension has been rescaled to a value of between 0 and 255 to facilitate the creation of RGB colours (best viewed online: spatialanalysis.co.uk/surnames). Stress values for 3 dimensions= 11.064 and 4 dimensions= 9.838.

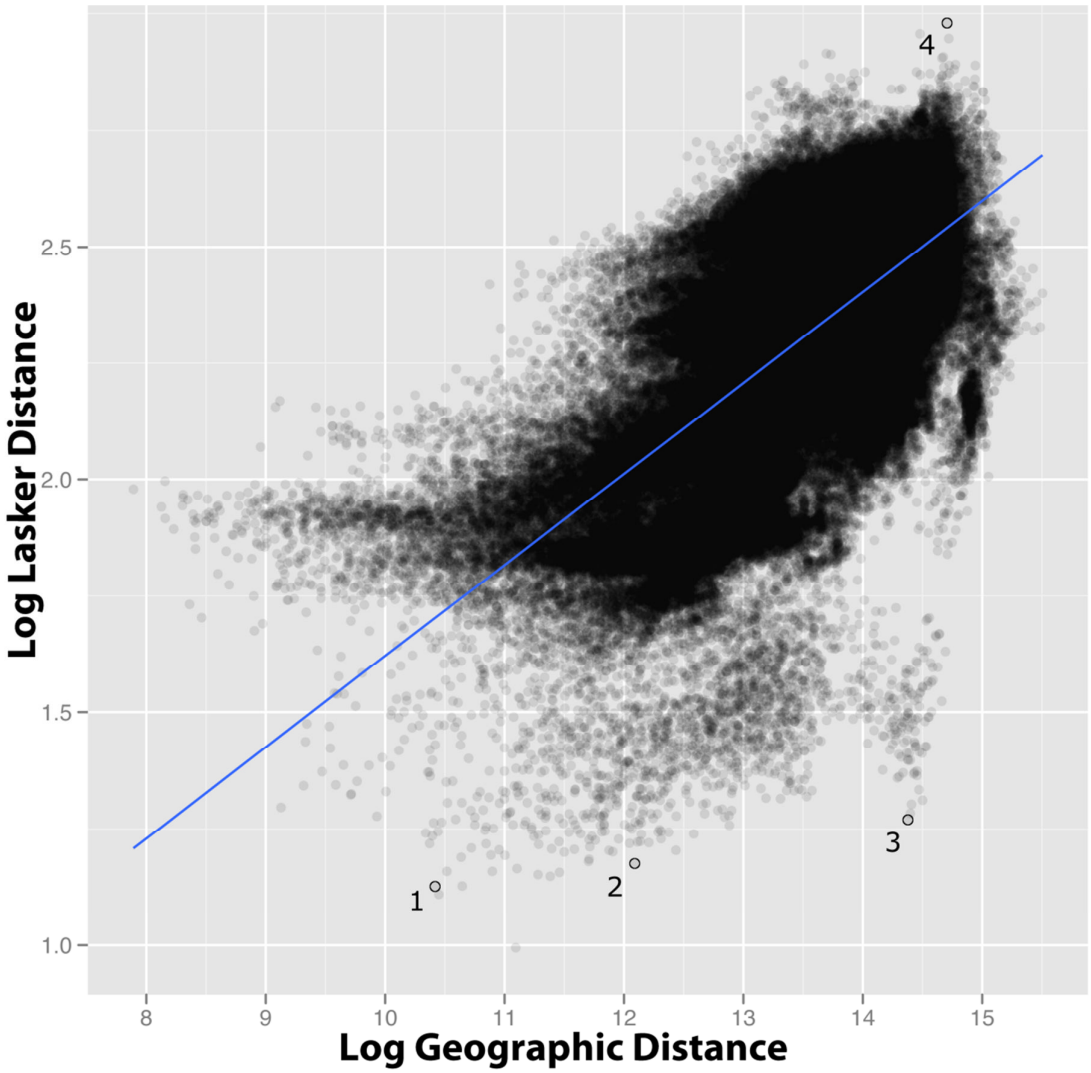

Figure 6: A plot showing the relationships between the Lasker Distance and log geographic distance (km). Taking the log of each axis creates a greater spread of points in the plot window. Every possible region-pair is represented. Point 1 is between a pair of neighbouring areas in northern Wales; Point 2 is between the areas of Asturias and Ourense in northern Spain; Point 3 is between Salamanca and Tenerife; Point 4 is between Crotone in the far south of Italy and Crotone in Northern Island.

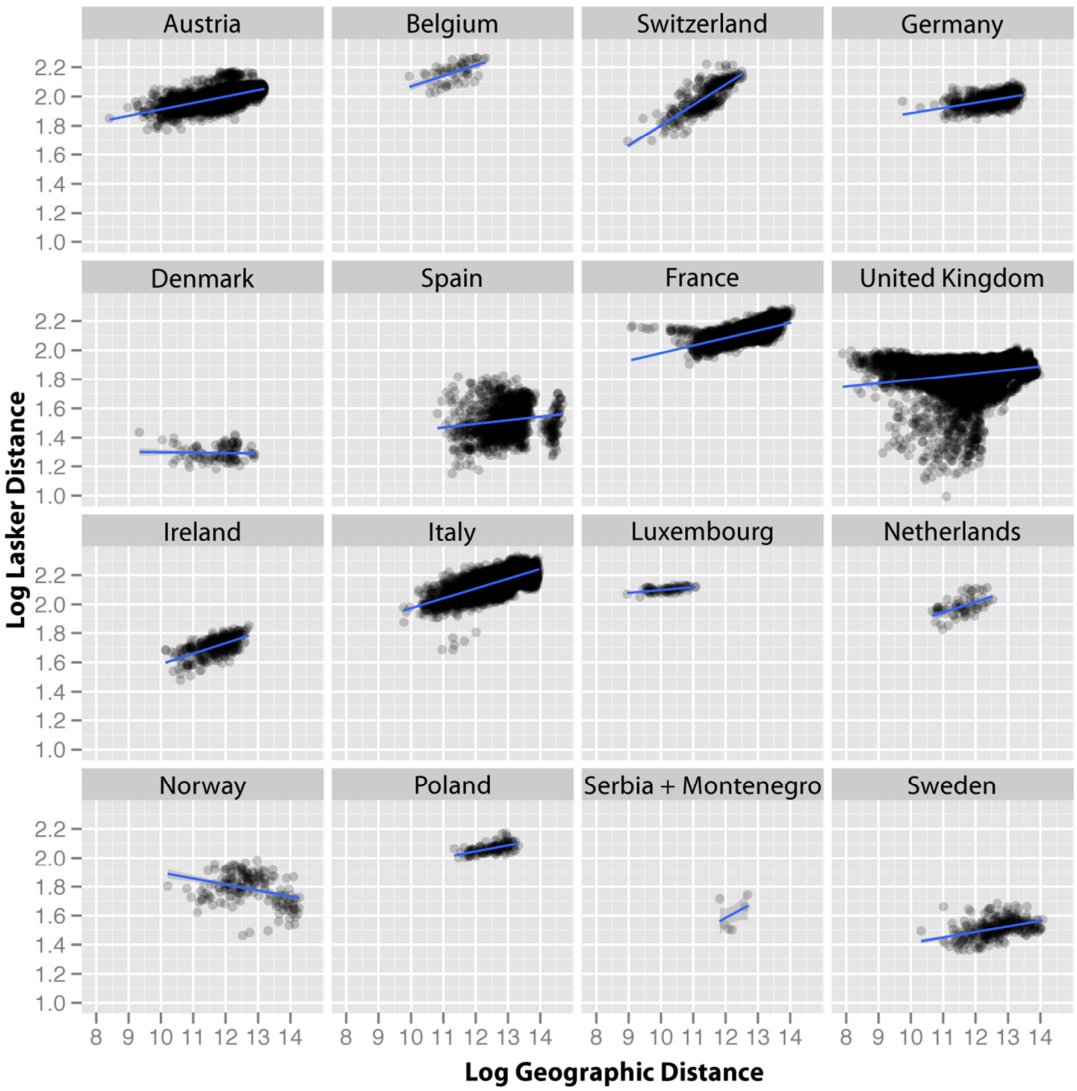

Figure 7: A plot showing the relationships between the Lasker Distance measures and log geographic distance (km) within each European country studied here. Every possible region-pair is represented.

Symbol	Description
$\boldsymbol{D} = \{\boldsymbol{e_1}, \ldots, \boldsymbol{e_N}\}$	Data, in this case surname distance matrix with geographic units (e_i's) to be clustered.
$\boldsymbol{N}$	The number of geographic units (or number of rows) in distance matrix.
$\boldsymbol{P} = \{\boldsymbol{P_1}, \ldots, \boldsymbol{P_K}$	Partition of D into K clusters.
$\boldsymbol{K}, \boldsymbol{K_{max}}$	Number of clusters, maximum number of clusters.
$\boldsymbol{N_k}$	Number of items in cluster k.
$\boldsymbol{H}$	Number of resampling iterations.
$\boldsymbol{D^{(h)}}$	Dataset obtained by resampling D (h-th iteration).
$\boldsymbol{M}, \boldsymbol{M^{(h)}}$	Connectivity matrix, corresponding to h-th iteration.
$\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}^{(K)}}$	Consensus matrix, corresponding to K clusters.
$\boldsymbol{I^{(h)}}$	*N x N* indicator matrix.

Table 1: Variables and definitions used in merged consensus clustering. Adapted from Monti *et al.* (2003).

Country	Data Year	Total Population	Worldnames Population	No. Unique Surnames	NUTS Level	No. Spatial Units
Poland	2007	38,518,241	8,015,455	339,339	2	16
Serbia, Montenegro and Kosovo	2006	10,159,046	1,704,559	69,977	2	4
Austria	1996	8,316,487	2,520,012	81,387	2	98
Belgium	2007	10,511,382	3,489,068	852,492	3	11
Denmark	2006	5,457,415	3,074,871	153,134	2	15
France	2006	64,102,140	20,280,551	1,197,684	3	96
Germany	2006	82,314,900	28,541,078	1,226,841	2	39
Great Britain	2001	60,587,300	45,690,258	1,612,599	3	131
Rep. of Ireland	2007	4,239,848	2,916,744	46,507	3	26
Italy	2006	59,131,282	15,927,926	1,305,554	3	103
Luxemburg	2006	480,222	117,619	75,267	3	12
Netherlands	2006	16,570,613	4,672,344	531,970	2	12
Norway	2006	4,770,000	3,536,524	123,240	3	19
Spain	2004	45,116,894	9,545,104	260,469	3	50
Sweden	2004	9,142,817	791,143	135,830	3	24
Switzerland	2006	7,508,700	1,565,098	19,270	3	26
	Totals	426,927,287	152,388,352	8,031,560		

Table 2: The countries and their data used in this study. "NUTS Level" refers to the geographic unit of analysis used. There are 495,059 hapax (occurring only once) surnames in the data.