doi 10.4436/JASS.92004

**JASs *I*nvited *R*eviews**

*Journal of Anthropological Sciences*
*Vol. 92 (2014), pp. 99-117*

# Analysing surnames as geographic data

**James Cheshire**

*UCL Centre for Advanced Spatial Analysis, UCL, Gower Street, London, WC1E 6BT, U.K.*
e-mail: james.cheshire@ucl.ac.uk

**Summary -** *With most surname research undertaken within the fields of anthropology and population genetics, geographers have overlooked surnames as a credible data source. In addition to providing a review of recent developments in surname analysis, this paper highlights areas where geographers can make important contributions to advancing surname research, both in terms of its quality and also its applications. The review discusses the emerging applications for surname research, not least in the mining of online data, and ends by suggesting three future research themes to ensure the building momentum of surname research continues to grow across disciplines.*

**Keywords –** *Surnames, Geography, Regions, Geneology.*

## Introduction

Family names, or surnames, provide almost every culture with a ubiquitous method for distinguishing between familial groups. Yet the routine use of surnames has meant that their cultural and geographical significance is often taken for granted. We frequently make judgments - either consciously or subconsciously - about a person's ancestry or origin based on their surname. It is obvious to those in the UK, for example, that surnames such as "Smit h", "Jones" and "Macleod" are English, Welsh and Scottish in origin, respectively. Placing surnames within a regional context becomes somewhat more specialist yet many people would, for example, relate the surname "Cheshire" to the county of the same name in England. Even within cities surnames unique to a particular ethnic or cultural group cluster in particular areas, reflecting the underlying population distribution. From such anecdotes alone, it is clear that surnames can contain spatial information on a variety of scales – from city neighbourhoods through to continents - relating to the probable origins, and areas of residence, for many of their bearers. The purpose of this review is to go beyond such anecdotal conjecture to demonstrate the ways in which researchers are able to generate insights into the geography of surnames.

This review will highlight areas where geographers have much to contribute to surname research, in addition to indicating a broadening range of applications that may help to increase awareness of the value of surname data. It will demonstrate the growing importance of surnames as a source of spatial data and argue for their increased utilisation both within and beyond geography. This review therefore differs from previous literature [see Colantonio *et al.* (2003) and Darlu *et al.* (2012)] that has traditionally focused on the use of surnames in population genetics. It will begin by outlining the utility of surnames as a source of geographic population data before covering recent developments in their analysis at both the individual and regional level. The final sections of the review are concerned with improving existing methodologies in the context of geography and also working towards a set of research themes to reflect the increasing range of surname research applications.

## Surnames and population geography

Surnames and geography dynamically intersect through the potential of genealogical data to enrich geographic research with regard to migration patterns, and through the construction of personal identity around ethnic origins or ancestral homes (Otterstrom & Bunker (2012). Surname adoption did not occur simultaneously in all places, and surnaming conventions have always been a product of both cultural (including linguistic) and legislative processes. Such processes are systematic but not geographically uniform, resulting in spatial structuring of surname distributions that may subsequently be obscured by population movements. In Britain for example, surnames were developed as a means to distinguish individuals (particularly men) from one another at a time when there were very few forenames. Such distinctions were essential for administrative purposes and surnames could be passed from one generation to the next to ensure a record of lineage. This inheritability is extremely important because it creates a self-maintaining, enduring and culturally significant surname geography. Patterns of inheritance form the basis for genealogical research and enable the creation of familial lineages that are often traceable to the point in time when the surname was first formalized. The extent to which surnames offer an unbroken line to the geographic past will therefore correspond to the different time horizons of surname adoption.

A further consideration in the use of surnames for population research is the fact that they are largely transferred along male lines and passed to women after marriage. This point is especially relevant in the context of sociological studies concerned with making broader generalisations about population mobility. In many countries the majority of women change their surname when they marry, thus adopting the same cultural marker as their partner (see Valetas, 2001). This creates minor problems in the classification of areas (according to their surname compositions) but more significant problems in the classification of individuals (according to their surname lineage). The surname may represent an entirely different culture to the one the woman was born into, and of course, cannot be used to apply the assumption that women who share the same surname are more likely to be related than those who do not. When classifying a region (as opposed to an individual) based on its surname composition, the inclusion of female surnames is less problematic because they could either have adopted a surname from a local male, or still hold a surname inherited from their father. Such conclusions are also true at a European level, and certainly in rural areas, with many matrimonial migrations being less than a few kilometres (Manni *et al.,* 2008).

In some contexts, such as population genetics, it makes sense to exclude females from fine scale studies of particular surnames (see Bowden *et al.,* 2007) but at the more aggregate scale their impact is less clear. Winney *et al.* (2011), for example, found only a minor effect after the exclusion of females from their sample design suggesting that the phenomenon of women marrying locally continues in many areas of the UK. More pragmatically, it is often challenging to systematically remove females from large surname databases without the gender of each bearer being recorded. In several papers, such as Longley *et al.* (2011), it was considered best to treat surnames with male and female bearers in the same way. Aside from the uncertainty inherent in assigning genders to individuals based on their forenames (which are more likely to be available), a strength of this is the use of the most complete population data available which could be compromised with poor gender-based sampling. In some countries, such as the Czech Republic (see Novotny & Cheshire, 2012), female derivations of a name are easily identified and can therefore be systematically removed if required.

The following two sections will outline recent developments in the geographic analysis of surnames. Developments discussed in the first seek to create a geographic classification of individual surnames by identifying their point of origin or their areas of highest concentration; those in the second seek to identify regions with similar surname compositions. Almost all recent studies have benefited from the increasing availability of comprehensive digital databases and the computational power to process them: compare, for example, Otterstrom & Bunker's (2012) access to 800 million surnames with Lasker's (1985) access to a few thousand.

## The classification of individual surnames

Many early studies into the geography of individual surnames were little more than visual interpretations of point distributions (or counts) of surname occurrences (see Guppy, 1890). In the past few decades more sophisticated statistical approaches have been applied to offer a range of metrics such as a surname's geographic concentration, its flows and probable area of origin. All methods rely on the fact that surnames are not geographically random – the vast majority exhibit some kind of spatial patterning. For example, surnames derived from placenames within a 50km radius of Manchester and Birmingham in Great Britain occur at 145% of the expected frequency, reducing to 124% 50-99km away and 82% of expected over 150km away (Kaplan & Lasker, 1983). These statistics offer important insights into the spatial dynamics of surnames and demonstrate the idea that frequencies often follow standard models of distance decay. There are now several websites, for example worldnames.publicprofiler.org, that enable the simple production of surname maps alongside some basic descriptive statistics, offering specialists and non-specialists alike the facility to undertake similar analysis of the surnames of interest to them.

One of the biggest challenges in this area remains the creation of a consistent approach to the inductive analysis of the millions of surnames in circulation. Until relatively recently researchers have selected surnames based on manual empirical research, rather than objectively selecting them utilising a range of consistent metrics. To address these limitations, several entirely automated approaches to the geographic classification of individual surnames have been proposed. The first, outlined by Tucker (2005), does not require geographic information; it simply uses a reference list of "diagnostic forenames" that have been manually classified into ethnic groups. These can then be used as a template for the classification of an entire names register based on the frequency of surnames assigned to a particular forename. This method is surprisingly efficient because the heavily skewed nature of forename and surname distributions ensures relatively few names can cover a large proportion of the population (Tucker, 2001).

Expanding this principle, Mateos *et al.* (2011) have constructed a large naming network based on surname and forename pairs. This approach reveals information about the names because certain pairs appear more commonly within ethnic groups and therefore become clustered together in network space. These so-called cultural-ethnic-linguistic (CEL) clusters can then be assigned labels to tie them to a specific region or cultural group (Mateos *et al.,* 2011). An example output from this is provided in Figure 1. This work marks a shift away from name dictionaries and towards automated classification with clustering. For a full review of other approaches to name-based ethnicity classification see Mateos (2007). As is outlined in the applications section, such approaches have a wide range of contemporary applications from classifying patients admitted to hospital (see Petersen *et al.,* 2011) to users of social media (Chang *et al.,* 2010).

The implicit assumption in name-based ethnicity classification research is that there is a single area of origin for that surname. This may be true at the global level, but at the sub-national level surnames have more complex spatial distributions with many exhibiting multiple points of origin and altered distributions. To capture this, a different approach has been developed by Novotny & Cheshire (2012) who apply the Dice coefficient; a comparative measure that builds a network of surnames based on the degree of similarity in their spatial distributions. The process is computationally intensive since it produces an adjacency matrix with a column/ row for every surname in the population. It is therefore less easily scaled to very large datasets of the kind processed by Mateos *et al.* (2011). In the case of the Czech Republic (the country used in the study) over 200 million surname proximity observations were calculated to create the "Czech surname space". Nevertheless, the method was able to identify clusters of surnames that shared

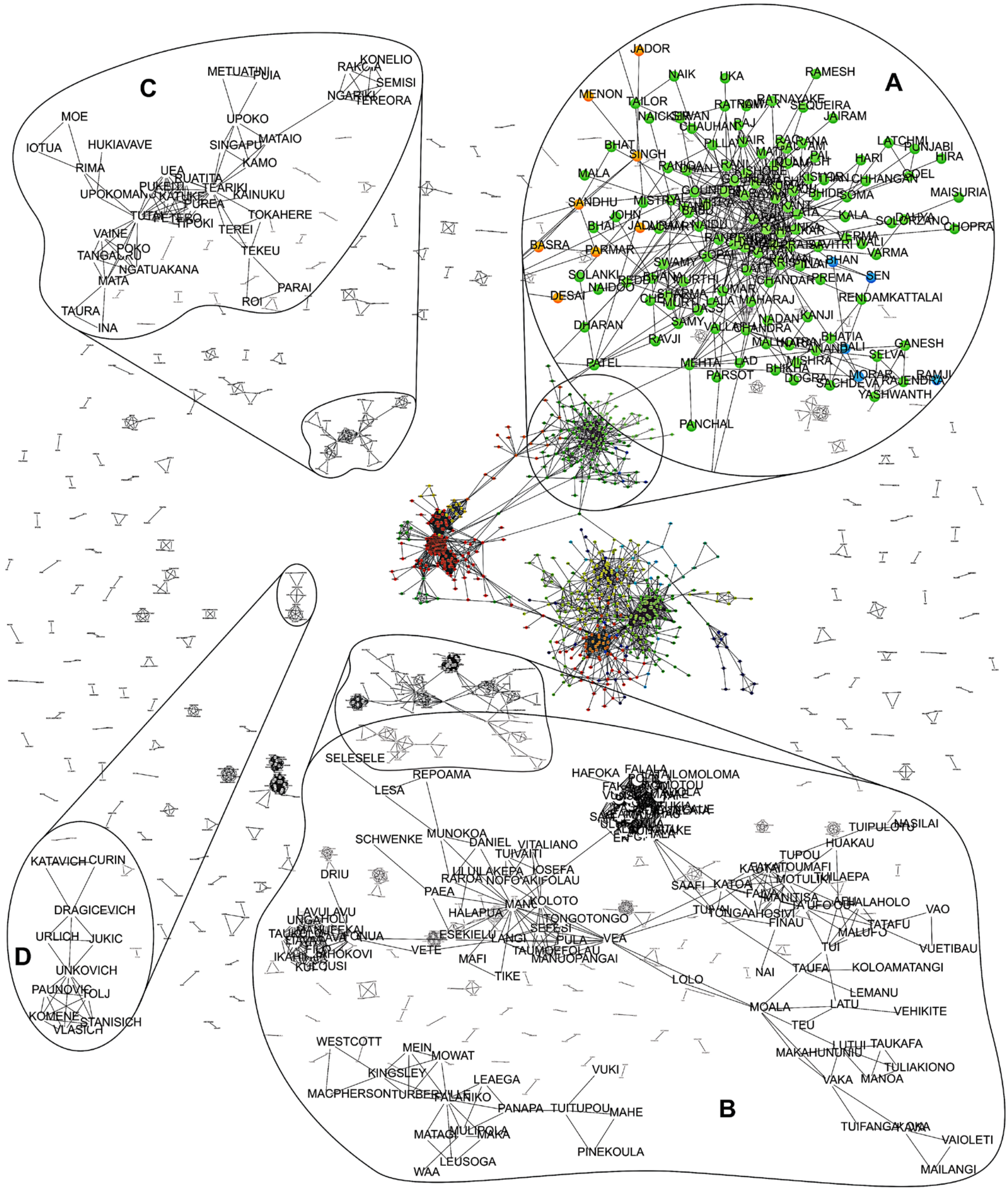

*Fig. 1 - A sample of the Auckland surnames network. The graph shows the highly structured outcome of naming practices in a city with high rates of immigration from all over the world. The giant component in the centre of the graph has been classified with* fastcommunity *algorithm into 22 clusters, each depicted by a different node colour. Four subgraphs are magnified to show the tightly knit internal structure of some CEL communities. (A) is classified as part of the giant component (and is South Asian/Indian), the others are Tongan (B), Samoan and other Pacific Islanders (C), and Eastern European (particularly Dalmatian: D). The last three are disconnected from the network's giant component. Taken from Mateos* et al.*,2011, doi:10.1371/journal.pone.0022943.g002. The colour version of this figure is available at the JASs website.*

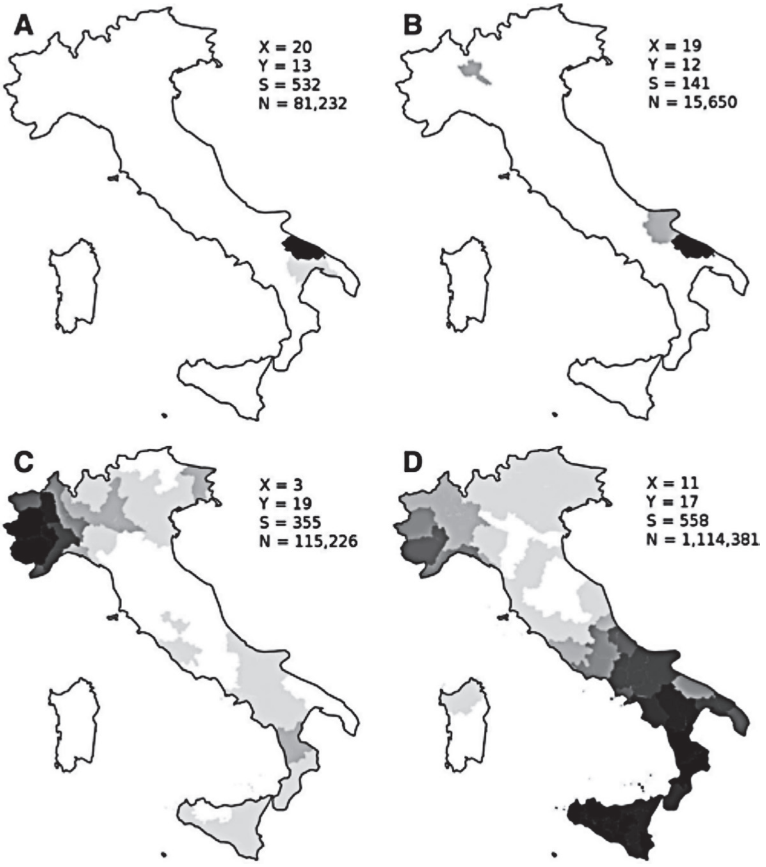

***Fig. 2 - Some of the 400 maps displayed in Boattini** et al. **(2012). In the two maps at the top the two clusters of surnames are most frequent in the province of Bari. For Polyphyletic surnames (those with multiple areas of origin) it is less clear-cut. (C) concerns regional polyphyletic surnames (Piedmont); whereas (D) shows macroregional polyphyletic surnames (all southern Italy besides Sardinia). Regional polyphyletic surnames are quite frequent in Italy and point to long-lasting regional socio-cultural identities and dialects. S is the number of surnames clustered in the SOM cell and N is the total number of individuals bearing one of the S surnames. Taken from Boattini et al., 2012, p. 254).***

common origins or cultural characteristics, in addition to providing a basis for the creation of surname regions (a topic explored below). The paper also highlights the impact of large-scale enforced migrations on the geographic structure of surnames as well as their surprising resilience in many cases despite such population changes.

The approach of Novotny & Cheshire (2012) abstracted surnames into a network space whereas Manni *et al.* (2005) and Boattini *et al.* (2012) maintain geographic relations through their use of self-organising maps (SOMs) on Dutch and Italian surnames respectively. The Manni *et al.* (2005) study identified areas of highest concentration and probable origin for 9000 of the most popular surnames in the Netherlands. This approach was further developed by Boattini *et al.* (2012) in the context of Italian surnames and is proposed as a general method to unravel population structures. In this study the origins of 49,117 different

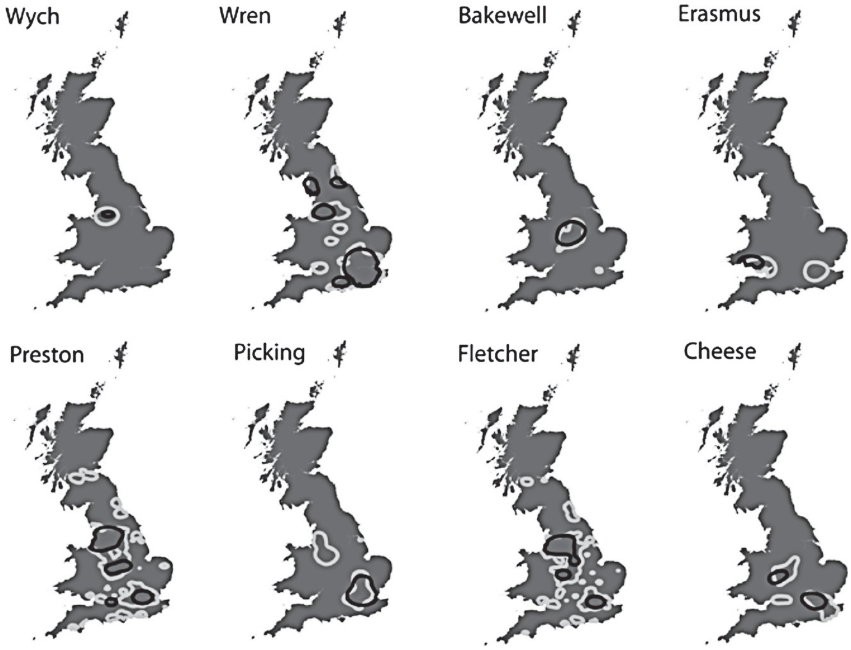

***Fig. 3 - The varying spatial extents of 8 British surnames between 1881(black) and 2001 (grey) as defined by the methodology of Cheshire and Longley (2012). In line with known population changes, there is a general trend towards more geographically extensive areas.***

surnames were identified and, crucially, validated against proven supervised methods of determining surname origins. The geographic information captured by the SOM methodology can be easily stored and queried in a database by researchers interested in constraining the spatial extent of their study or for excluding certain types of surname, such as those with multiple origins. Figure 2 offers an example of Boattini *et al.*'s (2012) outputs.

Darlu *et al.* (2011) (and in other papers) have developed an approach to calculate the "probability of geographic origin" of surnames. Whilst it is often presumed that the area of a surname's highest concentration is its likely place of origin, Darlu and colleagues calculate the weighted mean probability of geographic origin and feed it into a Bayesian formula, which is iteratively recalculated until a convergence criterion is met (see Degioanni & Darlu, 2001; Chareille & Darlu, (2010). This method is limited in that it requires data from two or more time periods, but it does offer the potential for more accurate insights into the origin of surnames than some of those previously discussed. The Bayesian approach can also differentiate between male and female migrations through the use of marriage registers and has the potential for gathering information about the genetic diversity of a particular area (Degioanni & Darlu, 2001). It seems well adapted to the analysis of hundreds, or even several thousand surnames, but the computational power required for the iterative Bayesian approach combined with the need for data from multiple time periods may make it impractical to deploy on large population registers.

The aforementioned research is reliant on pre-determined and discrete spatial units (such as administrative regions or prefectures). Discrete conceptualisations of geography are those bound to a particular set of underlying spatial units and can only show transitions at the borders of such units. These are the most common form of representation of surname data, and population data more broadly, and are therefore more widely used in analysis. The criteria for the creation of such units is often administrative and therefore subject to change over time, making it challenging to undertake temporal analysis.

The work presented by Cheshire & Longley (2012) overcomes some of the above challenges by combining a surface approach with the practical advantages of discrete spatial units. Cheshire & Longley (2012) achieve this through the use of kernel density estimation (KDE) to produce a surname density surface, or heat map. KDE transfers the input data (at a fine spatial scale) onto a regular grid of density estimates that can then be compared over time. To simplify the outputs, and facilitate comparison, a contour line is drawn around the areas of highest surname concentration [see Cheshire & Longley (2012) for full methodology]. Figure 3 shows the extent of 8 surnames in Great Britain in 1881 and 2001 delineated by their respective areas of highest concentration. It is clear that the majority of surnames in Britain have remained relatively static in terms of their highest concentrations according to this measure and that their spatial extents have also changed little. There are many more metrics that can be applied to surnames as a result of this methodology and they provide a useful means to partition large surname databases according to a range of pre-defined criteria.

It is clear that the identification of surname origins, either geographical or cultural, can have important applications in a broad range of fields. They are, however, limited to the study of surnames on an individual basis and therefore cannot offer comprehensive insights into the regional contexts in which the surnames, and their bearers, exist. The next section therefore addresses the ways in which surname distributions can be aggregated to reveal regional geographies.

## Surname regions

Studies that create regional geographies from surnames, like those interested in individual surnames, conceptualise geographic space as either continuous or discrete. Sokal *et al.* (1992) take a continuous approach to produce frequency surfaces of 100 surnames that are then combined to find common boundaries (dramatic changes in the frequency distribution). This study attempted direct comparisons between genetics and surnames by suggesting that abrupt surname boundaries were the result of barriers to population movement and mixing. Interestingly, the researchers found no such relationship, instead suggesting that the boundaries were the product of historical factors related to the origin of surnames (Sokal *et al.,* 1992). Nonetheless, stronger associations between surname transitions and physical barriers to gene flow have since been found in other studies beyond Britain through the use of barrier algorithms, such as Monmonier's algorithm (see Manni *et al.,* 2004). The most compelling is the implementation of the algorithm in the Ferrara Province of Italy, where a number of the identified barriers closely match topographic features known to restrict population movement. The approach, however, has only been published in a few studies from the same authors (Manni & Barrai, 2001; Manni *et al.,* 2004, 2008) and would therefore benefit from further research.

The use of discrete spatial units to create uniform surname regions is a common approach. Almost all studies follow a process of inductive generalisation to produce surname regions through the creation of a dissimilarity matrix. This is based on the quantitative comparison of the observed mix of surnames in spatial units used. One of the most popular methods for creating such a matrix was pioneered by Lasker during the 1970s and 1980s (Lasker, 1985). Lasker and colleagues selected surnames from telephone directories or marriage records to undertake studies of both regional [for example in Henley-on-Thames (Fox & Lasker, 1983) or Oxfordshire (Lasker, 1999)] and national level [see Mascie-Taylor & Lasker (1990) a Lasker (1985)]. Other popular methods exist, not least the Nei Distance (see for example Manni *et al.,* 2008), that serve to indicate the degree of similarity (or difference)

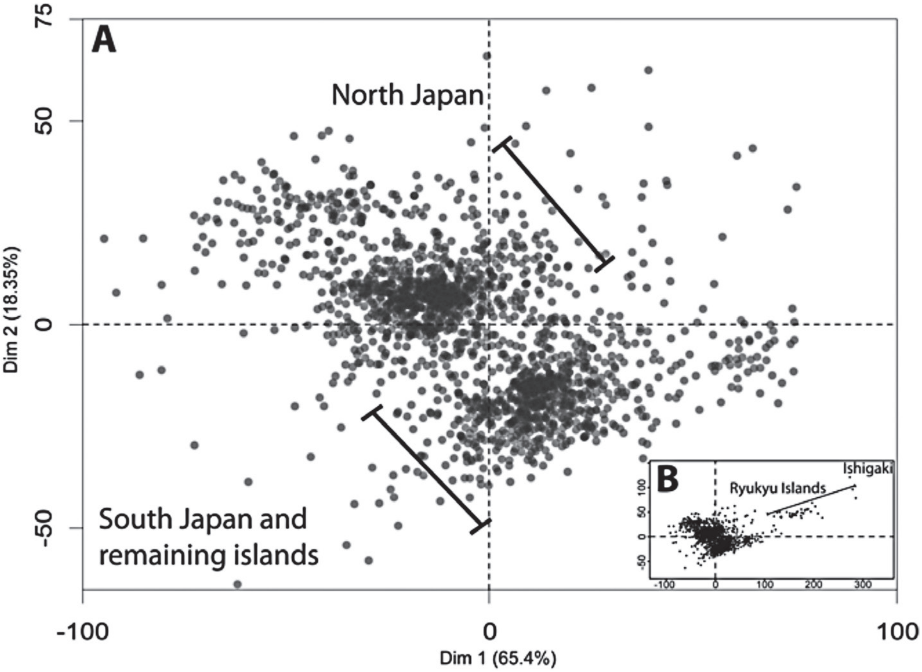

***Fig. 4 - Plots of the first two dimensions from PCA performed on a Lasker Distance matrix of Japanese municipalities. Approximate regions have been labeled as they appear in the cloud of points. The main plot (A) excludes the Ryukyu Islands that, as the inset (B) shows, exhibit significant variation from the majority of municipalities in Japan. Taken from Cheshire*** **et al.*****, 2013, p. 10).***

in the surname composition between geographic areas. It is not uncommon to calculate a range of measures as each are sensitive to different aspects of the surname distribution, such as the skew of the underlying population and the importance of small numbers of rare names in particular spatial units (see Manni *et al.,* 2008).

Surname dissimilarity matrices of the sort generated by the above measures can become extremely large given the number of pairwise comparisons required and therefore need summary measures to reveal key patterns. Hierarchical clustering algorithms, such as Wards, or stochastic approaches, such as K-Means, are popular in this context, as are data reduction techniques such as principal components analysis (PCA) and multidimensional scaling (MDS). Results from PCA and MDS can be plotted to give an impression of how well the geographic configuration of the spatial units matches their configuration based on surname dissimilarity. Figure 4, taken from Cheshire *et al.*'s (2013) study of Japan shows how PCA can reveal a great deal about the relationship between surname composition and physical barriers. Results from cluster analysis can be mapped, as each spatial unit is assigned a grouping to maximise within-region similarities and between- region differences.

These methods have been successfully used at the broad European level (Scapoli *et al.,* 2007; Cheshire *et al.,* 2011) through to the small-island level (Branco & Mota-Vieira, 2004). The bulk of this research has been undertaken at the national-level by one group of authors. Examples of their European papers include: Austria (Barrai *et al.,* 2000); Switzerland (Rodriguez-Larralde *et al.,* 1998a); Germany (Rodriguez-Larralde *et al.,* 1998b); Italy (Manni & Barrai, 2001); Spain (Rodriguez-Larralde *et al.,* 2003); Belgium (Barrai *et al.,* 2004); the Netherlands (Manni *et* *al.,* 2006); and France (Scapoli *et al.,* 2005). All studies demonstrate the spatial structure of surnames in the countries studied, with fewer commonalities between populations the further away they were.

Such differences are thought to be largely indicative of the different linguistic and cultural histories within or between the countries studied. This has been confirmed, on the European level at least, by a number of studies. In France (Scapoli *et al.,* 2005) and Belgium (Barrai *et al.,* 2004) the dialect transitions closely match those of surnames, meaning that measures of each can be used interchangeably. In the Netherlands, however, Manni *et al.* (2008) found no statistically significant relationship between surnames and dialect. This finding is interesting given the dialects of the country, and suggests that other factors can influence the surnames chosen. In the case of the Netherlands, Manni *et al.* (2008) cite religion as a possible explanation, with surname transitions occurring along the border between Protestant and Roman Catholic areas. The extent to which such differences (between surnames and language) are visible depends on the scale and linguistic diversity of the populations studied. If establishing broad surname regions in Europe, as Scapoli *et al.* (2007) have done, it is clear that language is the key determinant.

The Scapoli *et al.* (2007) study uses data from 8 European countries to create a continental-level surname regionalisation. It selected data for 2094 towns and cities grouped into 125 spatial units. Clear regionalisation patterns in surname compositions emerged, closely matching the national borders for the eight countries included, with exceptions showing the geo-historical distribution of languages. Whilst being extensive, both geographically and in terms of the number of surnames sampled, the work is still limited by its partial sampling of “representative” locations. The study by Cheshire *et al.* (2011) (see Figure 5) offers a couple of advances on previous work. The first relates to the volume and geographic extent of the surnames analysed, whilst the second is methodological, through the use of consensus clustering to create the regions. The paper covers 16 European countries - 8 million surnames from 152 million individuals - with data drawn from a range of sources including national population registers and telephone directories. The resulting regionalization, shown in Figure 5, shows 14 key surname regions in Europe inductively generated from the Lasker Distance calculation and subsequent consensus clustering.

The use of consensus clustering (see Monti *et al.,* 2003) has been used in a number of studies within the genetics community. It offers a range of metrics that help to determine both the optimal cluster solution, in terms of the number of clusters or regions, and the clustering method selected. These are important issues within the classification literature as different methods produce different results and the point at which the differences between groups cease to be significant is context dependent. What may be optimal in an image classification sense will not, for example, be the same for population datasets because clustering algorithms can be blind to a range of cultural factors identified in more qualitative research. The promise of consensus clustering, as discussed by Cheshire *et al.* (2011), is that it can offer the researcher a range of outcomes alongside a series of metrics that can be used to inform the final classification.

Classifications of surnames both on an individual basis, and on a regional basis, as outlined above, have been key drivers in the geographic analysis of surnames to date. Since much of this work is situated in the population genetics literature, more can be done to fully exploit surnames as a source of geographic data. The remainder of this paper will outline possible methodological advances before discussing the expansion of surnames research for a wider range of applications.

## Enhancing current approaches to geographic surname analysis

This next section seeks to highlight the merits of treating surnames as a source of geographic data and to outline potential improvements that will be of benefit to many of the studies undertaken to date. It is clear that surname research has undergone a transition from relatively data poor to extremely data rich and this, as Darlu *et al.* (2012) also note, will require the development and application of new statistical methods to better handle sampling and lemmatisation (the grouping of related surnames). In addition, this abundance of data offers the chance to address some of the challenges associated with geographically- referenced data.

## Refinements to the geographic analysis of surnames

Two broad approaches to the geographic analysis of surnames were discussed above. As yet, there has been no large-scale attempt to feed the outputs from the individual surname studies into those seeking broad regional geographies. Such a combination could be of significant value in the exclusion of polyphyletic surnames, for example.

Even if the capability exists to do so, the analysis of a complete population register is not always desirable. If surnames can be selected for study based on a number of relevant attributes this may lead to improved results in such contexts. The ability to process individual surnames automatically, in the way that Boattini *et al.* (2012), for example, have demonstrated with SOMs, offers the chance to generate such attributes en masse and will enable an informed sampling procedure based on a scalable classification methodology. The exclusion of polyphyletic surnames, for example, is commonplace in genetic studies but not in all studies of surname regionalisation. This may largely be due to the labour intensive way in which polyphyletic surnames have been identified through the study of each surname in turn. To create a representative regional geography, many thousands (if not millions) of surnames should be used, rendering manual approaches impractical. In some cases, therefore, the impact of polyphyly is intentionally overlooked.

Surname databases can also be further refined for specific applications based on a combination of the approaches outlined in this review. The work of Mateos *et al.* (2011) to create a global database of surname cultural-ethnic-linguistic (CEL) groups would, for example, benefit from the inclusion of some kind of spatial validation or comparison. This would serve to further disaggregate the bearers of a given surname based on their population history. For example in the USA the surname “Lee” is relatively common and classified as “British” but many of its bearers on the Pacific Coast are of Asian origin. Given that more recent migrant groups have different geographic distributions from early settlers (in the case of the USA) or native groups, a spatial clustering approach, such as that taken by Novotny and Cheshire (2012) or Degioanni & Darlu (2001) would identify these and could inform the CEL classification. One could therefore imagine the inclusion of “recent” or “established” migrant groups in the list of categories.

A further application of the kinds of geographic surname analyses described in this review is to sample design because they reveal, and therefore help to exclude (or include), members of a population who live in areas subject to large-scale changes. Such areas are most likely to be urban and have, in the past, been excluded based on arbitrary criteria – typically a distance threshold – in studies seeking to identify the genetic characteristics of a native population (see Winney *et al.,* 2011). Urban areas, however, are not uniform in either their shape or exposure to the currents of migration over time. It therefore follows that a more sophisticated approach to their identification, at least in terms of population structure, stands to improve existing practices. Surnames offer a means to do this so long as the geographic units used in their analysis are sufficiently small to enable urban areas to be meaningfully subdivided. As Figure 6 shows, this was the case in Longley *et al.* (2011) who were able to identify areas of similarity between urban areas due to their diverse populations comprising a large number of international migrants. Such areas can be easily excluded from further iterations of a sampling or regionalization strategy if necessary in to reduce the impacts of such groups.

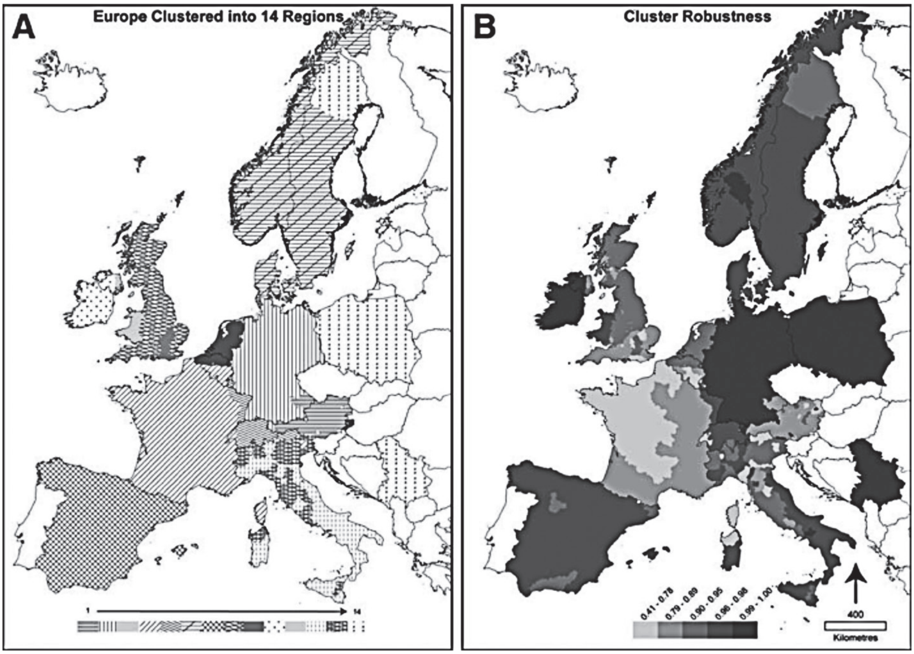

***Fig. 5 - Maps showing the spatial distributions of each of the 14 cluster allocations (A) and their respective robustness values (B). Higher robustness values represent a better result. On the left hand plot each cluster has been assigned a unique pattern. A full-color version can be found at spatialanalysis.co.uk/surnames. Taken from Cheshire*** **et al.*****, 2011, p.586.***

## Integrating spatial concepts

Given the relative lack of interest in surnames from geographers (Zelinsky, 1997) few studies have given serious consideration to a fundamental issue with spatial data - the Modifiable Areal Unit Problem (MAUP) (see Openshaw, 1984). The MAUP refers to the way in which the input geographic units of any analysis can impact the outcome of that analysis, especially in relation to the strength of correlation between variables. The significance of the geographic units used in surname studies cannot be overstated for this reason. In many cases the input geographic units are utilised, often because they are the only geographic referencing information available or have been created for administrative convenience (such as municipalities and government districts) and are therefore not subject to the same cultural influences that surnames have been. The spatial partitioning of the data before the analysis may therefore mask culturally distinct populations, for example if they comprise relatively small towns grouped together, whilst increasing the impact of relatively uniform areas if they are partitioned into a large number of geographic units. This is illustrated in Figure 7.

Address-level surname data (especially from sources such as digital telephone directories and government databases) are now easily obtainable so there is a reduced need to aggregate surname counts to large administrative areas. Whilst in many cases high levels of granularity are unnecessary (and may in fact be detrimental due to small numbers), a range of spatial units should be explored in any analysis to assess their impact on the results. This was the approach taken by both Longley *et al.* (2011) in their regionalisation of Great Britain and also Novotny & Cheshire (2012) in their analysis of Czech surnames. In both studies comparisons

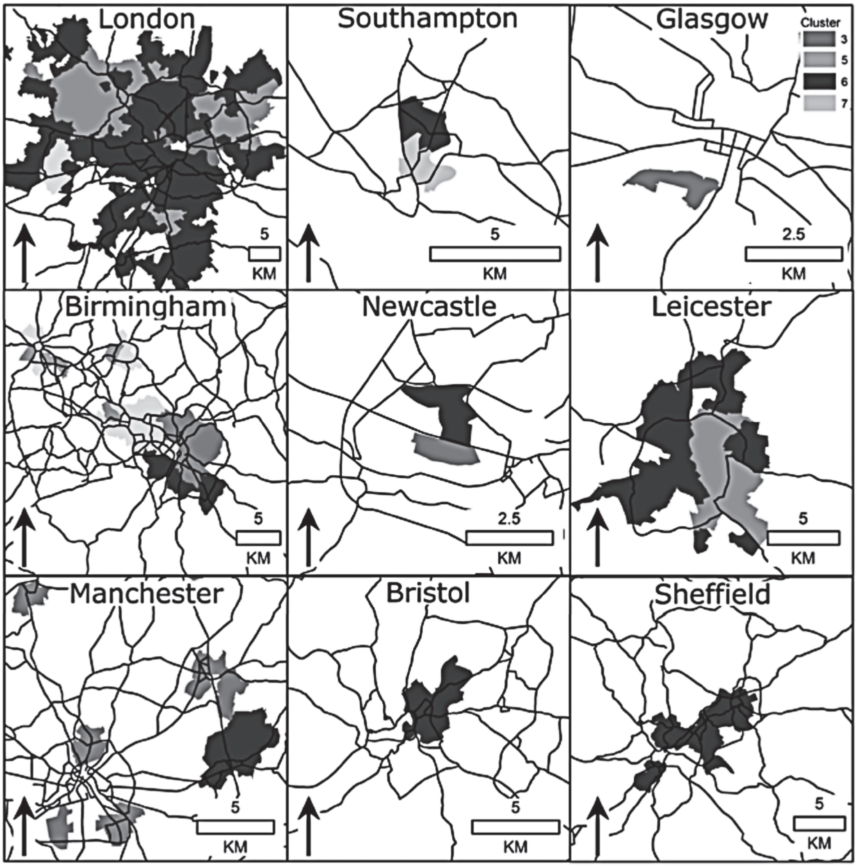

***Fig. 6 - Maps taken from Longley et al., 2011 (p. 512) illustrating the similarity in surname composition between 9 urban areas of Great Britain.***

were made between fine scale and broader scale spatial units. Such comparisons serve to remove speculation about the impact of geographic units on the results and can confirm the presence of distinct surname regions resulting from legitimate transitions in population.

In addition, a methodology for assessing the cultural significance of administrative geographies based on surnames has been proposed by Cheshire *et al.* (2013) using a comprehensive address level database. They compare Japan's surname regions to its current administrative geography and also its system of historic prefectures. Such work gives insights into the cultural significance of the geographic units used in the study of population data; something that surnames are well-placed to do with their clear links to past populations. Given the relative ease with which such comparisons can now be made, one could imagine such work as a routine precursor to any in-depth large-scale analysis of population data, provided the data are available.

## Surnames as geodemographic indicators

Beyond the technicalities of spatial units, geography and its sub-disciplines can offer further insights into population structure. Geodemographics, for example, is the analysis of people by where they live (Sleight, 1997) and is premised on the idea that the location of an individual, or group of similar individuals, provides useful demographic information (Harris *et al.,* 2005). Geodemographics in this broad sense is relevant for a range of reasons. For example, the work of Winney *et al.* (2011) can be thought of as producing a geodemographic classification based on what is revealed by the spatial distributions of people's surnames. This is one of a number of studies to explore how surnames can be used, for example, to improve sample designs in population genetics through targeting specific areas of surnames. This work, in principle, is little different from a retailer or health care provider that uses geodemographics to target specific population groups based on socio-economic characteristics. It can also be used as a basis to attach relevant genetic attributes to conventional social surveys.

Previous surname research has sought to represent similarities or differences between contiguous regions, largely in the context of clinal (or gradual) changes in the distributions punctuated by the occasional abrupt transition. There has been relatively little interest in the apparently close linkages, in terms of surname compositions, between regions that are geographically distant. Only one study, Longley *et al.* (2007), has sought to establish any causality and impacts of these linkages. Such regions are often quite small and therefore likely to have been missed by the relatively coarse granularity of previous research. The idea that similarity/ difference is more than the product of geographical distance has long been recognised in geodemographics and represents important exceptions to the idea that populations in close geographical proximity are likely to be more similar than those further apart (based on the often quoted "Tobler's First Law of Geography" (Tobler, 1970). In Longley *et al.* (2011), for example, the town of Corby in central England appears to have more in common with Scotland in terms of its surname composition than with its neighbouring areas. This reflects the large number of Scottish migrants and their descendants who live in the town. Corby's link to Scotland would be overlooked if there is the presumption that geographic proximity is the only driver of social similarity.

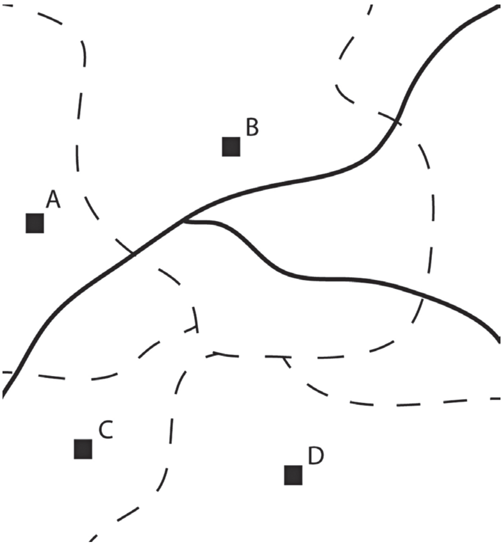

***Fig. 7 - A simple demonstration of the impact of differing administrative units on the geographic partitioning of data. If the dashed boundaries are used, all settlements (A, B, C and D) are treated separately in any pairwise comparisons or other analysis, whilst the solid boundaries group A with B and C with D. The solid boundaries may therefore smooth any variation if, for example, C and D were settlements with different naming conventions whilst analysis using the dashed boundaries will detect it.***

The link between surnames and demographic characteristics more broadly is worthy of further research. Some, such as Clark (2013), believe they can be a measure of social mobility in specific cases, but on a more general level surnames can be usefully applied in a range of contexts, not least healthcare. Petersen *et al.* (2011) were, for example, able to determine the utilisation of healthcare services by different ethnic groups as identified by their surnames. The authors concede that surname classification may not differentiate between recent and settled migrants (or those from later generations) - such a distinction may be important if there are different behaviours between these groups - but the extent to which this is the case is unclear (Petersen *et al.* 2011). Within the healthcare domain there are many studies that use surname lists to identify patients who are at risk from particular illnesses more prevalent in certain ethnic groups (see for example Nasseri, 2007; Fiscella & Fremont, 2006). The approaches taken in these studies are more simplistic than those described above and so healthcare presents a research domain that could stand to benefit a great deal from the geographic analysis of surnames.

Conceiving of surnames as geodemographic indicators facilitates a link to a broader range of research applications. Healthcare, discussed above, is one of just one domain that could benefit from the wider utilization of surnames and their associated metrics. The final two substantive sections of this paper touch on a few more research areas where surname data could offer significant benefits.

## Expanding applications

Technological innovations are facilitating access to new datasets that would benefit from the kinds of analysis honed over the many decades of surname research. At the very least, studies of surname histories conducted by academics or genealogists should be integrated into large-scale databases and used to offer depth to the breadth provided by inductive approaches. This is something that commercial genealogy companies will have and researchers could negotiate access to. The information carefully collated by enthusiasts may, for example, validate some of the automated approaches discussed above since there is a critical mass of individuals and groups keen to improve surname datasets. One excellent example of the depth that volunteered information can offer is the Church of Jesus Christ of Latter Day Saints' *FamilySearch* (familysearch.org) website that hosts over 800 million user-contributed surnames and other pieces of family history information (see Otterstrom 2008 for more information).

The explosion in online social networking websites creates a further means to discover genealogical connections between distant relatives with similar geographic histories (Otterstrom & Bunker, 2012; Timothy & Guelke, 2008). Whilst not all the data generated from websites such as Facebook and Twitter are easily accessible, companion applications are able to, with the users consent, collect data. In addition, many companies are keen to gather as many demographic indicators about their users as possible and surnames represent an enticing way to gauge ethnic and cultural groupings. Chang *et al.* (2010) have, for example, derived the ethnic composition of Facebook's users using a simple name-based ethnicity classification, whilst Ambekar *et al.* (2009) achieved a basic ethnicity classification from Wikipedia data. It is only a matter of time before similar methodologies are applied to other social media platforms, such as Twitter.

Surnames can also be used to gauge the representativeness of social media data of the population at large. One could, for example, perform simple comparisons between the distribution of surnames in a particular area from social media and the distributions from an official population register to reveal selectivity among certain cultural groups. In addition, the linkages between individuals on such sites could serve as a useful validation of whether those sharing surnames with similar geographic distributions are more likely to be in contact. Finally, the creation of an online network of relatively alike (in the context of surnames at least), individuals would make the recruitment for any genetic sampling study much more straightforward as the information about future events could be efficiently targeted and disseminated.

The ethics of such work are yet to be fully considered and likely to be subject to concerns similar to those leveled at research into population relatedness more generally. Concerns have been raised about the use of surnames and population genetics research for DNA databases compiled by governments and their security services. It would be unhelpful if the automated analysis of surnames from new forms of interaction, such as social media platforms, were to be perceived in this way as some kind of "biopolitical informational gathering" (Nash, 2005, p. 457). In practice, however, it appears that the enthusiasm for discovering personal ancestry outweighs concerns about this practice (and the tools used to do it) so the importance of social media in this domain is likely to grow.

The innovative classification of surnames has also been successfully applied to a number of more established datasets. In practice it is often not possible with conventional questionnaires to get a highly detailed sense of the ethnic/cultural composition of an area (Aspinall, 2009). The development of automated methods to discern surname origins offers the potential to create highly disaggregate classifications of names (and therefore their bearers) without the need for questionnaires or other contextual data. The results will not have the same level of accuracy, but they offer a marked improvement in representing smaller population groups. These can also be compared to, and validated against, more in-depth social surveys to get a sense of the accuracy of any name-based classification and to determine any systematic errors in the methodology. As has been discussed above, the inclusion of spatial information about the distribution of surnames at the sub-national level may serve to improve name-based ethnicity classification further and address some of its limitations related to cultural groups sharing similar naming preferences.

## Future trends

In the concluding part of their review Darlu *et al.* (2012) identify six future research trends that align well with what has been discussed here. The six areas include: the determination of the most probable geographical, temporal and cultural origin of surnames; the identification of common surname lineages; the identification of barriers to cultural and population interaction; and a synthesis of current advances to tease out different population episodes across space and time. Despite Darlu *et al.*'s (2012) call for an interdisciplinary approach, these appear quite specific and can still be confined to the domains of anthropology and population genetics. The further research proposed here therefore seeks to augment those listed above through three key themes: further development of new methodological approaches; more attention to non-Western countries and those with less stable population histories; an embracing and expanding of new applications for surname data.

With regard to the first, many of the methodological approaches applied to surname data are concerned with the identification of spatial patterns or pairwise comparisons between geographic areas. Such concerns are not unique to surname datasets, with fields such as semantics, dialectics, economics and regional science having much to offer. Novotny & Cheshire (2013) for example used the Dice and Jaccard coefficients from economics to build a surname network, and Longley *et al.* (2011) used a number of innovative visualisation techniques, such as word clouds, to demonstrate the surname compositions of particular regions. In addition, methodologies exist to create optimal configurations of geographic units at multiple scales (see for example Openshaw & Rao, 1995) and these may be of use if surname analysis is to move beyond the often rigidly defined administrative units it currently relies on. Sensitivity analysis of the results with different geographic units and explorations into their cultural significance (see Cheshire *et al.,* 2013) will offer interesting insights and a fresh perspective on the results of much of the analysis conducted to date. Arguably, the biggest limitation of the growing number of large datasets available to surname research is their relatively low signal to noise ratio in comparison to carefully compiled (but much less extensive) datasets. The integration of multiple techniques (as discussed in this review) offers the chance to address this through the iterative refinement of large-scale databases utilising informed data mining and surname classification.

The second theme - to move away from Western countries and those with stable population histories - seeks to widen the appeal of surnames as general demographic indicators. Western countries provide the focus for much of the research outlined in this review and so the move east or towards developing countries offers the potential for gaining fresh insights into the population structure of such countries. It is acknowledged that there has been surname research in many of these regions, not least in Japan (see Cheshire *et al.*, 2013; Takemitsu, 1998) and several countries in South America (see for example Barrai *et al.*, 2012; Rodriguez-Larralde *et al.*, 2011; Dipierri *et al.,* 2011), but the potential remains to gain the depth of insight achieved in the literature pertaining to European surnames. In addition, there have been few studies on the significance of surnames in places where there has been substantial population change either through forced migration, such as in the Czech Republic (Novotny & Cheshire, 2012), or though the gradual ebb and flow of population change. In parts of the world with large migrant populations the focus tends to be on excluding them in order that the native population can be studied (see for example Alonso & Usaqúen, 2012). The reasons for this stem from the interest in surnames as indicators of genetic structure, but as surname databases become more widely available they can be used as more general social indicators. Longley *et al.* (2007), for example, used surnames to identify migrants who moved from one part of Great Britain (Cornwall) to another (Middlesbrough) and numerous other studies have identified areas with relatively large migrant communities. From a social science perspective, many of these areas are the most dynamic and interesting. As with the healthcare examples discussed above, surname classifications can provide additional attributes to augment standard social datasets and thus offer a fresh layer of insight. It is up to existing surname researchers to demonstrate this and promote the wider use of surnames in the social sciences.

Darlu *et al.* (2012) also see the future of surname studies as lying more in the "rich information provided by the set of data preserved through the generations…and in well-defined communities, than in the accumulation of surnames on a wider geographical scale" (p. 172). It is hoped that this review demonstrates that whilst this may be a dominant direction for surnames research it need not be the only one. The final theme - an encouragement to move beyond traditional applications - reflects this. Interest in surname research is set to grow beyond what can be seen as its current core disciplines. There is much to offer fields such as social media analysis, volunteered surname data and social surveys. In addition, the challenges surrounding "big data" are something many surname researchers are well placed to address, being experienced in extensive data mining and classification methodologies. In short, there are benefits to learning new approaches from a range of disciplines, but this should not be a one-sided relationship; many of the lessons learnt from surname analysis can offer insights to other disciplines.

## Conclusion

This review paper has sought to demonstrate the importance of the analysis of surnames as a geographic dataset. It has discussed a range of studies that, through their surname analysis, have generated new insights into population structure. It is hoped that the wealth of research presented here inspires further work in order that surnames become a more widely used source of data both within the geography community and beyond. There is potential for integration of the two key research strands that focus on individual surname distributions and regional characteristics respectively. It is also clear that, aside from methodological developments, the breadth and depth of surname research is set to increase and become more interdisciplinary as new population datasets from the likes of social media become available and require the insights that surnames provide. Surname research is therefore in a very strong position with many new avenues to pursue, not least by geographers and anthropologists, over the coming years.

## Info on the web

http://worldnames.publicprofiler.org/

*Map the distribution of a surname across approximately 300 million people in 26 countries of the world.*

http://gbnames.publicprofiler.org/
*Map both historic (1881) and contemporary surname distributions for Great Britain.*

https://familysearch.org
*One of the largest repositories of surname data available, this website is a useful starting point for detailed surname studies.*

http://www.peopleofthebritishisles.org/
*This project is one of the largest and most detailed studies into the genetics' of the population of the British Isles.*

## References

Alonso L. & Usaqúen W. 2012. Y-Chromosome and surname analysis of the native islanders of San Andrés and Providencia (Colombia). *Homo* (in press).

Ambekar A., Ward C., Mohammed J., Male S. & Skiena S. 2009. Name ethnicity classification from open sources. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 49-58.

Aspinall P. J. 2009. The future of ethnicity classifications. *Journal of Economic and Migration Studies,* 35: 1417-1435.

Barrai I., Rodriguez-Larralde A., Mamolini E., Manni F. & Scapoli C. 2000. Elements of the surname structure of Austria. *Ann. Hum. Biol.,* 27: 607-622.

Barrai, I. Rodriguez-Larralde, A. Manni, F. Ruggerio, V. Tartari, D. Scapoli, C. 2004. Isolation by language and distance in Belgium. *Ann. Hum. Genet.,* 68, 1: 1-16.

Barrai I., Rodriguez-Larralde A., Dipierri J., Alfaro E., Acevedo N., Mamolini E., Sandri M., Carrieri A. & Scapoli C. 2012. Surnames in Chile: a study of the population of Chile through isonymy. *Am. J. Phys. Anthropol.,* 147: 380-388.

Boattini A., Lisa A., Fiorani O., Zei G., Pettener D. & Manni F. 2012. General method to unravel ancient populaton structures through surnames, final validation on Italian data. *Hum. Biol.,* 84: 235-270.

Bowden G.R., Balaresque P., King T.E., Hansen Z., Lee A.C., Pergl-Wilson G., Hurley E., Roberts S.J., Waite P., Jesch J., Jones A.L., Thomas M.G., Harding S.E. & Jobling M.A. 2008 Excavating past population structures by surname-based sampling: the genetic leagcy of the Vikings in Northwest England. *Mol. Biol. Evol.,* 25: 301-309.

Branco C. & Mota-Vieira L. 2004. Population structure of São Miguel Island, Azores: a surname study. *Hum. Biol.,* 75: 929-939.

Chang J., Rosenn I., Backstrom L. & Marlow C. 2010. ePluribus: ethnicity on social networks. *Proceedings of the Fourth International AAAI Conference of Weblogs and Social Media,* pp. 18-25.

Chareille P. & Darlu P. 2010. Anthroponymie et migration: quelques outlies d'analyse et leur application à l'étude des déplacements dans les domaines des Saint-Germain-des-Prés au IX$^{e}$ siécle. In Bourin M. & Martinez Sopena P. (eds): *Anthroponymie et migrations dans la Chrétienté Médiévale,* pp. 41-73. Casa de Velàzquez Madrid, Spain (Collection de la Casa de Velàzquez 116).

Cheshire J. A., Longley P. A., Yano K. & Nakaya T. 2013. Japanese surname regions. *Papers in Regional Science,* Doi: 10.111/pirs.12002.

Cheshire, J. A. and Longley, P. 2012. Identifying spatial concentrations of surnames. *International Journal of GIS*. 26: 309-325.

Cheshire J. A., Mateos P. & Longley P.A. 2011. Delineating Europe's cultural regions: population structure and surname clustering. *Hum. Biol.,* 83: 573-598.

Clark G. 2013. Surnames and social mobility. *American Economic Association Annual Meeting Papers.* http://www.aeaweb.org/aea/2013conference/program/meetingpapers.php (accessed April 2013).

Colantonio S., Lasker G., Kaplan B. & Fuster V. 2003. Use of surname models in human population biology: a review of recent developments. *Hum. Biol.,* 75: 785-787.

Darlu P., Bloothooft G., Boattini A. & Brouwer L. 2012. The family name as socio-cultural feature and genetic metaphor: from concepts to methods. *Hum. Biol.,* 84: 169-214.

Darlu P., Brunet G. & Barbero D. 2011. Spatial and temporal analyses of surname distributions to estimate mobility and changes in historical demography: the example of Savoy (France) from the eighteenth to the twentieth century. In Gutmann M.P., Deane G.D., Merchant E.R. & Sylvester K. (eds): *Navigating Time and Space in Population Studies,* International Studies in Population 9. Springer, New York.

Degioanni A. & Darlu P. 2001. A bayesian approach to infer geographical origins of migrants through surnames. *Ann. Hum. Biol.,* 28: 537-545.

Diperri J., Rodriguez-Larrralde A., Alfaro E., Scapoli C., Mamolini E., Salvatorelli G., Caramori G., De Lorenzi S., Sandri M., Carrieri A. & Barrai I. 2011. A study of the population of Paraguay through isonymy. *Ann. Hum. Genet.,* 75: 678-687.

Fox W. & Lasker G. 1983. The distribution of surname frequencies. *Int. Stat. Rev.*, 51: 81-87.

Fiscella K. & Fremont A. 2006. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research,* 41: 1482-1500.

Guppy H. 1890. *Homes of family names in Britain.* Harrison and Sons, London.

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

Kaplan B. & Lasker G. 1983. The present distribution of some English surnames derived from place names. *Hum. Biol.,* 55: 243-250.

Lasker G. 1985. *Surnames and genetic structure.* Cambridge University Press, Cambridge.

Lasker G. 1999. The hierarchical structure of an urban town, Kidlington, Oxfordshire examined by the coefficient of relationship by isonymy. *J. Biosoc. Sci.,* 31: 279-284.

Longley P., Cheshire J. & Mateos P. 2011. Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum,* 42: 506-516.

Longley P., Webber R. & Lloyd D. 2007. The quantitative analysis of family names: historic migration and the present day neighborhood structure of Middlesbrough, United Kingdom. *Ann. Assoc. Am. Geogr.,* 97: 31-48.

Manni F. & Barrai. I. 2001. Genetic structures and linguistic boundaries in Italy: a microregional approach. *Hum. Biol.,* 73: 335-347.

Manni F., Heeringa W., Toupance B. & Nerbonne J. 2008. Do surname differences mirror dialect variation? *Hum. Biol.,* 80: 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. *Hum. Biol.,* 76: 173-190.

Manni F., Heeringa W. & 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: 507-528.

Mascie-Taylor C. & Lasker G. 1990. The distribution of surnames in England and Wales: a model for genetic distribution. *Man,* 25: 521-530.

Mateos P. 2007. A review of name-based ethnicity classification methods and their potential in population studies. *Popul. Space Place,* 13: 243-273.

Mateos P., Longley P.A. & O'Sullivan D. 2011. Ethnicity and population structure in personal naming networks. *PLoSONE,* 6: e22943. doi:10.1371/journal.pone.0022943

Monti S., Tamayo P., Mesirov J. & Golub T. 2003. Consensus clustering: a resampling based method for class discovery and visualisation of gene expression microarray data. *Machine Learning,* 52: 91-118.

Nash C. 2005. Geographies of relatedness. *Trans. Inst. Br. Geogr.,* 30: 449-462

Nasseri K. 2007. Construction and validation of a list of common Middle Easter surnames for epidemiological research. *Cancer Detect. Prev.,* 31: 424-429

Novotny J. & Cheshire J. A. The surname space of the Czech Republic: examining population structure by network analysis of spatial co-occurrence of surnames. *PloSONE,* 7: e48568. Doi: 10.1371/ journal.pone.0048568

Openshaw S. 1984. *The modifiable areal unit problem*. CATMOG 38, GeoBooks, Norwich.

Openshaw S. & Rao L. 1995. Algorithms for reengineering 1991 census geography. *Environ. Plan. A*, 27:425-446.

Otterstrom S. 2008. Genealogy as religious ritual: the doctrine and practice of family history in the Church of Jesus Christ of Latter Day Saints. In Timothy G. & Guelke J. (eds): *Geography and genealogy: locating personal pasts,* pp. 137-151. Burlington VT, Ashgate.

Otterstrom S. & Bunker, B 2012. Genealogy, migration, and the intertwined geographies of personal pasts. *Ann. Assoc. Am. Geogr.,* Doi: 10.1080/00045608.2012.700607.

Petersen J., Longley P., Gibin M., Mateos P. & Atkinson P. 2011. Names-based classification of accident and emergency department users. *Health Place,* 17: 1162-1169.

Rodriguez-Larralde A., Scapoli C., Beretta M., Nesti C., Mamolini E. & Barrai I. 1998a. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. *Ann. Hum. Biol.,* 6: 533-540.

Rodriguez-Larralde A., Barrai I., Nesti C., Mamolini E. & Scapoli C. 1998b. Isonymy and isolation by distance in Germany. *Hum. Biol*., 70: 1041-1056.

Rodriguez-Larralde A., Gonzales-Martin A., Scapoli C. & Barrai I. 2003. The names of Spain: a study of the isonymy structure of Spain. *Am. J. Phys. Anthropol.,* 121: 280-292.

Rodriguez-Larralde A., Dipierri J., Gomez E.A., Scapoli C., Mamolini E., Salvatorelli G., De Lorenzi S., Carrieri A. & Barrai I 2011. Surnames in Bolivia: a study of the population of Bolivia through isonymy. *Am. J. Phys. Anthropol.,* 144: 177-184.

Scapoli C., Goebl H., Mamolini E., Rodriguez-Larralde A. & Barrai I. 2005. Surnames and dialects in France: population structure and cultural evolution. *J. Theor. Biol.,* 237: 75-86.

Scapoli C., Mamolini E., Carrieri A., Rodriguez-Larralde A. & Barrai I. 2007. Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. *Theor. Popul. Biol*., 71: 37-48

Sleight P. 1997. *Targeting customers: how to use geodeomographic and lifestyle data in your business.* NTC Publications, Henley on Thames.

Sokal R., Harding R., Lasker G. & Mascie-Taylor C. 1992. A Spatial Analysis of 100 surnames in England and Wales. *Ann. Hum. Biol.,* 19: 445-476.

Takemitsu M. 1998. *Surnames and Japanese.* Bungeishunju, Tokyo.

Timothy D. J. & Guelke J. K. (eds) 2008. *Geography and genealogy: locating personal pasts.* Ashgate, Aldershort, UK.

Tobler W. 1970. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.,* 46: 234-241.

Tucker D. K. 2001. Distribution of forenames, surnames and forename-surname pairs in the United States. *Names,* 49: 69-96.

Tucker D. K. 2005. The cultural-ethnic-language group technique as used in the Dictionary of American Family Names (DAFN). *Onomastica Canadiana,* 87: 71-84.

Webber R. & Longley P. 2003. Geodemographic Analysis of Similarity and Proximity: Their Roles in the Understanding of the Geography of Need. In Longley P. & Batty M. (eds): *Advanced Spatial Analysis: The CASA Book of GIS* 233. ESRI Press, Redlands.

Winney B., Boumertit A., Day T., Davison D., Echeta C., Evseeva I., Hutnik K., Leslie S., Nicodemus K., Royrvik E. C., Tonks S., Yang X., Cheshire J., Longley P., Mateos P., Groom A., Relton C., Bishop D. T., Black K., Northwood E., Parkinson L., Frayling T. M., Steele A., Sampson J. R., King T., Dixon R., Middleton D., Jennings B., Bowden R., Donnelly P. & Bodmer W. 2012. People of the British Isles: preliminary analysis of genotypes and surnames in a UK control population. *Eur. J. Hum. Genet.,* 20: 203–210Valetas M. F. 2001. The surname of married women in the European Union. *Bulletin Mensuel D'Inforamation De L'Intstitut National D'Etudes Demographiques*, 367.

Zelinsky W. 1997. Along the frontiers of name geography. *Prof. Geogr.,* 49: 465-466

Editor, Giovanni Destro Bisol