Annals of the American Association of Geographers

Routledge
Taylor & Francis Group

Family name origins and inter-generational demographic change in Great Britain

Journal:	*Annals of the American Association of Geographers*
Manuscript ID	AN-2019-1132.R2
Manuscript Type:	Revised Manuscript
Manuscript Category:	Geographic Methods
Key Words:	surnames, family names, population change, Great Britain

Family name origins and inter-generational demographic change in Great Britain

Abstract

We develop bespoke geo-spatial routines to typify 88,457 surnames by their likely ancestral geographic origins within Great Britain. Linking this taxonomy to both historic and contemporary population datasets, we characterise regional populations using surnames that indicate whether their bearers are likely to be long-settled. We extend this approach in a case study application, in which we summarise intergenerational change in local populations across Great Britain over a period of 120 years. We also analyse much shorter-term demographic dynamics and chart likely recent migratory flows within the country. Our research demonstrates the value of family names as markers of migration and change at regional and local scales. It also has important implications for understanding the population profiles of different places, as well as the inter-generational outcomes of migration.

Keywords

surnames; family names; population change; Great Britain

1. Introduction

Most long established family names (surnames) have remained regionally concentrated within Great Britain over centuries (Longley, Cheshire, and Mateos 2011; Cheshire and Longley 2012) and similar tendencies have been observed in other European countries, the United States, Japan, China, and Latin America (Rodríguez-Larralde et al. 1998; Rodríguez-Larralde, Morales, and Barrai 2000; Barrai et al. 2001; Cheshire et al. 2014; Herrera Paz et al. 2014; Orrù et al. 2018; Shi et al. 2019). Sociological and geographical surname research has demonstrated that this concentration is an outcome of regionally varying naming practices and socio-cultural processes of surname transmission (Cheshire 2014; Mazières and Roth 2018). As a result, surnames have been used in diverse applications, such as estimating genetic population structure (King and Jobling 2009; Darlu et al. 2012; Kandt, Cheshire, and Longley 2016), social mobility (Clark and Cummins 2015), ethnic segregation (Mateos 2007; Lan, Kandt, and Longley 2018) and health inequalities (Petersen et al. 2011; Lewis and Longley 2012).

While there is now significant experience in building surname classifications to infer national origins (Mateos, Longley, and O'Sullivan 2011; Kandt and Longley 2018), there has been much less systematic research into surname taxonomies of subnational, regional origins or their legacies of distinctive local and regional surname mixes. There are some exceptions mainly in the discipline of population genetics. The study by Manni et al. (2005) on Dutch surnames represents a rare attempt to use spatial concentrations of surnames as indicators for ancestral, geographic surname origins. (Degioanni and Darlu 2001; Bloothooft and Darlu 2013) use surname geographies to estimate short-term regional migratory flows within countries. These studies demonstrate the power of surname geographies to characterise population origins and change over time and contextualise a range of socio-spatial outcomes, such as inequalities in social mobility and health (Degioanni, Darlu, and Raffoux 2003;

Longley, Webber, and Lloyd 2007; Winney et al. 2012). In this paper, we develop a regional surname taxonomy for Great Britain and demonstrate its value in a research case study that couples family name records from a historic (1881) population census with contemporary Consumer Registers to describe inter- and intra-generational stasis and movement of local populations.

Fundamental to our analysis is the observation that the surnames of the long-established population groups of Great Britain possess properties that render them informative markers of geographic origins and processes. First, in many societies, surnames are hereditary, usually passed down along the patrilineal line (Jobling 2001). Second, most surnames have distinctive local or regional origins reflecting occupation, toponymic feature, or family lineage (Cheshire and Longley 2012). As a result, most surnames do not occur with equal frequency throughout Great Britain but are regionally or locally clustered. Past research has demonstrated that the distribution of individual surnames has remained remarkably stable ever since they first came into common parlance between the 12th and the 14th Centuries (Cheshire and Longley 2012). Given the historical and contemporary local nature of most intermarriage, individuals consequently inherit and pass on local surnames (Cheshire 2014). While the regional distinctiveness of surnames is well established at an international level for the purpose of inferring ethnic origin (e.g. Lan, Kandt, and Longley 2018), the use of surnames as historic regional and ancestral markers has rarely been systematically explored for entire populations isolated from effects of international migration.

In this study, we use historic census and consumer data to develop a two-stage surname classification scheme. We first use 1881 Census of Population data to identify a hierarchy of regions with distinct surname mixes and assign each of 88,457 British surnames to a region within this scheme. We then use this same regional classification to characterise local population change over several generations.

2. Data sources

Individual level historic census records including personal names and addressess are made publicly available under the ‘100-year rule’ governing release of UK Census data. Higgs and Schürer (2014) describe how these data and have been digitally encoded for most of the period 1851-1911. Addresses are also linked to parishes, the boundaries of which have been digitised (Southall 2012, 2014). It is thus possible to ascribe names to small areas and to estimate detailed surname geographies of the country.

In the analysis developed here, we use personal name and location data from the 1881 Census of Population (Schürer and Woollard 2000a, 2000b) to trace the geographic origins and subsequent diffusion of family names, choosing this source over data for earlier time periods because of the high standards that were used in digital encoding. Use of this source also enables direct comparison of our findings with previous research (e.g. Cheshire and Longley 2012; Cheshire 2014; Kandt, Cheshire, and Longley 2016).

Today, commercial Consumer Registers (DataTalk, St. Ives: www.datatalk.co.uk; CACI Ltd., London: www.caci.co.uk) compiled from the public version of the Electoral Register and consumer files record the names and addresses of adult individuals who consent to their inclusion (Lansley, Li, and Longley 2018). These Consumer Registers can be used for research purposes under the provisions of the EU General Data Protection Regulation and, although not achieving universal adult coverage, comparisons with the Office for National Statistics Mid Year Population Estimates confirm that they include a very large and representative proportion of the adult population (Lansley, Li, and Longley 2019). Recorded addresses include postcodes, which can be geocoded using the postcode directory available at census.ac.uk. In the research reported here, we use the Consumer Registers for 1998 and 2016, being one of the earliest, and the most recent, Consumer Registers available to us, respectively (see also Longley, Cheshire, and Singleton 2018).

Temporal comparison for small areas requires a consistent zonal design across different points in time. To this end, we develop an algorithm to create temporally harmonised zones based on 1881 parishes and contemporary (2011) Office for National Statistics Middle Layer Super Output Areas (MSOAs). The geographical extents of the 15,748 parishes used in the 1881 Census vary considerably: whereas parishes are typically of much smaller extent in urban areas than MSOAs, the reverse is frequently true in rural areas. We begin by merging 1881 parishes with their neighbours until their resulting population sizes reaches a minimum of 750 inhabitants. Only parishes located in the same Census District are merged in this procedure. The centroid of each of the merged parishes is then assigned to a 2011 MSOA using a point-in-polygon procedure. Subsequently, we assign MSOA centroids to the merged parish groups. In the next step, we merge parish data that were assigned to a common MSOA and, in parallel, merge MSOAs that were assigned to the same parish. Zones — either parishes or MSOAs — are merged with neighbouring zones when they do not have a single centroid within their boundary.

This procedure generates two lookup tables, the first assigning individual 1881 parishes to MSOAs and the second assigning MSOAs to groups of parishes. We combine these into a single lookup table assigning each individual parish to an MSOA group, which in turn is assigned to a parish group. It then becomes an arbitrary decision whether MSOA or parish groups are to be used as the temporally harmonised zones (THZ); we choose MSOA groups, of which there are 3,828. Finally, we assign the unit postcodes recorded in the Consumer Registers to the appropriate THZs.

3. A regional surnames taxonomy

To prepare the design of a regional surname taxonomy, we first filter out all individuals who bore surnames that do not originate in Great Britain, as defined using the Onomap classification tool of Mateos, Longley, and O'Sullivan (2011). This is necessary because these surnames tend to cluster in larger cities and conurbations, and including them would skew the analysis. The largest groups of names classified as having origins outside Great Britain in the 1881 Census include individuals with 'Jewish' and 'Sikh' surnames as well as surnames having origins in Italy, Germany and France. The presence of these names in the 1881 Census is unsurprising given the various waves of immigration that 'Great Britain' has witnessed in the many centuries before 1881. For instance, large numbers of French Huguenot refugees arrived at the end of the seventeenth century, while during the Nineteenth Century many Germans came to Great Britain fleeing war and searching for new opportunities. Fallout of the Napoleonic wars and outbreaks of typhus and cholera in northern Italy also led to importation of surnames from Continental Europe. The presence of bearers of Sikh names can be largely ascribed to activities of the East Indian Company which recruited many lascars to work on its ships. Jewish names have a longer lineage in Great Britain, often traceable to attempts to escape persecution elsewhere in Europe (Winder 2014).

Surnames were considered to have originated from the British Isles if they fell into one of the following Onomap categories: 'English', 'Celtic', 'Cornish', 'Irish', 'Northern Irish', 'Scottish', or 'Welsh'. From a total of 420,192 unique surnames, a total of 88,457 surnames were classified as having roots in Great Britain in 1881 (Table 1). While, according to this classification, just 21 per cent of all surnames found in the 1881 Census are "British", this share encompasses 96 per cent of the 1881 Census population. The population with non-British surnames, who nevertheless may be long-settled and British, is therefore sparse and we expect that their exclusion has little impact on further analysis. Of all British surnames in 1881, 76,089 (86 per cent) were still present in the 2016 Consumer Register. This 14 per cent drop likely reflects intergenerational extinction of rare surnames, which is a common phenomenon in societies with unilineal surname transmissions (Yasuda et al. 1974). In our case, 95 per cent of all surnames becoming extinct by 1998 had fewer than 35 bearers in

1881. Additionally, this period of 117 years spans two world wars entailing casualties of younger, male soldiers, which further increased the risk of surname extinction. Slightly varying data collection methods used in successive Consumer Registers may explain some limited deviations of results, such as the very slight increase in the number of surnames by 0.3 per cent between 1998 and 2016. For all years, the average number of bearers of British names (323.07 in 1881) is much higher than the average number of bearers of non-British names (3.70 in 1881).

<< Table 1 >>

Defining ancestral surname regions

In order to define suitable regions representing ancestral surname concentrations, we estimate the frequencies of each British surname in each THZ in 1881. We then compare surname frequencies across THZs and generate a pair-wise matrix measuring the dissimilarity in surname composition for each pair of THZs. This procedure is based on the concept of 'area isonymy'. Isonymy refers to "*the recurrence of the same surnames in different ancestral lines in the same pedigree*" (Lasker 1969, 309), and this idea can be transferred to zonal populations. The similarity, η, between the names of individuals resident in any pair (i,j) of zones is calculated as:

$$\eta_{ij} = \sum_s \frac{n_{s.i} n_{s.j}}{2 n_i n_j} \quad (1)$$

where $n_{s.i}$ and $n_{s.j}$ are the number of bearers of surname s in zonal populations i and j respectively. There are other measures of isonymy, such as those developed by Hedrick (1971, 1975) and Nei (1973), who adjust for within-population surname diversity. However, these measures have been shown to be usually correlated with Lasker's measure. Therefore, we use the simpler specification by Lasker, which has been successfully applied in various studies (e.g. Cheshire 2014). The degree of isonymy measured for each THZ can be converted to a distance by taking the negative logarithm.

<< Figure 1 >>

<< Figure 2 >>

Pairwise distances between zones are ideal inputs for clustering algorithms and we use Ward's hierarchical clustering algorithm to identify isonymy groups, i.e. groups of zones with similar surname compositions (Figure 1). As has been previously reported (Cheshire and Longley 2012; Kandt, Cheshire, and Longley 2016), there is a very high level of geographic contiguity in the cluster assignments of long-settled names, and the results of the Ward cluster analysis can be used to derive ‘isonymy regions’ at any convenient level of granularity or recursion, as illustrated for British names from the 1881 Census in Figure 2. The stability of cluster solutions can be further estimated by using computationally intensive bootstrapping techniques (Jain and Moreau 1987), but in view of the near-complete inclusion of the 1881 population, we expect that the cluster solution represents a robust classification of areas. We explore the emergent regionalisation at various levels of the clustering hierarchy and visually identify a parsimonious number of contiguous isonymy regions (*r*) as 77, 39 of which occurred at the lowest level shown in Figure 2C, nesting within the 38 coarser granularities shown in Figure 1. This zonal configuration is used for the next step, in which each of the 88,457 surnames is assigned to an appropriate isonymy region.

Assigning surnames to their ancestral regions

For each surname, we calculate the location quotient for all 77 isonymy regions. The location quotient LQ_s^i measures the prevalence of surname s in area i relative to the overall prevalence of the surname across all zones (Burt, Barber, and Rigby 2009). It is defined as follows:

$$LQ_s^i = \frac{n_s^i}{\sum_s n_s^i}\left(\frac{n_s}{\sum_s n_s}\right)^{-1} \tag{2}$$

where n_s^i is the frequency of surname s in region i, and n_s is the frequency of surname s in England, Scotland, and Wales. We use the location quotient to initially assign each surname to the region showing the highest LQ at the lowest level of the clustering hierarchy. This initial step ensures that each surname is assigned to the region with its highest concentration at the lowest level. If the assigned region encompasses 75 per cent of all bearers of this surname, the region is defined as the surname's geographic origin. If the region encompasses fewer than 75 per cent of all occurrences, then we proceed to the next, higher levels in the clustering hierarchy. We successively repeat this procedure until the threshold of 75 per cent for each surname is reached. As a result, each surname is assigned to a region encompassing the vast majority of its occurrences, and across surnames, these regions vary according to a surname's spatial diffusion, indicating, in turn, the uncertainty in defining geographic origins. Rare and geographically concentrated surnames are thus assigned to smaller regions, while common and geographically dispersed names are assigned only at higher levels in the clustering trees. For instance, Brown is assigned to region *r*02 covering all of Great Britain except Wales (*r*03), while Rees is assigned to southern Wales (*r*29) at the bottom of our truncated clustering tree.

Figure 3 details the cumulative percentages of the number of surnames that are allocated to the different isonymy regions, and the numbers of individuals that these names account for. Overall the number of surnames in the highest tiers is relatively small, but their coverage of the population is high. The opposite holds for the lower tiers (i.e. at lower levels in the truncated clustering tree). This procedure generates the surname taxonomy of regionally identifiable origins.

<< Figure 3 >>

The distribution of surnames across the clustering hierarchy is sensitive to the threshold of 75 per cent. Higher thresholds lead to more surnames being allocated to larger regions and vice versa. We chose the threshold such that the region of a surname encompasses the vast majority of its bearers, while not being too sensitive to possible migration of some surname bearers to places (predominantly cities) outside its 'heartland'.

Identifying local populations

The surname taxonomy allows population profiling in various ways. First, in each THZ, we establish the proportion of the population with local, regional and national surnames. Surnames deemed 'local' are those in a THZ that fall within one of the 39 most granular zones for which the 75 per cent threshold is passed. We define surnames as 'regional' when they are more widely dispersed at successively higher levels of the clustering tree until the threshold of 'country-wide' surnames is reached – defined by isonymy regions *r*03, *r*04 and *r*05 shown in Figure 2A. Finally, surnames are 'GB-wide' if their cores are confined only to either regions *r*01 (Great Britain) or *r*02 (England and Scotland).

We extend these classes by accounting for names prevalent in neighbouring areas. Some of the isonymy regions are not wholly contiguous and have small outliers from their cores. These outliers are removed and regions are only considered neighbours where their core areas are adjacent to one another. We define all names that belong to immediately adjacent isonymy regions as 'locally adjacent', while names that belong to adjacent higher level aggregations are defined as 'regionally adjacent'. These classes of names are used to eliminate artificial boundary effects that arise because we define surname origin regions using only 75 per cent of occurrences and some parts of the remaining 25 per cent frequently occur in adjoining areas. Thus, for example, a surname that is identified with the Cornish isonymy region (*r*30) is likely to also occur in adjacent Devon (*r*71).

<< Table 2 >>

4. Measuring intergenerational population change

Trends and patterns of population change

We use the 1881 regional taxonomy of surnames as a baseline against which to measure population changes as recorded in the 1998 and 2016 Consumer Registers. In order to measure population change, we adopt the least restrictive definition of locality, which considers surname's locality at all levels in the clustering tree and in adjacent regions. (The Consumer Registers differ from the 1881 Census in their inclusion only of individuals aged 19 or older, and omission of individuals who do not consent to inclusion.) In the discussion that follows, we consider only the names identified as British as set out in Table 1.

The share of the population that resides in their surname's 1881 isonymy region decreases in all parts of Great Britain (Figure 4). Viewed across all 77 isonymy regions the average share dropped from 95.6 per cent to 88.3 per cent in 1998 and further to 79.0 per cent in 2016. We can interpret from this plausible trend that, over time, an increasing number of people leave their region of origin to settle elsewhere. This process results in an increased mixture of populations and a consequent relative reduction of local populations in a given area. Yet, the overall decline masks significant regional variations. In the region roughly encompassing Greater London (*r*62), the proportion dropped sharply from 91.7 per cent in 1881 to 55.9 per cent in 2016. Accordingly, the higher-level isonymy regions encompassing London in the clustering tree (*r*16, *r*22, *r*36) also experience significantly stronger declines than the rest of the country. These regions were also among the regions with the lowest share of local population in 1881. The isonymy regions that largely comprise of Cumberland (*r*72), the

Scottish Lowlands (*r*68), Cheshire (*r*48) and the North East and Cumbria (*r*50) have the highest proportions of local population in 2016 (Table 3). In addition, the distribution as a whole has become less centred around the mean, indicating diverging trends among regions in recent years.

<< Figure 4 >>

<< Table 3 >>

The temporal trends exhibit a clear geographical pattern beyond the scale of the isonymy regions (Figure 5). Measured at the level of THZs, the share of the local population declined most dramatically in the cores of larger metropolitan areas, such as London, Birmingham, Leeds and Manchester over the inter-generational period 1881 to 1998. This pattern is similarly manifest over the more recent period of 1998 to 2016, in which the areas recording high population change extended to the metropolitan suburbs of London, as well as smaller cities and other urban areas such as Leicester and Newcastle upon Tyne. In contrast, rural areas in northern England, Scotland and Wales remain largely unchanged over both time periods. There is also evidence of regional difference, with southern metropolitan centres experiencing the more pronounced changes.

<< Figure 5 >>

Characterisation of ancestral population change

Knowledge of surnames through their regional origins also makes it possible to characterise the population movements that underpin demographic composition at migration destinations. We do this by estimating an inter-generational origin-destination matrix of population-weighted flows of surnames over the period of 1881 to 1998, alongside a similar matrix for the more recent 1998 to 2016 period. For each time period, we define the origins of the matrix as the entire set of the 77 nested isonymy regions and regional destinations as the 12 more aggregate isonymy regions shown in Figure 2B. The choice of the level of *k*=12 isonymy regions is arbitrary and any level could be chosen depending on the level of granularity that is of interest. We choose 12 regions because they are sufficiently granular to meaningfully chart regional migration flows while avoiding very large numbers of nuanced results. For both time periods, we proceed by associating the number of British surname bearers identified as non-local to each of these 12 regions. In the following, we have assigned names to each of these 12 regions to indicate their approximate location, however, they may not be exactly contiguous with the administrative extent that they imply.

For each of the 12 regions, we calculate the difference in population sizes broken down by surname class between 1998 and 1881. We adjust the absolute increase by population growth that has occurred during this period and determine the excess population over the expected population. For instance, the total non-local population residing in *r*22 increased by 238,223 from 78,709 to 316,932 individuals between 1881 and 1998 (Table 4). The proportion of the non-local population originating in *r*44 was 16.7 per cent in 1881 (13,136/78,709). C*eteris paribus,* we expect this population to be 16.7 per cent of the total increase in the non-local population in 1998. Thus, the expected non-local population increase with surnames originating in *r*44 is 39,578 in region *r*22. Compared to the actual increase of 44,454, we infer a net flow of 4,696. We repeat this process for all populations in each of the 12 regions between 1881-1998 and 1998-2016. A positive net result is taken as an indicator of net migration flow, regardless that only the 1881 figures include children, and assuming that the potentially confounding effects of patronymic name retention cancel each other out and that population fertility is geographically invariant for migrants and non-migrants alike. Although based on our assumptions, it is inconceivable that the estimated flow represents that actual population movement, it can nevertheless be understood as an index of relative magnitude of migratory movements among different, regional populations.

<< Table 4 >>

The inferred inter- and intragenerational flows reveal selective population movements with diverse geographical patterns. Exploring the top 10 net flows between regions (Tables 5a, 5b), we find, for example, that by 1998, the surname region encompassing Wales (*r03*) had received a large inflow of British surname bearers from southern parts of England (*r*16). South East England [excl. Sussex and Kent] (*r*22), on the other hand, predominantly received British surname bearers from within the region (*r*62) and from southern and central parts of Scotland [excl. Aberdeenshire] (*r*44). The western parts of England also received high numbers of people with 'southern Scottish' names (e.g. 'Northamptonshire' (*r*17), 'Greater Yorkshire and the Humberside' (*r*18) and 'East Anglia' (*r*13)).

<< Table 5a >>

<< Table 5b >>

A comparison of all flows across the *k*=12 destination regions (Table 6) reveals that the size of the migration varies considerably. In both time periods, South East England [excluding Sussex and Kent] (*r*22), but including London, is by far the most popular destination region in terms of both absolute and relative flows, when adjusted for time period and population size. The Northern Highlands (*r*20) in northern Scotland, on the other hand, received overall the lowest influx of bearers of non-local surnames, although the influx appears to have increased more recently. In all other regions, flows per year per 100,000 people appear to have decreased in the latter period 1998-2016. Although this trend may indicate that the intergenerational period 1881-1998 may have witnessed times of unusually strong migration movements, the significantly larger population denominator of 1998 may skew these relative estimates.

<< Table 6 >>

Selected population movements are associated with pronounced compositional differences among the *k*=12 destination regions (Figure 6). For example, the non-local population in regions in the northern parts of Scotland (*r*20) has surnames from only three isonymy regions (Figure 6A), whereas in the southern parts of Scotland (*r*21) there is a more diverse mix of non-local surnames. These results may be slightly underestimated because Scotland only breaks into four isonymy regions, each of which are defined as 'local' according to the adjacency criterion defined above. Therefore, an increase in Northern Scottish names in Southern Scotland and vice versa are not considered to be newly imported. The same applies to Wales (*r*03). Nevertheless, Wales much like other regions such as the South East of England reveal a diverse, regional composition of the newly settled population.

Across the two time periods, there are distinct shifts in the origins of newly arrived populations in almost all regions. The Northern Highlands region (*r*20) received most of its new population from a confined sample of English regions (Figure 6A), although total absolute change is very small compared to other regions, as indicated by the bar chart on the left. The regional origin of movers into Southern Scotland (*r*21, Figure 6B) indicates greater diversity of origins within England, alongside a small proportion from Wales. In Northern England (*r*11, Figure 6C), we detect a higher share of Scottish in-movers, which increases in regions further south, such as Greater Yorkshire and the Humberside (*r*18, Figure 6D). Cornwall (*r*09, Figure 6L) particularly stands out in this respect: whereas in the intergenerational period, surnames of newcomers suggested a diverse mix of regional origins, this composition shifted exclusively to Scottish regions in the 1998-2016 period. All of these trends suggest a strong southward movement from Scotland over the last decade. The largest flow can be observed for South East England (*r*22, Figure 6K), which was mainly composed of people of diverse English origins between 1881 and 1998 and subsequently shifted towards people of Scottish origins between 1998 and 2016.

<< Figure 6 >>

The patterns and trends suggest selective migration flows causing regionally differentiated patterns of diversification. In other words, destination regions do not receive the same mix of internal movers in terms of their geographic origin. Neither do we observe a clear relationship between proximity and strength of migration flow, as shown by the influx of people with Scottish names into England's most southern regions. These tendencies are likely to increase diversification both within and between regions.

5. Discussion and Conclusion

Characterising populations through surname geographies

Surname geographies have significant potential to infer ancestral, regional origins of populations. Our historical, isonymy-based regionalisation of Great Britain provides a baseline against which subsequent diffusion, both hierarchical and contagious, can be measured to indicate migration and associated socio-spatial processes. The local and regional origins of most surnames underpin the clear definition of isonymy regions in 1881, albeit that some common surnames, principally metonyms derived from common occupations, such as Smith or Baker, are almost ubiquitous. In 1881, two in three individuals bore surnames that were of subnational extent; that is their surname could be assigned to an isonymy region at the local, regional or country-wide level. As a consequence, the taxonomy adds significant information over existing national classifications of surnames and can be used as to provide place-based indicators to profile populations with regard to the areal extent of 'localities' at predefined thresholds.

Regionalisation based on isonymy is thus relevant to the study of the definition and evolution of place. Our analysis reveals a steady decrease of the share of population with local surnames in all parts of Great Britain, although the vast majority of people continue to reside in regions defined by the geography of their surnames. The overall decline in the share of population that bears local surnames masks considerable regional and local variation. The strongest decline occurs in English cities, whereas change in rural areas and certain parts of northern England, in particular, is much less marked. The geographical diffusion of surnames also suggests that there has been significant migratory movement from southern Scottish regions to the southern parts of England. During the more recent intra-generational period, London and other English metropolitan areas continue to experience the highest levels of change. In sum, our findings chart a trend of increasing regional diversity and differentiated compositional change of historic population structures.

In future research, we hope to consider whether faster declines in the shares of long-established populations of local surname bearers are associated with other social and demographic characteristics, such as higher levels of neighbourhood population churn, different community structures or distinctive patterns of economic activity. However, there are important caveats to such analysis. Return migration within or between generations is undetectable and is likely to be geographically variable in its effects upon the creation and maintenance of place effects. Changing surname geographies also manifest factors other than migration, such as the propensity for women to change surnames upon marriage, and different propensities for men and women to undertake regional, national, and international migration. Nevertheless, evidence suggests that local marriages are still common; therefore, the regional origin of newly adopted surnames of many women remain highly correlated with the origin of their maiden names (Mazières and Roth 2018). We also believe that this work enables a new and more disaggregate approach to the analysis of migrant origin – destination flows. This in turn enables richer ‘place profiles’ of population mix, and analysis of the inter-generational outcomes of migration of family groups.

Limitations of our case study

The gold standard of characterising local populations by their origins would be to directly trace the individuals that moved as well as their offspring. While this may be partially accomplished by linking historical census micro-data, this would strongly rely on inter-census linkage which, in turn, would also be based on surnames alongside other identifying characteristics. Moreover, since data protection regulations require a 100-year time lag on the release of full census micro-data, direct, population-wide tracking of individual movement remains unfeasible over inter-generational time periods. Therefore looser coupling of historical censuses with comparatively recent Consumer Registers to trace geographic shifts in the patterning of surnames presents an alternative way to study settlement histories over extended time frames.

Because the time points in this study remain unequally spaced with missing data for a large part of the Twentieth Century, changes during intervening periods remain undetected. Our own agenda is to assimilate further historical censuses up until 1911, but in view of access procedures and issues of data quality, we expect that considerable time and effort will be required to hone these sources to this research. We do, however, believe that this is attainable, particularly if techniques used to link recent annual Consumer Registers can be adapted to historical census sources (see Lansley, Li, and Longley 2018).

The results of inferring migratory flow over the inter-generational period may also be affected by differential fertility rates within Great Britain at different times. Although it can be assumed that fertility rates differ geographically, there is very little data on regional variation of this characteristic. If such figures were available, the index could be adjusted accordingly, although potentially differential fertility rates between domestic migrants and the local population or variations in time may render such adjustments speculative.

A further limitation is that the baseline Welsh and Scottish populations are less diverse in terms of surnames: that is, there are fewer surnames per thousand population than in the rest of Great Britain. The lower level of surname diversity may be reflected in the clustering solution, in which Wales only breaks into two sub-regions and Scotland into four. These outcomes might indicate that observed changes in these regions may either be underestimated, for instance, because of undetected within-zone migration of bearers of regionally common names; or overestimated, for instance, by a relatively small influx of bearers of non-local names that manifest in disproportionate changes in the stock of a region's surnames.

It should also be remembered that the flow estimates of regional populations indicate relative magnitudes; they should not be taken as precise indicators of absolute change. Although there are more sophisticated approaches to modelling population flows based on surname geographies (Degioanni and Darlu 2001; Bloothooft and Darlu 2013), we have chosen a simpler, descriptive method, since our study focuses on long-term population change, which does not permit direct attribution of population change to individual movements. More methodological work to model intergenerational movements based on surname geographies, possibly within a spatial interaction framework, is needed.

Finally, uncertainties result from the commercial nature of the Consumer Registers. The sources of data and ways in which they are assembled are not fully disclosed and there may be unknown unevenness in coverage – albeit that there are no obvious hypotheses to suggest that some British surnames will be less well covered than others. Our own research attempts to triangulate Consumer Registers with conventional statistical sources such as the 2001 and 2011 Censuses, and the Office for National Statistics Mid-Year Population Estimates (Lansley, Li, and Longley 2019), but it is clear that the representation of the adult British population remains incomplete. Notwithstanding these limitations, Consumer Registers remain the best available data source of individually georeferenced name records.

Concluding remarks

Our study demonstrates that geographical surname taxonomies can be used to infer the composition of regional and local populations, present day and historic. We have developed a hierarchical, explicitly spatial framework to infer regional origins to varying levels of precision for every surname. This makes it possible to characterise population change at multiple scales across Great Britain over inter- as well as intra-generational time horizons. Analysis of changing surname geographies allows us to characterise local populations in terms of their baseline characteristics and subsequent population inflows arising from contagious diffusion from adjacent locations or hierarchical diffusion cascading through the British settlement hierarchy. Regional surname taxonomies can be used as indicators of the uniqueness of regions, since names manifest the unique ancestral, socio-cultural and biogenetic contexts of their bearers (see, for example, Winney et al. 2012), and thus offer value in exploring the long-term context of prominent societal phenomena, such as persistent inequalities in health, social mobility or labour market outcomes.

References

Barrai, I., A. Rodriguez-Larralde, E. Mamolini, F. Manni, and C. Scapoli. 2001. Isonymy structure of USA population. *American Journal of Physical Anthropology* 114 (2):109–123. doi: 10.1002/1096-8644(200102)114:2%3C109::AID-AJPA1011%3E3.0.CO;2-I.

Bloothooft, G., and P. Darlu. 2013. Evaluation of the Bayesian method to derive migration patterns from changes in surname distributions over time. *Human Biology* 85 (4):553–568. doi: 10.3378/027.085.0403.

Burt, J., G. Barber, and D. Rigby. 2009. *Elementary statistics for geographers* 3rd ed. London: The Guilford Press.

Cheshire, J. 2014. Analysing surnames as geographic data. *Journal of Anthropological Sciences* 92:99–117. doi: 10.4436/JASS.92004.

Cheshire, J., and P. A. Longley. 2012. Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science* 26 (2):309–325. doi: 10.1080/13658816.2011.591291.

Cheshire, J., P. A. Longley, K. Yano, and T. Nakaya. 2014. Japanese surname regions. *Papers in Regional Science* 93 (3):539–555. doi: 10.1111/pirs.12002.

Clark, G., and N. Cummins. 2015. Intergenerational wealth mobility in England, 1858-2012: Surnames and social mobility. *The Economic Journal* 125 (582):61–85. doi: 10.1111/ecoj.12165.

Darlu, P., G. Bloothooft, A. Boattini, L. Brouwer, M. Brouwer, G. Brunet, P. Chareille, J. Cheshire, R. Coates, K. Dräger, B. Desjardin, P. Hanks, P. A. Longley, K. Mandemakers, P. Mateos, D. Pettener, A. Useli, and F. Manni. 2012. The family name as socio-cultural feature and genetic metaphor: From concepts to methods. *Human Biology* 84 (2):169–214. doi: 10.3378/027.084.0205.

Degioanni, A., and P. Darlu. 2001. A Bayesian approach to infer geographical origins of migrants through surnames. *Annals of Human Biology* 28 (5):537–545. doi: 10.1080/03014460110034265.

Degioanni, A., P. Darlu, and C. Raffoux. 2003. Analysis of the French National Registry of unrelated bone marrow donors, using surnames as a tool for improving geographical localisation of HLA haplotypes. *European Journal Of Human Genetics* 11:794. doi: 10.0.4.14/sj.ejhg.5201038.

Hedrick, P. W. 1971. A new approach to genetic similarity. *Evolution* 25 (2):276–280. doi: 10.1111/j.1558-5646.1971.tb01880.x.

———. 1975. Genetic similarity and distance: Comments and comparisons. *Evolution* 29 (2):362–366. doi: 10.2307/2407223.

Herrera Paz, E. F., C. Scapoli, E. Mamolini, M. Sandri, A. Carrieri, A. Rodríguez-Larralde, and I. Barrai. 2014. Surnames in Honduras: A study of the population of Honduras through isonymy. *Annals of Human Genetics* 78 (3):165–177. doi: 10.1111/ahg.12057.

Higgs, E., and K. Schürer. 2014. Integrated Census Microdata (I-CeM), 1851-1911 [data collection] UK Data Service SN: 4178. doi: 10.5255/UKDA-SN-7481-1.

Jain, A. K., and J. V. Moreau. 1987. Bootstrap technique in cluster analysis. *Pattern Recognition* 20 (5):547–568. doi: 10.1016/0031-3203(87)90081-1.

Jobling, M. A. 2001. In the name of the father: Surnames and genetics. *Trends in Genetics* 17 (6):353–357. doi: 10.1016/S0168-9525(01)02284-3.

Kandt, J., J. A. Cheshire, and P. A. Longley. 2016. Regional surnames and genetic structure in Great Britain. *Transactions of the Institute of British Geographers* 41 (4):554–569. doi: 10.1111/tran.12131.

Kandt, J., and P. A. Longley. 2018. Ethnicity estimation using family naming practices. *PLOS ONE* 13 (8):e0201774. doi: 10.1371/journal.pone.0201774.

King, T. E., and M. A. Jobling. 2009. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends in Genetics* 25 (8):351–360. doi: 10.1016/j.tig.2009.06.003.

Lan, T., J. Kandt, and P. A. Longley. 2018. Ethnicity and residential segregation. In *Consumer Data Research*, eds. P. A. Longley, A. Singleton, and J. A. Cheshire, 71–83. London: UCL Press.

Lansley, G., W. Li, and P. A. Longley. 2018. Consumer registers as spatial data infrastructure and their use in migration and residential mobility research. In *Consumer Data Research*, eds. P. A. Longley, J. Cheshire, and A. Singleton, 15–28. London: UCL Press.

———. 2019. Creating a linked consumer register for granular demographic analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. doi: 10.1111/rssa.12476.

Lasker, G. W. 1969. Isonomy (recurrence of the same surnames in affinal relatives): A comparison of rates calculated from pedigrees, grave markers and death and birth registers. *Human Biology* 41 (3):304–321.

Lewis, D. J., and P. A. Longley. 2012. Patterns of patient registration with primary health care in the UK National Health Service. *Annals of the Association of American Geographers* 102 (5):1135–1145. doi: 10.1080/00045608.2012.657500.

Longley, P. A., J. A. Cheshire, and P. Mateos. 2011. Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum* 42 (4):506–516. doi: 10.1016/j.geoforum.2011.02.001.

Longley, P. A., J. Cheshire, and A. Singleton eds. 2018. *Consumer Data Research*. London: UCL Press.

Longley, P. A., R. Webber, and D. Lloyd. 2007. The quantitative analysis of family names: Historic migration and the present day neighborhood structure of Middlesbrough, United Kingdom. *Annals of the Association of American Geographers* 97 (1):31–48. doi: 10.1111/j.1467-8306.2007.00522.x.

Manni, F., B. Toupance, A. Sabbagh, and E. Heyer. 2005. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Ychromosome sampling. *American Journal of Physical Anthropology* 126 (2):214–228. doi: 10.1002/ajpa.10429.

Mateos, P. 2007. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place* 13 (4):243–263. doi: 10.1002/psp.457.

Mateos, P., P. A. Longley, and D. O'Sullivan. 2011. Ethnicity and population structure in personal naming networks ed. D. O'Rourke. *PLoS ONE* 6 (9):e22943. doi: 10.1371/journal.pone.0022943.

Mazières, A., and C. Roth. 2018. Large-scale diversity estimation through surname origin inference. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 139 (1):59–73. doi: 10.1177/0759106318778828.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* 70 (12):3321–3323.

Orrù, A., S. De Iasio, P. Frederic, M. Girotti, R. Boano, and E. Sanna. 2018. Spatial diffusion of surnames by long transhumance routes between highland and lowland: A study in Sardinia. *Homo* 69 (3):127–138. doi: 10.1016/j.jchb.2018.06.006.

Petersen, J., P. Longley, M. Gibin, P. Mateos, and P. Atkinson. 2011. Names-based classification of accident and emergency department users. *Health & Place* 17 (5):1162–1169. doi: 10.1016/j.healthplace.2010.09.010.

Rodríguez-Larralde, A., J. Morales, and I. Barrai. 2000. Surname frequency and the isonymy structure of Venezuela. *American Journal of Human Biology* 12 (3):352–362. doi: 10.1002/(SICI)1520-6300(200005/06)12:3%3C352::AID-AJHB5%3E3.0.CO;2-S.

Rodríguez-Larralde, A., C. Scapoli, M. Beretta, C. Nesti, E. Mamolini, and I. Barrai. 1998. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. *Annals of Human Biology* 25 (6):533–540. doi: 10.1080/03014469800006772.

Schürer, K., and M. Woollard. 2000a. 1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced version). [data collection]. UK Data Service SN: 4177. doi: 10.5255/UKDA-SN-4177-1.

———. 2000b. 1881 Census for Scotland. [data collection]. UK Data Service SN: 4178. doi: 10.5255/UKDA-SN-4178-1.

Shi, Y., L. Li, Y. Wang, J. Chen, Y. Yuan, and H. E. Stanley. 2019. Regional surname affinity: A spatial network approach. *American Journal of Physical Anthropology* 168 (3):428–437. doi: 10.1002/ajpa.23755.

Southall, H. 2012. Rebuilding the Great Britain Historical GIS, Part 2: A geo-spatial ontology of administrative units. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 45 (3):119–134. doi: 10.1080/01615440.2012.664101.

———. 2014. Rebuilding the Great Britain Historical GIS, Part 3: Integrating ualitative content for a sense of place. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 47 (1):31–44. doi: 10.1080/01615440.2013.847774.

Winder, R. 2014. *Bloody foreigners: The story of migration to Britain*. London: Little, Brown.

Winney, B., A. Boumertit, T. Day, D. Davison, C. Echeta, I. Evseeva, K. Hutnik, S. Leslie, K. Nicodemus, E. C. Royrvik, S. Tonks, X. Yang, J. Cheshire, P. Longley, P. Mateos, A. Groom, C. Relton, D. T. Bishop, K. Black, E. Northwood, L. Parkinson, T. M. Frayling, A. Steele, J. R. Sampson, T. King, R. Dixon, D. Middleton, B. Jennings, R. Bowden, P. Donnelly, and W. Bodmer. 2012. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics* 20 (2):203–210. doi: 10.1038/ejhg.2011.127.

Yasuda, N., L. L. Cavalli-Sforza, M. Skolnick, and A. Moroni. 1974. The evolution of surnames: An analysis of their distribution and extinction. *Theoretical Population Biology* 5 (1):123–142. doi: 10.1016/0040-5809(74)90054-9.

Table 1: Long-standing British surnames in 1881, 1998, and 2016.

	1881	**1998**	**2016**
Data	Population Census	Consumer Register	Consumer Register
Source	Censuses of England, Scotland, and Wales	Public Electoral Register	Public Electoral Register, Consumer Transactions
Extent	Entire population	Adult Population	Adult Population
Number of surnames	420,962	710,155	1,386,583
of which classified as British	88,457	75,850	76,089
Number of individuals	29,740,607	45,404,725	50,163,224
who bear a surname classified as British	28,577,866	41,083,610	40,060,165

Table 2: Populations defined as local or otherwise in 1881.

	Excluding adjacency			Including adjacency		
	Individuals	**Proportion**	**Cumulative**	**Individuals**	**Proportion**	**Cumulative**
Local	514,401	1.8	1.8	542,979	1.9	1.9
Regional	7,515,979	26.3	28.1	8,744,827	30.6	32.5
Country-wide	8,916,294	31.2	59.3	9,859,364	34.5	67.0
GB-wide	7,801,757	27.3	86.6	7,973,224	27.9	94.9

Table 3: The four isonymy regions with the highest and lowest percentage of individuals in 2016 that bear a surname that is local to the region in which they are residing.

r	**Area**	**1881**	**1998**	**2016**
01	Great Britain	94.9	86.9	75.5
	high			
72	Cumberland	95.6	92.5	86.7
68	Scottish Lowlands	94.1	90.3	86.6
48	Cheshire	95.6	92.6	86.5
50	North East and Cumbria	94.8	92.1	85.9
	low			
16	South East England	93.2	82.3	68.1
22	South East England [excl. Sussex and Kent]	92.9	80.4	65.2
36	Greater London, Surrey, and Hampshire	92.4	78.4	62.1
62	Greater London	91.7	73.8	55.9

Table 4: Example calculation of selected net intergenerational flows 1881-1998 for *r*22. The expected population is defined as the percentage of 1881 non-local population times the population growth of the non-local population in the entire region *r*22.

Origin	Population 1881	Population 1998	Expected Population Change 1881 - 1998	Actual Population Change 1881 1998	Difference (Net Flow)
*r*11	…	…	…		…
*r*44	13,136	57,590	39,758	44,454	4,696
*r*62	278	7,114	841	6,836	5,995
*r*73	2,474	13,721	7,488	11,247	3,759
*r*76	…	…	…		…
Total	78,709	316,932			

Table 5a: The 10 largest flows estimated for 1881-1998

		1881 - 1998			
r	**From**	***r***	**To**	**Net flow**	**Per Year**
16	South East England	03	Wales	6,792	58
62	Greater London	22	South East England [excl. Sussex and Kent]	5,995	51
44	Southern Scotland [excl. Aberdeenshire]	22	South East England [excl. Sussex and Kent]	4,696	40
44	Southern Scotland [excl. Aberdeenshire]	18	Greater Yorkshire and the Humberside	4,567	39
44	Southern Scotland [excl. Aberdeenshire]	13	East Anglia	4,464	38
62	Greater London	18	Greater Yorkshire and the Humberside	4,385	38
62	Greater London	03	Wales	4,358	37
22	South East England [excl. Sussex and Kent]	03	Wales	4,314	37
62	Greater London	13	East Anglia	4,234	36
44	Southern Scotland [excl. Aberdeenshire]	17	Northamptonshire	4,081	35

Table 5b: The 10 largest flows estimated for 1998-2016

1998-2016					
r	**From**	***r***	**To**	**Net flow**	**Per year**
29	Southern Wales	22	South East England [excl. Sussex and Kent]	1,030	57
44	Southern Scotland [excl. Aberdeenshire]	22	South East England [excl. Sussex and Kent]	870	48
44	Southern Scotland [excl. Aberdeenshire]	18	Greater Yorkshire and the Humberside	856	48
21	Southern Scotland	17	Northamptonshire	787	44
21	Southern Scotland	18	Greater Yorkshire and the Humberside	658	37
21	Southern Scotland	22	South East England [excl. Sussex and Kent]	657	36
44	Southern Scotland [excl. Aberdeenshire]	23	Sussex and Kent	636	35
44	Southern Scotland [excl. Aberdeenshire]	17	Northamptonshire	585	32
44	Southern Scotland [excl. Aberdeenshire]	09	Cornwall	580	32
62	Greater London	13	East Anglia	552	31

Table 6: Inter- and intragenerational flows of the population with surnames that are non-local to the destination region; annual flow per 100,000 people in between parentheses.

		1881-1998		1998-2016	
r	**Destination Area**	**Total flow**	**Per year**	**Total flow**	**Per year**
20	Northern Highlands	1,638	14 (6.0)	612	34 (14.6)
21	Southern Scotland	24,570	210 (5.3)	1,602	89 (2.3)
11	Northern England	7,020	60 (2.7)	972	54 (2.4)
18	Greater Yorkshire and the Humberside	23,985	205 (4.0)	2,502	139 (2.7)
15	Lancashire	11,115	95 (3.7)	1,476	82 (3.2)
19	West Midlands	11,466	98 (2.7)	1,080	60 (1.7)
03	Wales	30,186	258 (7.3)	1,710	95 (2.7)
17	Northamptonshire	20,241	173 (3.0)	1,890	105 (1.8)
13	East Anglia	22,581	193 (5.4)	1,620	90 (2.5)
23	Sussex and Kent	10,998	94 (3.5)	1,080	60 (2.2)
22	South East England [excl. Sussex and Kent]	48,087	411 (3.9)	3,582	199 (1.9)
09	Cornwall	11,232	96 (6.3)	900	50 (3.3)
	Total	221,481	1,907 (4.2)	19,026	1,057 (2.1)

Figure 1: Truncated clustering dendrogram showing the hierarchy of the first 77 isonymy regions identified in the 1881 Census.

Figure 2: Isonymy regions at different levels of the clustering hierarchy as identified by the Ward's clustering algorithm shown in Figure 1.

Figure 3: Cumulative percentage of surnames and individuals in 1881 assigned to the 77 isonymy regions. Isonymy region numbering is as shown in Figure 1.

Figure 4: Boxplots representing the share of people within all isonymy regions that bear a surname that is local to the region they are living in (1881-2016).

Figure 5: Ratio of relative change of population with local surnames for the THZs between a) 1881 and 1998 and b) 1998 and 2016. The earlier years are each indexed at 1.

Figure 6: Proportion of regional origins of 'imported' surnames by the size of their population; colours reflect broader regions (England [E], Scotland [S], Wales [W]) with darkers shades showing regions of more local extent; the bar charts show the total annual inflows presented in Table 6 across all 12 destination regions to put the relative flows in perspective.

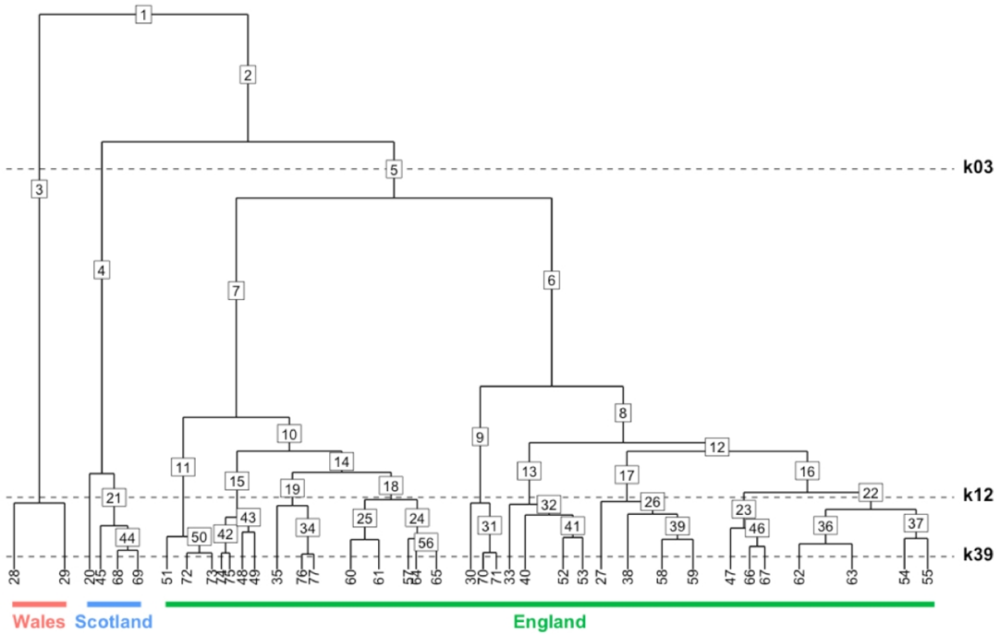

Figure 1: Truncated clustering dendrogram showing the hierarchy of the first 77 isonymy regions identified in the 1881 Census.

432x274mm (72 x 72 DPI)

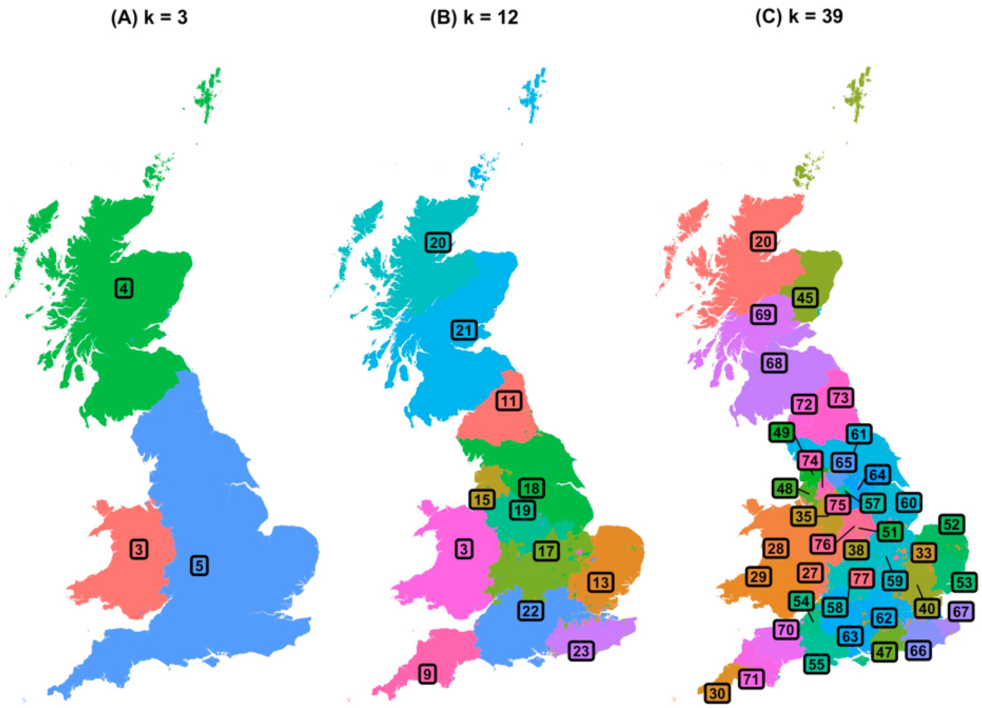

Figure 2: Isonymy regions at different levels of the clustering hierarchy as identified by the Ward's clustering algorithm shown in Figure 1.

334x237mm (72 x 72 DPI)

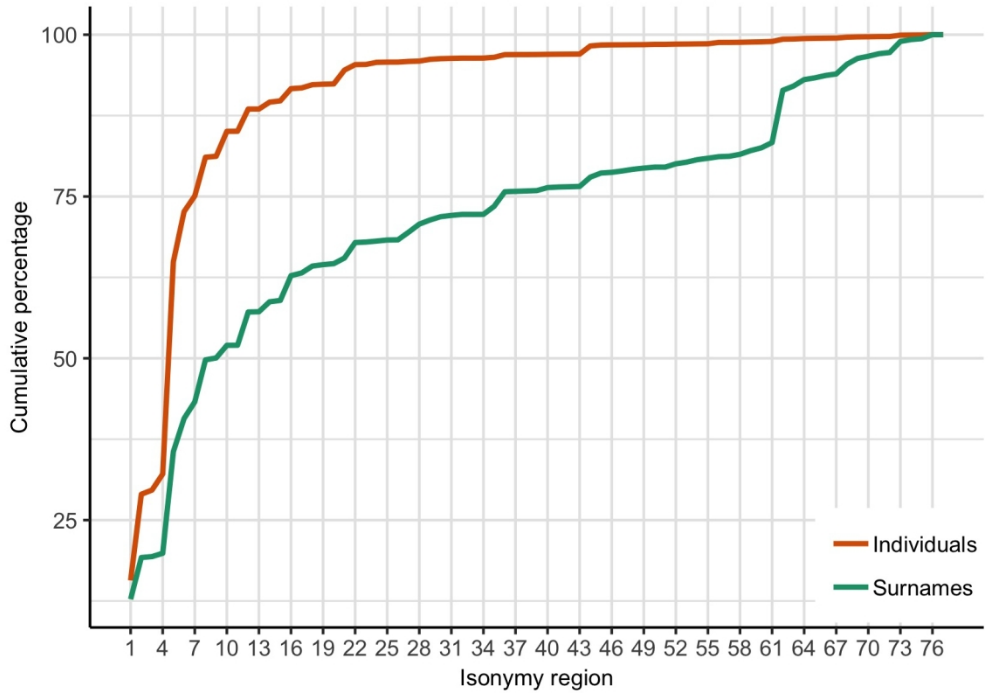

Figure 3: Cumulative percentage of surnames and individuals in 1881 assigned to the 77 isonymy regions. Isonymy region numbering is as shown in Figure 1.

583x416mm (72 x 72 DPI)

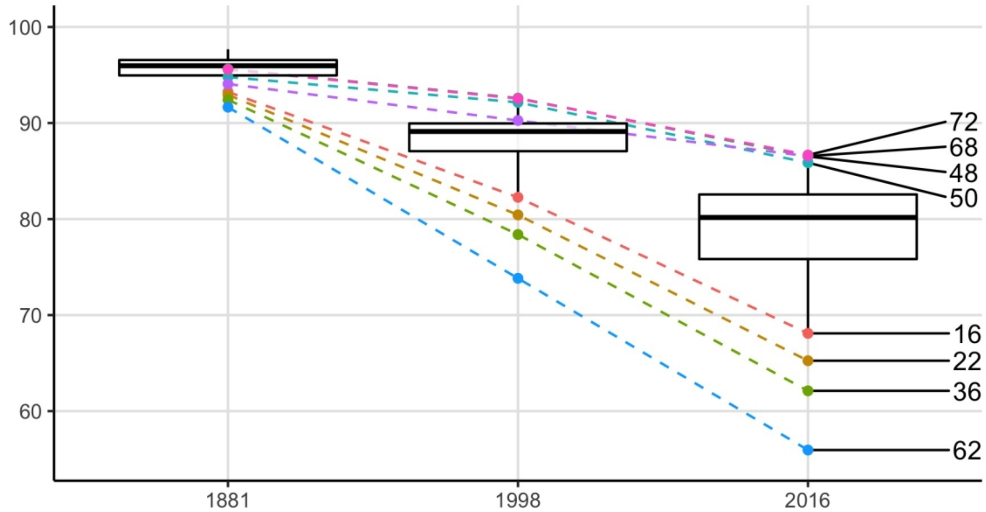

Figure 4: Boxplots representing the share of people within all isonymy regions that bear a surname that is local to the region they are living in (1881-2016).

624x333mm (72 x 72 DPI)

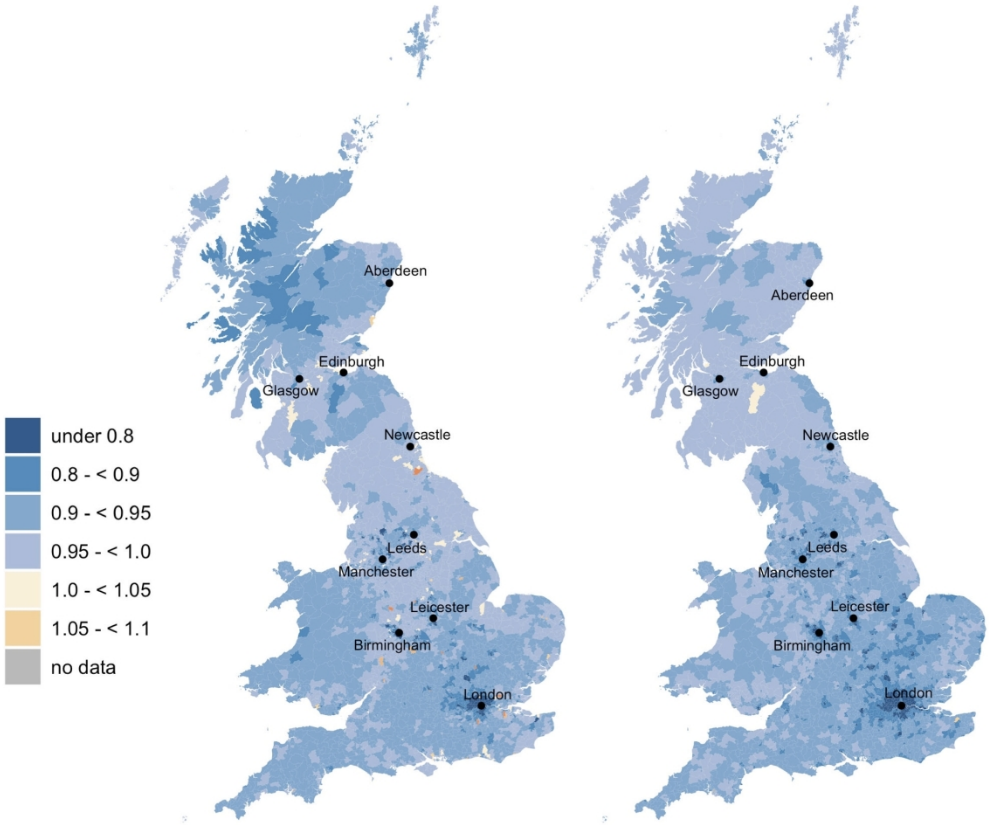

Figure 5: Ratio of relative change of population with local surnames for the THZs between a) 1881 and 1998 and b) 1998 and 2016. The earlier years are each indexed at 1.

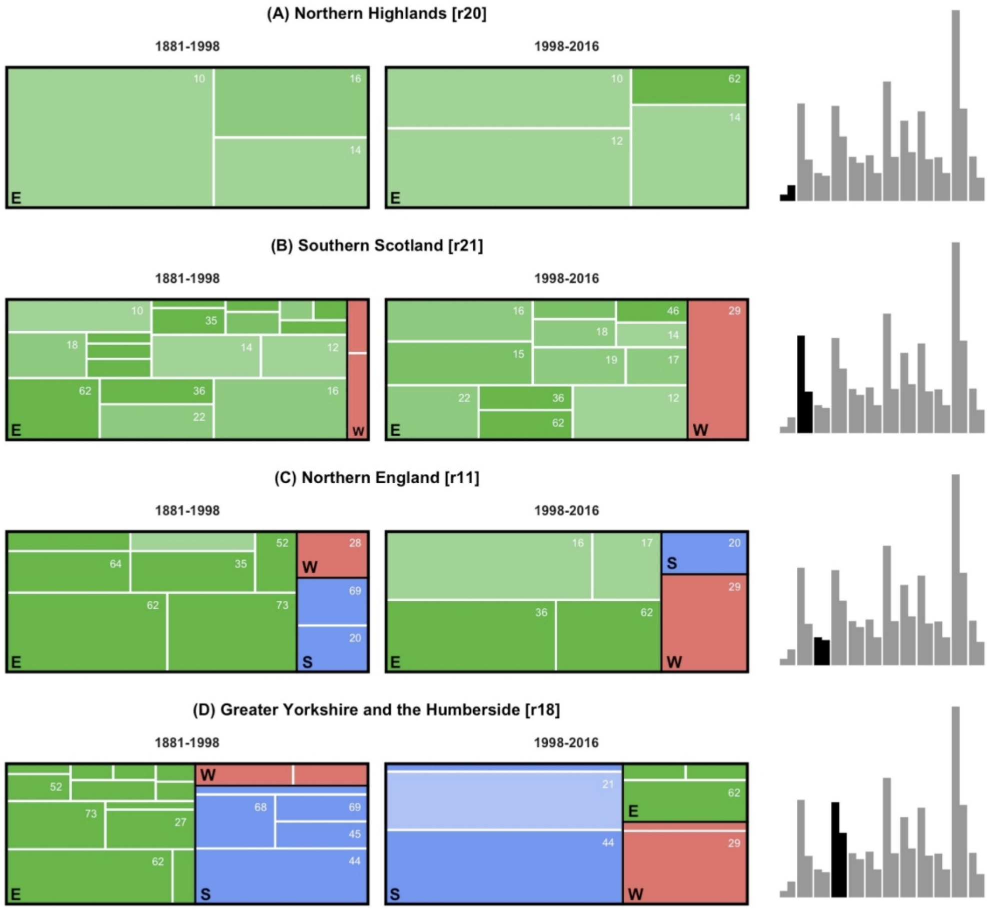

Figure 6: Proportion of regional origins of ‘imported’ surnames by the size of their population; colours reflect broader regions (England [E], Scotland [S], Wales [W]) with darkers shades showing regions of more local extent; the bar charts show the total annual inflows presented in Table 6 across all 12 destination regions to put the relative flows in perspective.

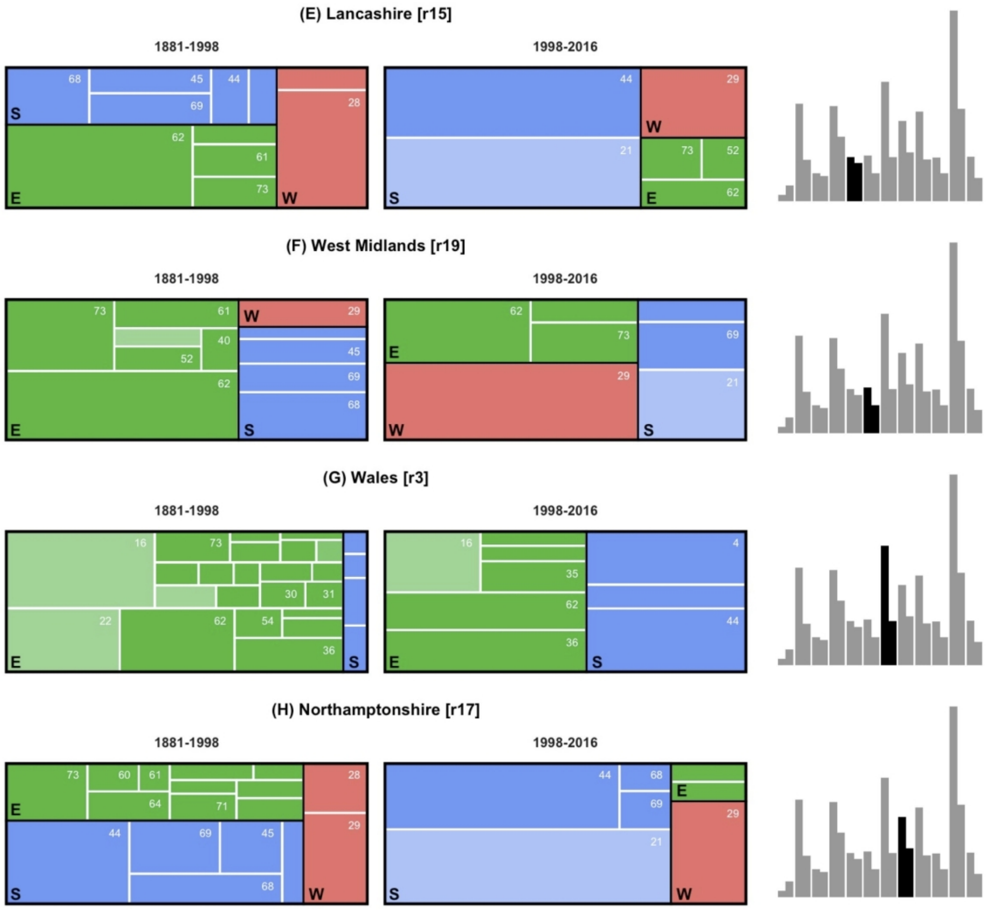

Figure 6: Proportion of regional origins of 'imported' surnames by the size of their population; colours reflect broader regions (England [E], Scotland [S], Wales [W]) with darkers shades showing regions of more local extent; the bar charts show the total annual inflows presented in Table 6 across all 12 destination regions to put the relative flows in perspective.

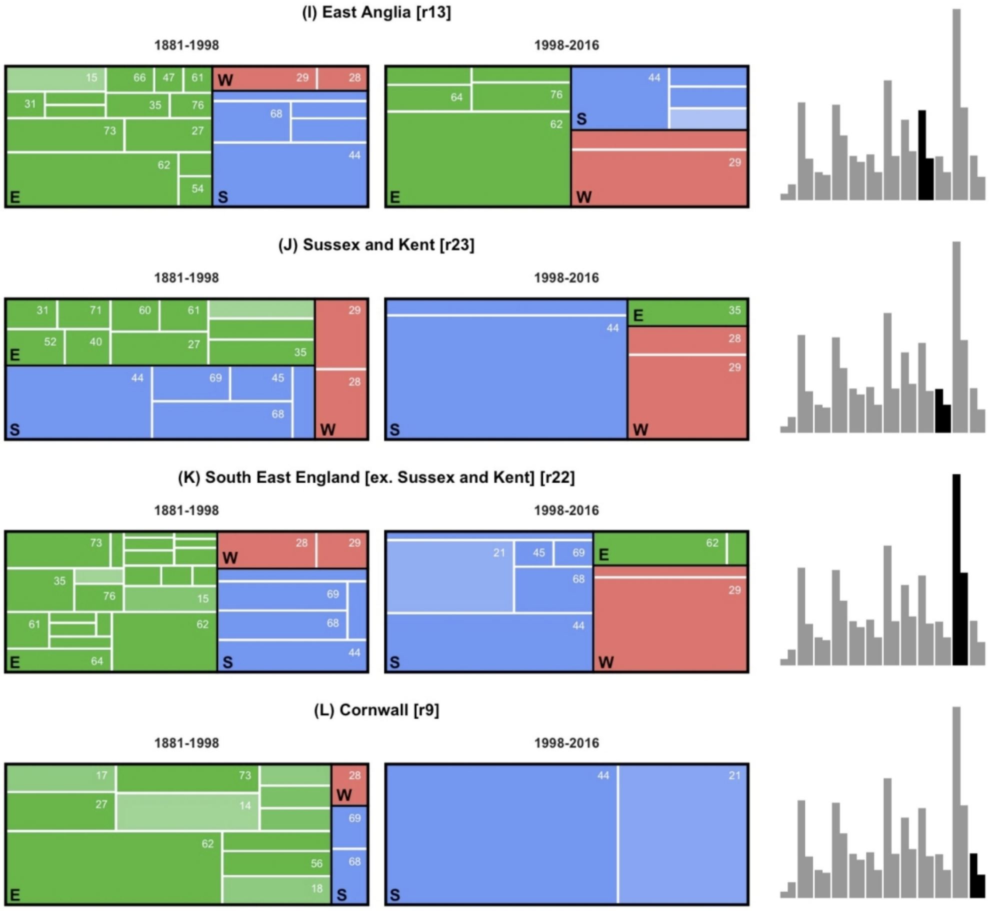

Figure 6: Proportion of regional origins of 'imported' surnames by the size of their population; colours reflect broader regions (England [E], Scotland [S], Wales [W]) with darkers shades showing regions of more local extent; the bar charts show the total annual inflows presented in Table 6 across all 12 destination regions to put the relative flows in perspective.