

Platial Geo-Temporal Demographics Using Family Names

Justin van Dijk^{ORCID} and Paul A Longley^{ORCID}

Department of Geography, University College London, UK

We introduce platial geo-temporal demographics as a novel way to describe places using family names as markers of migration and change at sub-national scales. By identifying the likely origins of 59,218 surnames in Great Britain, we create platial profiles of surname mixes in terms of the distance their forbears have likely migrated between 1881 and 1998/2016. By combining individual-level data derived from historic censuses of population with near-complete contemporary population registers of enfranchised adults, we demonstrate how locally and regionally distinctive surname mixes can be used in characterizing places in terms of demographic change and stasis. The results suggest that a hierarchy of places arises in Great Britain, with larger conurbations (e.g., London and Birmingham) having more surnames that can be traced back to other parts of Great Britain and beyond, as opposed to places that are characterized by the presence of a larger share of surnames that have a more local origin. These regional differences are likely linked to processes of social mobility and economic activity.

Keywords: surnames; platial profiling; geodemographics; population change

History: received on 5 July 2019; accepted on 16 August 2019; published on 27 January 2020

1 Introduction

Geodemographics has been defined as the analysis of people by where they live (Harris et al., 2005). The field relates to platial geographies in that residential neighbourhoods are in the local sense the outcome of ‘birds of a feather flocking together’, in ways that are replicated over entire national (and, arguably, international) settlement systems. Residential structure is thus seen as the interplay of locational proximity and social similarity played out across the urban and regional system. Geodemographics is a geography of night-time residence (Martin et al., 2015), typically measured using a melange of variables sourced from censuses and other conventional statistical or industrial sources. Some data make it possible to shift the focus from place of residence to place of work, resulting in geodemographic classifications of workplaces (e.g., Singleton and Longley, 2019). The re-use of statistics to characterize broader activity patterns than residence alone adds a temporal dimension into classification, and the term ‘geo-temporal demographics’ has been used to recognize that common or shared activity patterns measured over timescales ranging from the diurnal to the inter-generational can be used alongside conventional social, economic, and demographic data in order to better understand community structure. Such classifications may or may not be anchored to geographies of night-time residence. In what follows, we view inter-generational migration patterns as formative in notions of place formation.

Here we will use aggregated geographies of individuals’ family names to create geo-temporal demographic platial profiles. Surnames in the British Isles came into common parlance between the 12th and 14th Centuries and have subsequently been passed down male bloodlines. Most Anglo-Saxon

J van Dijk and PA Longley (2020): *Platial Geo-Temporal Demographics Using Family Names*. In: FB Mocnik and R Westerholt (eds.), *Proceedings of the 2nd International Symposium on Platial Information Science (PLATIAL'19)*, pp. 23–31

<https://doi.org/10.5281/zenodo.3628863>



Second International Symposium on Platial Information Science (PLATIAL'19)
Coventry, UK; 5–6 September 2019

Copyright © by the author(s). Licensed under Creative Commons Attribution 4.0 License.

family names can be described as either metonyms (pertaining to occupation, e.g., Smith and Weaver), toponyms (e.g., Hill and Gill) or diminutives (e.g., Williamson and Williams), and most have clearly identifiable geographic origins, albeit to varying levels of geographic precision (Cheshire and Longley, 2012; Longley et al., 2007). Examination of the evolving geographies of multiple surnames makes it possible to chart the outcomes of historic processes of population change and intergenerational migration (Kandt et al., 2020). From a more recent perspective, forename–surname pairings can be used to infer issues of ethnicity (Kandt and Longley, 2018) and residential segregation (Lan et al., 2018) of bearers of names that have more recently been imported from abroad. However, these foundations have not as yet been used to examine the mixes of surnames that characterize the residents of distinctive places and the cumulative effects of such mixes in the accretion and diminution of place effects.

The aim of this paper is to create data-driven Great Britain-wide platial profiles by relating the geographic origins of surnames to the locations of their bearers at different points in time. From a geodemographic perspective, this entails sifting contemporary residential structure into residents whose roots in a locality likely extend over many generations, and those whose surnames indicate bloodlines that are less established in localities. The defining characteristics of places are by no means seen as invariably grounded in inter-generational history, but our approach seeks to decompose the repetitive social similarities that characterize different locations within the national settlement system from the underlying historic structure of communities that underpins this more recent socio-spatial differentiation. Our motivation is thus to address the question: ‘what place is like this place?’ We conclude with some brief speculations concerning the ways in which this allows us to address issues of regional development and functional inter-dependence within the settlement system.

2 Creating Platial Surname Profiles

2.1 Data Sources

Individual-level historic census records are made publicly available 100 years after their collection date. Higgs and Schurer (2014) have brought together and standardized digital transcriptions of the censuses for England and Wales collected between 1851 and 1911 excluding 1871; data for Scotland are available for the full period 1851 to 1901. The individual census records are linked to parishes, the boundaries of which have been digitized according to two sets of consistent parish geographies. While disclosure of individual names and addresses is presently not possible from post-1911 UK censuses, a near complete set of linked addresses-level consumer registers has been assembled for the period 1997–2016 by the UK Consumer Data Research Centre (Lansley et al., 2019). This corpus of data comprises the public version of the UK Electoral Register supplemented with various consumer data sources from 2002 onwards to capture many of those that opted out of inclusion of the public version of the register or were ineligible to vote in any elections. While the linked consumer registers are incomplete and are of uncertain provenance, they have been subjected to extensive computational address matching procedures and have been partially validated by triangulation with annual Office for National Statistics’ mid-year population estimates. Through these processes of internal and external validation, the Consumer Registers are found to comprise the vast majority of the UK’s adult population (Lansley et al., 2019). In what follows, we examine the local, regional, national, and international origins of surnames for one historic period (1881) and two much more recent points (1998 and 2016).

2.2 Surname Origins

Using the 1881, 1998, and 2016 data we begin by locating the recorded residences of individuals bearing British surnames relative to the probable historic seed point of the name. To this end, each surname with at least 30 bearers¹ in 1851, 1861, or 1881 is assigned a geographic centroid, calculated using kernel density estimation (KDE).² For each surname, the isotropic fixed bandwidth is estimated using a likelihood cross-validation method, constrained by a minimum bandwidth of 5 kilometres and a maximum bandwidth of 40 kilometres. For computational reasons, we estimate the bandwidth on a sample of the surname population in cases where a surname has more than 5,000 bearers. Although, it is theoretically possible to vary the bandwidth according to the distribution of the background data to accommodate local variations, for large data sets this is extremely challenging (cf. Zhang et al.,

2017). All KDE calculations are executed using the R programming language (R Core Team, 2019) and the *Sparr* package (Davies et al., 2018). To speed up processing times, calculations are parallelized using GNU Parallel (Tange, 2011) and distributed over a high-performance Linux cluster. This process resulted in 59,218 centroids, each defined as the maximum relative density value of the associated KDE for the qualifying surnames. This set is defined as our ‘long-settled surname stock of British names’ in the subsequent analysis (for full details on the KDE calculations and its parameters, see van Dijk et al., 2019; van Dijk and Longley, 2020).

2.3 Spatial-Temporal Comparison

A temporally consistent zonal design is required in order to compare changing surname mixes over time. We create this by combining the consistent parishes from the historic censuses with the contemporary 2011 Office for National Statistics (ONS) Middle Layer Super Output Area (MSOA) geography. Because some of the parishes are relatively small, particularly in urban areas, we start by iteratively merging historic parishes until they reach a minimum threshold of 750 inhabitants.³ In a second step, we assign MSOA centroids to these aggregated parishes using a point in polygon procedure, and MSOAs assigned to the same parish are merged. Parishes that do not have a single MSOA centroid within their boundaries are merged with neighbouring parishes. This procedure results in an MSOA-based temporally consistent zones (TCZs) of Great Britain consisting of 3,370 zones (cf. Kandt et al., 2020). On average each TCZ is 67 square kilometres (standard deviation: 150 square kilometres). The average population sizes of these zones in 1881, 1998, and 2016 are 8,914, 13,473 and 15,256 respectively. Once the TCZs have been defined, we create two lookup tables: one assigning MSOAs to TCZs and one assigning historic parishes to TCZs.

2.4 Origin-Destination Matrix

Next, surname migration distance is approximated using a Delaunay triangulation that connects the 3,370 TCZ centroids. Edges that are not consistent with the UK’s coastal geography are removed manually. This Delaunay network is then used to create an origin–destination matrix in which each TCZ is combined with all other TCZs using Dijkstra’s shortest path algorithm, in which edge length is used to measure impedance. The Delaunay triangulation is executed using the *spatstat* package in R (Baddeley et al., 2015). The creation of the origin–destination matrix is achieved through the *NetworkX* Python library (Hagberg et al., 2008).

2.5 Surname Migration Distances

Our platial profiles are based upon the mix of ‘surname migration distances’ of those residing within each TCZ. We thus assign each of the 59,218 long-settled surname origins to the centroid of the closest TCZ, effectively linking each surname to the origin–destination matrix. For each TCZ we then extract the individuals bearing the long-settled surnames and use the origin–destination matrix to calculate the distance that the bearers have apparently migrated from their likely origin region to their TCZ of residence. These calculations are made for 1881, 1998, and 2016. The resulting tables for 1881, 1998, and 2016 record the distances migrated by all long-term surname bearers in each TCZ.

In a last step, for every surname and TCZ combination, the log value of every migration distance is weighted by the log value of the average distance between the TCZ and all other TCZs. This yields a log-based ratio between the average distance to all other TCZs and the distance over which the bearers of the surnames have migrated. This compensates for the degree of centrality of the TCZ within Great Britain. Figure 1 shows the distribution of these values for a 1 per cent sample of the 1881 data set (about 3,000,000 rows).

2.6 Platial Profiles

The weighted distances are then classified into four origin categories: local, regional, national, and international, where the international category is comprised of surnames that do not appear in the dictionary of long-settled surnames. The classification is informed by the distribution shown in Figure 1 to estimate groups of roughly similar size. For example, surname y in TCZ k with a log-distance under

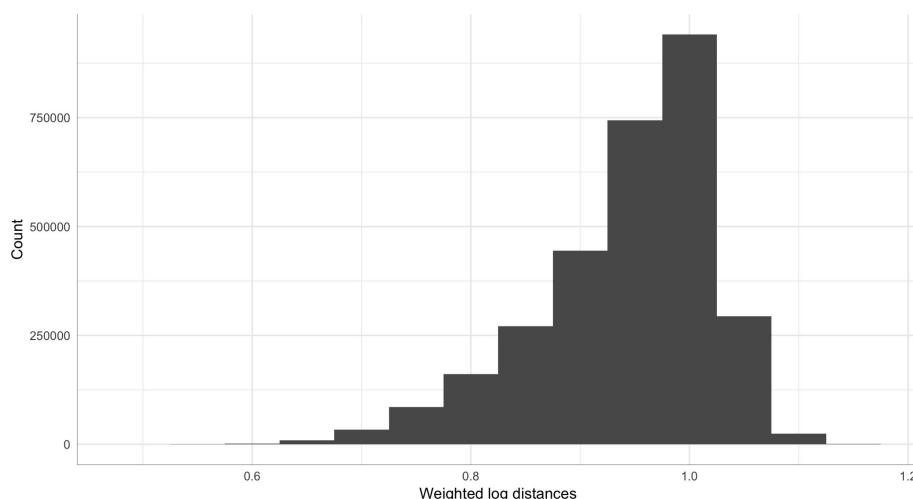


Figure 1: Weighted log-distances. Distribution of a random sample of individual long-settled surname log-distances weighted by the average log-distances per TCZ.

0.85 is considered local to this TCZ. However, this also implies that when surname y is simultaneously present in TCZ m it could have a different log-distance ratio and therefore, e. g., be considered as having a regional or wider national origin. That is, the values represent a level of locality that is relative to the TCZ centroid under consideration. Table 1 shows the distance classification used for these weighted log-distances.

3 Temporal and Regional Differences in Platial Profiles

The surname distance classification can be used to map the proportions of the population designated as having local, regional, national, or international roots. This is shown in Figure 2 for the three time periods. In 1881 areas with a high proportion of local names are found throughout the country. Particularly high proportions (over 70 per cent) of the population that are local to the area in which they reside are found in some parts of Scotland and north Wales. By contrast some major urban areas such as London and Birmingham record only 10 to 20 per cent of individuals as bearing local surnames. In 1998 and 2016, the proportions of individuals with local names have dropped significantly for all areas, indicating an increased mixture of populations.

The populations classified as having more regional roots, as presented in Figure 2b, show a similar trend over time. Other parts of Scotland and Wales have zones in which more than 70 per cent of the population is classified as regional in 1881. However, even though in general terms the regional population has decreased for most areas in 1998 and 2016, the decline is less pronounced than that of the local population. Low (below 20 per cent) proportions of individuals with regional names are found in English cities, with London being the primary exemplar.

Because the different classifications are mutually exclusive, the maps of the proportion of the population with national and international names (Figures 2c–2d) present inversed pictures of the local and regional maps. It is therefore not very surprising that the areas with the highest proportions (40 to 50 per cent) of individuals bearing national surnames in 1881 are predominantly situated in England. Rural areas typically exhibit lower proportions of surnames drawn from other parts of the country. In 1998 and 2016 it can be clearly seen that national surnames have spread throughout most

Table 1: Classification. Distance classification based on individual long-settled surname log-distances weighted by the average log-distances per TCZ.

Local	< 0.85
Regional	0.85–0.95
National	> 0.95

of the country – again manifesting the increased mixing of the long-settled population in Great Britain. London stands out as a location with a low share of individuals with national surnames (20 to 30 per cent), especially in 2016. This may seem counter-intuitive but is explicable by examining the results for international names shown in Figure 2d; by 2016 some areas in London host populations for which 60 to 70 per cent of names are not deemed to have origins in Britain.⁴

4 Discussion

Our study demonstrates that geographical surname classifications can be used to decompose local populations into local, regional, national, and international components. These classifications can be used to underpin platial geo-temporal classification of community structure. Similar to Christaller's Central Place Theory, the results cautiously suggest that a hierarchy of places exists within the British settlement system. Large conurbations, predominantly situated in England, are characterized by surnames that can be traced back to most other areas in Great Britain. London is clearly at the top of this settlement hierarchy, as identified by the high preponderance of international names. Potentially this hierarchy can be used in describing the relationship between surname diffusion and economic outcomes, such as those related to social mobility.

Some limitations of this analysis should be acknowledged. First, spatial heterogeneity in local, historic naming conventions is not fully accommodated. For example, the baseline Welsh and Scottish populations have less diverse ranges of surnames, making it possible that small increases in surnames imported from elsewhere will lead to disproportionately large apparent changes. Secondly, our surname classification is somehow arbitrary – adjusting the class boundaries of the classification will have an impact on the results. Thirdly, extremely widespread surnames (e.g., Smith) are assigned to a single origin on the basis of tiny marginal density values that bear little or no correspondence with place effects. The same applies too many other names because secondary peaks in the KDE density surface are currently not taken into consideration. Further research into this topic is likely to treat popular names, especially surnames pertaining to widely practised occupations, as having less rigidly defined origins, for instance, by employing a fluid concept of origins that defines origins at different geographical scales (cf. Kandt et al., 2020). Another solution would be to quantitatively determine the informational content of each surname, and use this informational content as a criteria for inclusion or exclusion (see Güell et al., 2018, 2015).



Notes

1. Because a KDE requires sufficient input data points, the choice for a minimum of 30 bearers is pragmatic.
2. In 1871 historic census records are only digitally available for Scotland and 1871 is therefore excluded from analysis. For surnames that have 30 or more bearers in multiple years, the earliest available year is used to determine the surname's origin.
3. Code to iteratively merge (aggregate) adjacent polygons is stored in a Postgres/PostGIS database using a minimum threshold. It is available on GitHub (<https://github.com/jtvandijk/pg-polygon-merge-repo>).
4. These surnames may nevertheless have British origins, such as surnames that had a frequency under 30 in 1851, 1861, and 1881.

Funding

This work is supported by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1 and EPSRC grant EP/M023583/1 (UK Regions Digital Research Facility).

ORCID

Justin van Dijk  <https://orcid.org/0000-0001-5496-425X>
Paul A Longley  <https://orcid.org/0000-0002-4727-6384>

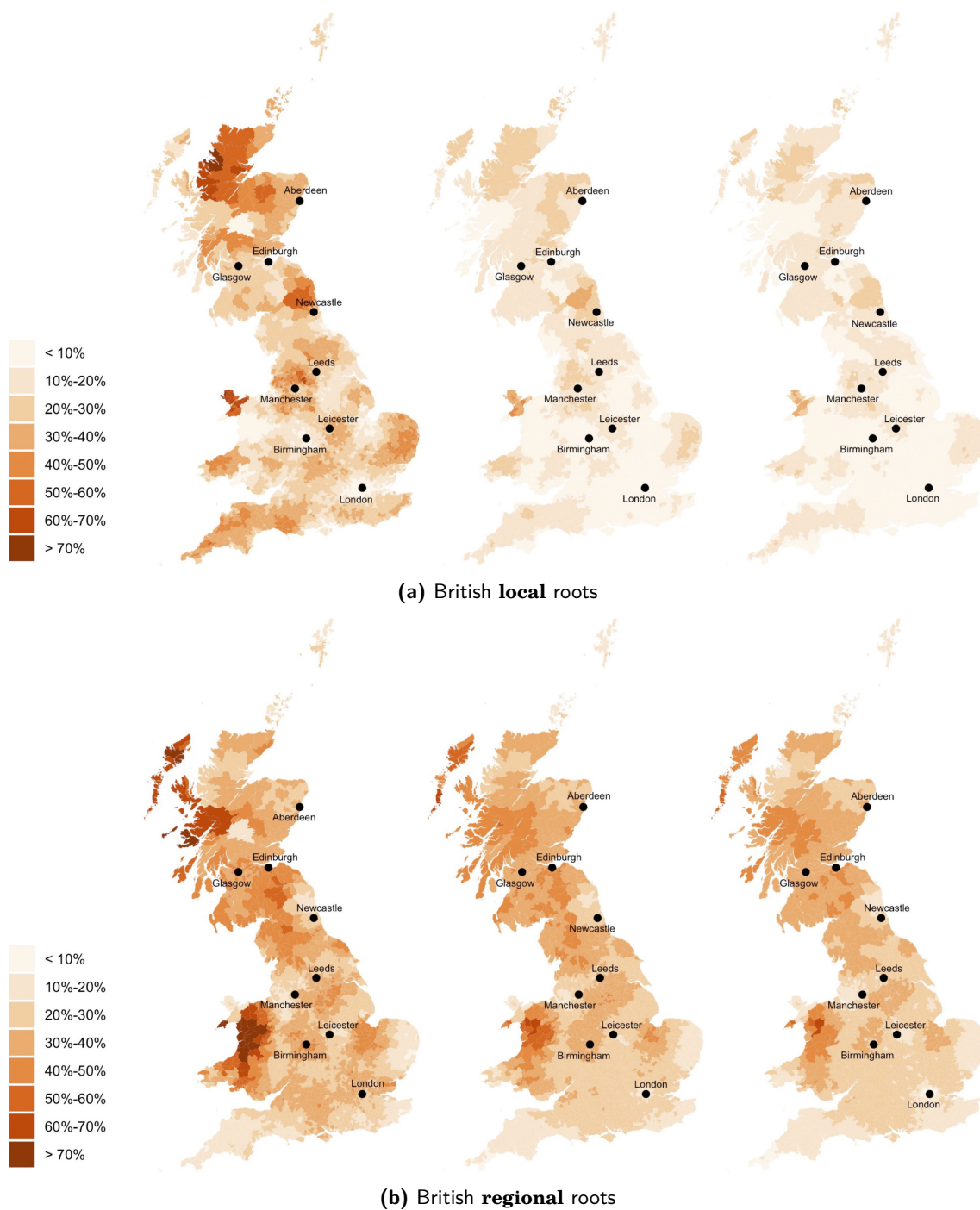
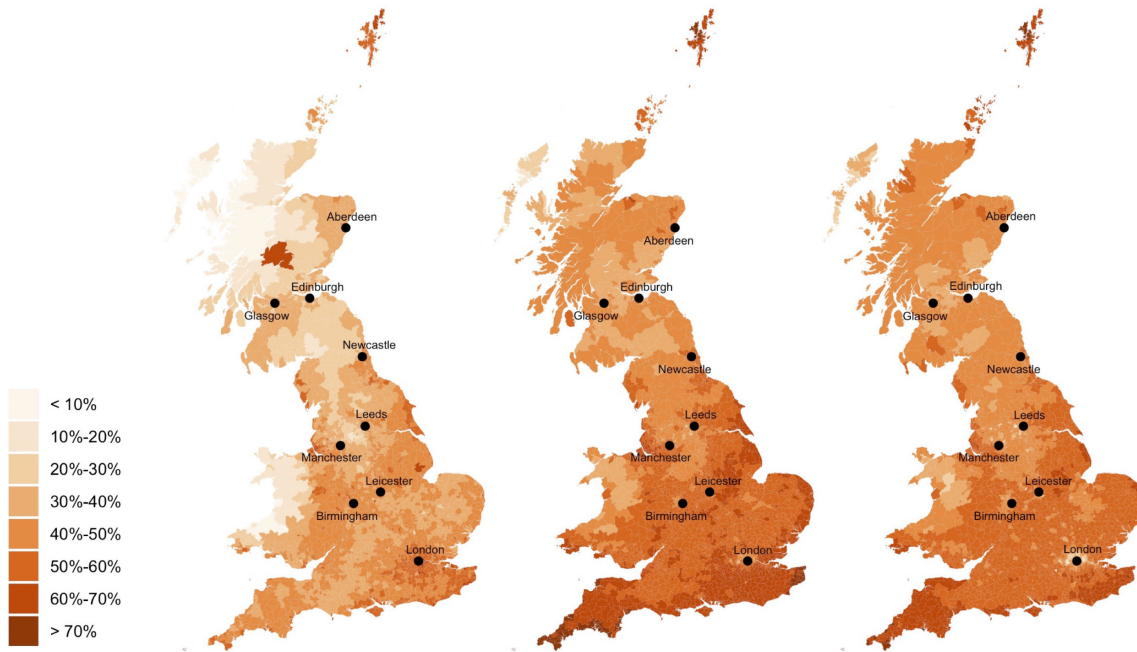
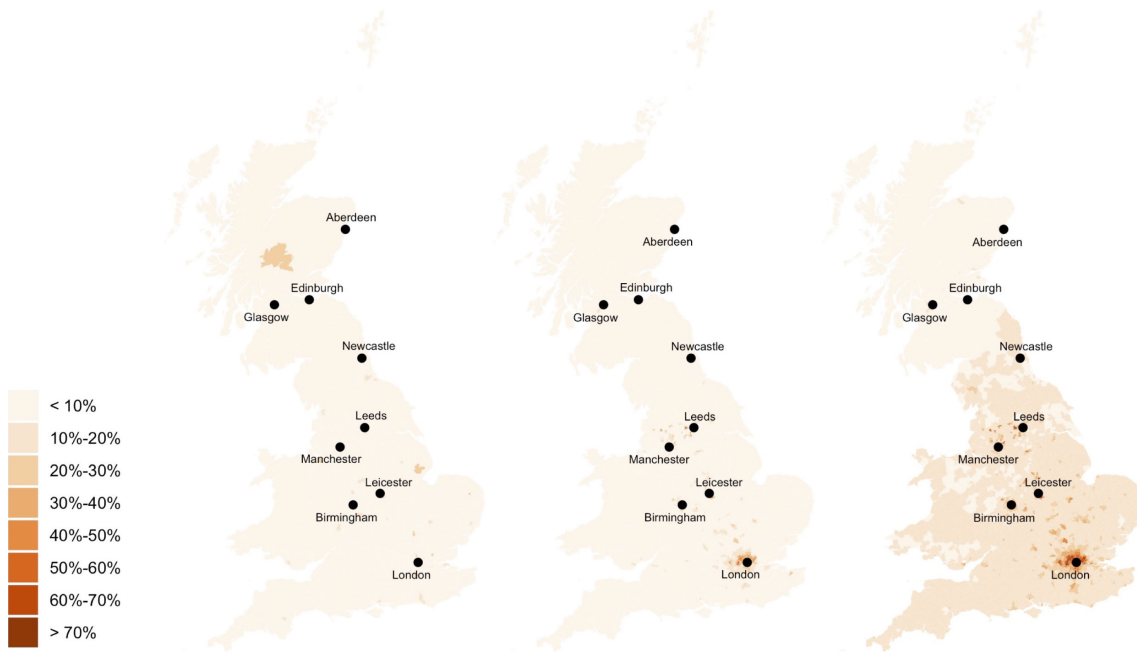


Figure 2: Platial profile. Population classified by their roots in 1881, 1998, and 2016.



(c) British **national** roots



(d) British **international** roots

Figure 2 (continued): **Platial profile**. Population classified by their roots in 1881, 1998, and 2016.

References

- Baddeley, Adrian; Rubak, Ege; and Turner, Rolf: *Spatial point patterns: methodology and applications with R*. London, UK: CRC Press, 2015
- Cheshire, James and Longley, Paul A: *Identifying spatial concentrations of surnames*. *International Journal of Geographical Information Science*, 26(2), 2012, 309–325. doi: 10.1080/13658816.2011.591291
- Davies, Tilman M; Marshall, Jonathan C; and Hazelton, Martin L: *Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk*. *Statistics in Medicine*, 37(7), 2018, 1191–1221. doi: 10.1002/sim.7577
- van Dijk, Justin; Lansley, Guy; Lan, Tian; and Longley, Paul A: *Using the spatial analysis of family names to gain insight into demographic change*. Proceedings of the 27th Conference on GIS Research UK (GISRUK), 2019
- van Dijk, Justin and Longley, Paul A.: *Interactive display of surnames distributions in historic and contemporary Great Britain*, n. d. Manuscript submitted for publication
- Güell, Maia; Pellizzari, Michele; Pica, Giovanni; and Rodriguez Mora, Jose Vicente: *Correlating social mobility and economic outcomes*. *The Economic Journal*, 128(612), 2018, F353–F403. doi: 10.1111/eoj.12599
- Güell, Maia; Rodriguez Mora, Jose Vicente; and Telmer, Christopher I: *The informational content of surnames, the evolution of intergenerational mobility, and assortative mating*. *The Review of Economic Studies*, 82(2), 2015, 693–735. doi: 10.1093/restud/rdu041
- Hagberg, Aric; Schult, Daniel; and Swart, Pieter: *Exploring network structure, dynamics, and function using NetworkX*. Proceedings of the 7th Python in Science Conference (SciPy), 2008, 11–15
- Harris, Richard; Sleight, Peter; and Webber, Richard: *Geodemographics, GIS and neighbourhood targeting*. Chichester, UK: Wiley, 2005
- Higgs, Edward and Schurer, Kevin: *Integrated census microdata (I-CeM), 1851-1911 [data collection]*. Data record in the UK Data Service data catalogue, 2014. doi: 10.5255/UKDA-SN-7481-1
- Kandt, Jens; van Dijk, Justin; and Longley, Paul A: *Family name origins and inter-generational demographic change in Great Britain*. *Annals of the American Association of Geographers*, 2020, in press
- Kandt, Jens and Longley, Paul A: *Ethnicity estimation using family naming practices*. *PLOS ONE*, 13(8), 2018, e0201774. doi: 10.1371/journal.pone.0201774
- Lan, Tian; Kandt, Jens; and Longley, Paul A: *Ethnicity and residential segregation*. In: Longley, Paul A; Singleton, Alex; and Cheshire, James A (eds.), *Consumer Data Research*, London, UK: UCL Press, 2018. 71–83
- Lansley, Guy; Li, Wen; and Longley, Paul A: *Creating a linked consumer register for granular demographic analysis*. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4), 2019, 1587–1605. doi: 10.1111/rssa.12476
- Longley, Paul A; Webber, Richard; and Lloyd, Daryl: *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), 2007, 31–48. doi: 10.1111/j.1467-8306.2007.00522.x
- Martin, David; Cockings, Samantha; and Leung, Samuel: *Developing a flexible framework for spatiotemporal population modeling*. *Annals of the Association of American Geographers*, 105(4), 2015, 754–772. doi: 10.1080/00045608.2015.1022089
- R Core Team: *R: a language and environment for statistical computing*. <https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing>, 2019. Retrieved 27 November 2019

Singleton, Alex D and Longley, Paul A: *Data infrastructure requirements for new geodemographic classifications: the example of London's workplace zones*. *Applied Geography*, 109, 2019, 102038. doi: 10.1016/j.apgeog.2019.102038

Tange, Ole: *GNU Parallel – the command-line power tool*. ;login: *The USENIX Magazine*, 36(1), 2011, 42–47

Zhang, Guiming; Zhu, A-Xing; and Huang, Qunying: *A GPU-accelerated adaptive kernel density estimation approach for efficient point pattern analysis on spatial big data*. *International Journal of Geographical Information Science*, 31(10), 2017, 2068–2097. doi: 10.1080/13658816.2017.1324975