Using the spatial analysis of family names to gain insight into demographic change

Justin van Dijk[*1], Guy Lansley[†1], Tian Lan[‡1], and Paul A. Longley[§1]

[1] Department of Geography, University College London

January 18, 2019

Summary

This paper describes the steps involved to prepare the largest ever quantitative analysis of the distribution of surnames in Great Britain. We describe the method to estimate approximately 1.2 million surname distributions using Kernel Density Estimates (KDEs) for seven years of historic census data and twenty years of contemporary Consumer Registers. We argue that these surname distributions could offer valuable insight into processes of contagious and hierarchical diffusion of populations as well as the regional distinctiveness of demographic change and stasis.

KEYWORDS: surname distributions; kernel density estimates; historic census; consumer registers

1. Surnames as socio-spatial indicators

In the United Kingdom and many other countries, surnames contain at least two properties that make them informative markers for a number of socio-spatial processes. First, in many instances, surnames are hereditary through the patrilineal line. Second, as a result of regionally varying naming practices, many surnames can be traced back to a national or regional origin (Cheshire & Longley, 2012). As such, family names have been used in various applications, such as inferring ethnicity (Lan et al., 2018) and measuring ethnic segregation (Kandt & Longley, 2018). However, there has been less systematic research into regional surname origins and long-term changes in distinctive surname mixes at a sub-national level. This paper argues that surname distributions can offer insights into the long-term impact of demographic change on place composition by combining family naming records from historic population censuses with contemporary Consumer Registers.

2. Data sources

Digitally encoded Historic Census data for England, Scotland, and Wales provide population-wide micro data of individuals' names and addresses (Higgs & Schürer, 2014). This integrated collection of historic census microdata covers the decennial census of England and Wales for 1851, 1861, 1881, 1891, 1901, 1911, and for Scotland for the period 1851 to 1901. The addresses are linked to parishes, which have been digitised into a set of two consistent parish geographies. Because contemporary censuses do not disclose names, extensive databases of names at the addresses-level were sought from public versions of the electoral register from 1997 until 2016, with supplements from consumer data from 2002 onwards to capture those that opt-out or are not eligible to vote. The 'Consumer Registers' are linked to bolster their coverage and are found to be representative of the vast majority of the UK's adult population (Lansley et al., 2018).

[*] j.t.vandijk@ucl.ac.uk
[†] g.lansley@ucl.ac.uk
[‡] tian.t.lan@ucl.ac.uk
[§] p.longley@ucl.ac.uk

3. Kernel Density Estimation

One way to analyse and compare surname distributions over time without being hindered by changing administrative areas is by point pattern analysis. We first assign every individual found in the historic census data to the centroid of the parish with which they are associated. Similarly, we geocode all individuals found in the Consumer Registers directly through the coordinates associated with each postcode. We subsequently map the spatial patterns of these point events on a surname-by-surname basis through a process called Kernel Density Estimation (KDE). KDE is a non-parametric method that places a search window (kernel) over a point and uses the information within this kernel to estimate point densities. A KDE applied over two-dimensional space can be formally described as follows (Shi, 2010, p.643):

$$\hat{f}(x,y) = \frac{1}{nh^2}\sum_{i=i}^{n} K(\frac{d_{i,(x,y)}}{h}) \quad (1)$$

where $\hat{f}(x,y)$ is the estimated density at location (x,y), n is the number of point events that fall within the bandwidth h, $d_{i,(x,y)}$ is the distance between the location (x,y) and an event point i. Lastly, K is the density function that describes the contribution of point i to the estimated density at location (x,y).

Two major disadvantages of using KDEs is that they are relatively slow to calculate and that they are calculated over a regular grid. This leads to long processing times and requires exponentially more storage with an increasing grid resolution. To mitigate these storage issues, we develop a method to compress the raster information by using some of the raster grid's properties. First, because the raster grids are applied on the same area over and over again, they are consistent in shape, extent, and resolution for all surnames. As such, we only have to store the XY-coordinates of each grid cell once. We simply extract all XY-coordinates from the grid and store them together with their index position Second, we extract the point density estimates for each cell for every calculated surname distribution. We start by rescaling the point density estimates on a scale ranging from 0-100 so that the results between different years and surname distributions are comparable. Subsequently, because the KDEs often result in sparse matrices, we only retain the index position and the value of the cells with a value greater than or equal to 1. In turn, we write the string of indices and values to a database. When we need the KDEs again, the XY-coordinates of the grid can easily be joined with the point density estimates using index value as a key. This process allows for an effective raster compression and deconstruction, and an on the fly raster reconstruction with minimal resources.

All KDEs are calculated in R (R Core Team, 2014) using the "Sparr" package. To speed up processing times all calculations are parallelised using GNU Parallel (Tange, 2011) and executed on a high-performance computing cluster before they are stored in a Postgres database. To account for the spatial heterogeneity of the population throughout the United Kingdom, for each surname all the individuals bearing this name are weighted by the population. Lastly, for the Historic Censuses, KDEs are only calculated for surname populations of at least 30 individuals and for the contemporary Consumer Registers the surname populations consisting of at least 50 individuals. The total number of calculated and successfully stored KDEs for each year, the number individuals that are represented by these KDEs, and the size of the entire population that is within our dataset is shown in **Table 1** and **Table 2** for the Historic Censuses and the Consumer Registers, respectively.

4. KDE examples

Figure 1 and **Figure 2** give an example of two of the KDE-based surname geographies. The black contour lines indicate the areas that have the highest relative density of surname occurrences for "Rossall" and "Lansley", respectively. It is clear that both surnames have quite a different geography; with particularly the contemporary Rossall's still being mostly concentrated in the same area as their ancestors in 185,1 whereas the Lansley's has spread out over the South-West of England.

Table 1 Overview Kernel Density Estimates Historic Censuses

Year	Extent	# Surnames (n >= 30)	# Individuals	# Total Individuals
1851	England, Scotland, Wales	37,923	19,072,209	20,511,998
1861	England, Scotland, Wales	49,480	20,985,507	22,430,761
1871	Scotland	5,725	3,116,540	3,347,851
1881	England, Scotland, Wales	39,872	28,235,641	29,821,393
1891	England, Scotland, Wales	47,865	27,626,365	32,978,628
1901	England, Scotland, Wales	44,820	35,386,469	36,868,783
1911	England, Wales	45,307	34,325,533	36,276,903

Table 2 Overview Kernel Density Estimates Consumer Registers

Year	Source	# Surnames (n >= 50)	# Individuals	# Total Individuals
1997	Electoral register	41,250	42,347,715	45,128,534
1998	Electoral register	42,299	44,136,104	46,982,474
1999	Electoral register	42,630	44,490,347	47,382,611
2000	Electoral register	42,679	44,337,450	47,218,923
2001	Electoral register	42,118	43,296,297	46,100,648
2002	Consumer sources	43,178	44,313,592	47,269,669
2003	Consumer sources	43,014	43,301,593	46,302,577
2004	Consumer sources	43,425	43,403,302	46,542,176
2005	Consumer sources	43,746	42,919,401	46,207,146
2006	Consumer sources	44,426	43,091,549	46,561,515
2007	Consumer sources	45,366	43,590,413	47,234,394
2008	Consumer sources	46,112	43,939,841	47,722,361
2009	Consumer sources	47,348	45,210,895	49,181,333
2010	Consumer sources	49,766	46,396,443	50,578,969
2011	Consumer sources	48,411	45,781,304	49,971,710
2012	Consumer sources	48,798	45,178,124	49,578,069
2013	Consumer sources	49,814	46,320,748	50,862,892
2014	Consumer sources	50,415	46,991,971	51,622,349
2015	Consumer sources	50,264	46,992,768	51,637,090
2016	Consumer sources	50,488	47,387,529	52,109,263

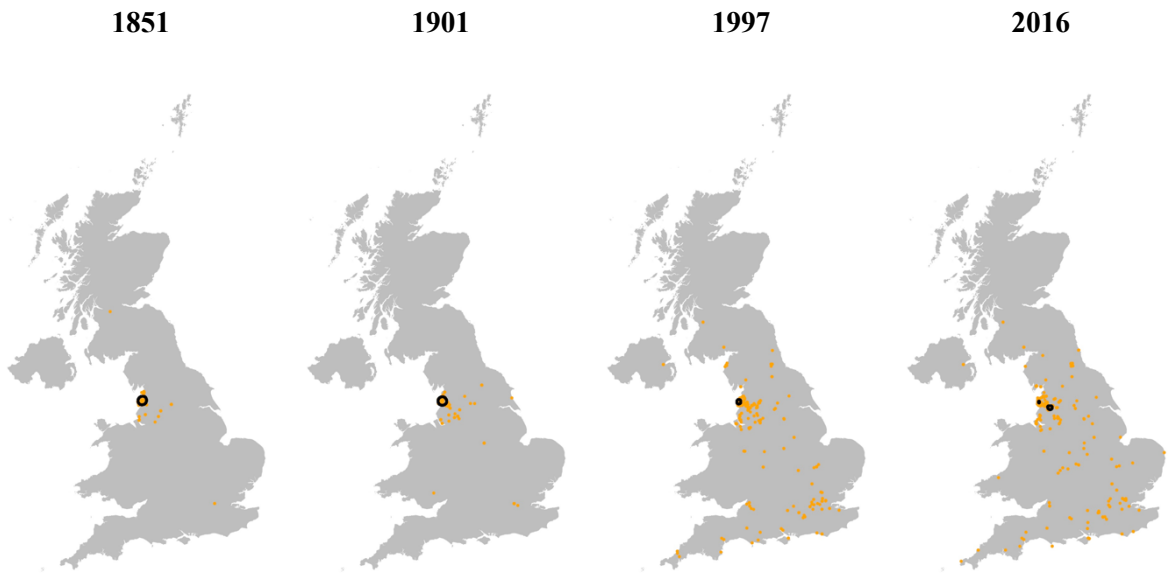

Figure 1 Kernel Density Estimate "Rossall"

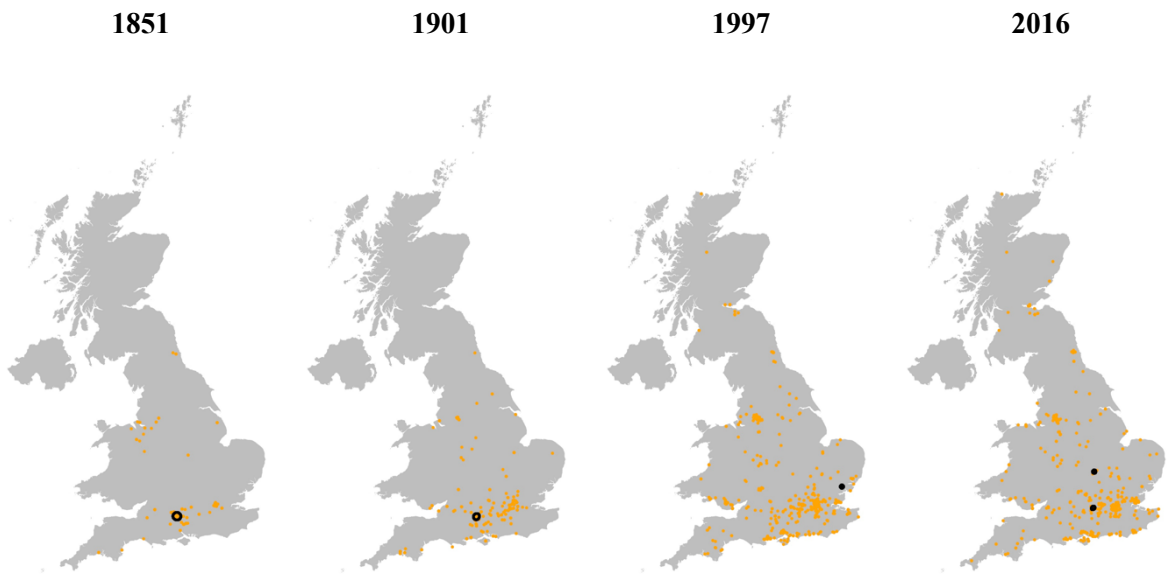

Figure 2 Kernel Density Estimate "Lansley"

5. Conclusion

The calculation of approximately 1.2 million KDEs allows for the linkage of places through ancestral lines. As such, the spatial analysis of surnames can potentially be used to gain insight into demographic change. First, the degree of concentration of surnames within their most likely region of origin (as could be defined by the area of highest relative density in 1851), could provide useful to identify surnames that can be considered truly regional and surnames that are common in a larger area (e.g. metonyms like "Smith"). The changing distribution of these regional names throughout the country could be an effective proxy for migration. Second, places can be characterised by looking at the mixture of different surnames that are present; e.g. areas largely dominated by individuals that bear a surname that has 'always been there' are rather different from areas that have a mix of surnames that can be traced to local, regional, national, and international regions. Furthermore, the methodology that we developed to

compress and store raster grids makes it feasible to create a database that is linked to a website that allows anyone to explore their own surname geography; this website is currently under development. Added advantage of this method is that the website at no point requires a connection to a database with sensitive individual level data.

6. Acknowledgements

This work is funded 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').

Biographies

Justin van Dijk is a Research Associate at the Urban Dynamics Lab and the Department of Geography at University College London. His primary research interests are grouped around the analysis and visualisation of large-scale spatial data, urban mobility, and geographic information systems in general.

Guy Lansley is a Research Associate at the UK Consumer Data Research Centre and the Department of Geography at University College London. His research is primarily focused on harnessing geodemographic insight from big consumer datasets of unknown provenance.

Tian Lan is a Research Associate, working in the Geospatial Analytics and Computing group at the Department of Geography at University College London. He mainly works on the research of contemporary residential segregation using Consumer Registers and the linkage of historical census records.

Paul Longley is Professor of Geographic Information Science at University College London and director of the UK Consumer Data Research Centre at UCL. His publications include 18 books and more than 150 refereed journal articles and book chapters.

References

Cheshire, J.A. & Longley, P.A. (2012) Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*. 26 (2), 309–325.

Higgs, E. & Schürer, K. (2014) *Integrated Census Microdata (I-CeM), 1851-1911*, [data collection]. UK Data Service SN: 748. DOI: 10.5255/UKDA-SN-7481-1

Kandt, J. & Longley, P.A. (2018) Ethnicity estimation using family naming practices. *PLOS ONE*. 13 (8), e0201774.

Lan, T., Kandt, J. & Longley, P.A. (2018) Ethnicity and residential segregation, in Paul A. Longley, Alex Singleton, & James A. Cheshire (eds.) *Consumer Data Research*. London: UCL Press. pp. 71–83.

Lansley, G., Li, W. & Longley, P.A. (2018) Modelling small area level population change from administrative and consumer data, in *Proceedings of the 26th Conference on GIS Research UK (GISRUK)*. Leicester: University of Leicester.

Shi, X. (2010) Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. *International Journal of Geographical Information Science.* 24 (5), 643–660.

Tange, O. (2011) GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*. 36 (1), 42–47.