Some challenges for updating neighbourhood geodemographic in the context of the 2021/2 Output Area Classification.

Jakub Wyszomierski[*1], Paul A. Longley[†1], Christopher G. Gale[‡2]

[1]Department of Geography, University College London, UK
[2]Office for National Statistics, Titchfield, Fareham, Hampshire, UK.

February 10, 2021

Summary

This paper explores ongoing efforts to develop a framework for regularly updating geodemographic classifications with national coverages. We discuss the motivation and methods for improving the temporal granularity of geodemographic segmentations and necessity for doing so in light of de-harmonisation of the 2021/2 Census. We bring focus to the deployment of spatial microsimulation techniques for improving the completeness of a highly granular and frequently updated UK-wide dataset. We discuss the contribution of this research to develop the 2021/2 Output Area Classification.

KEYWORDS: geodemographics, Output Area Classification, 2021 UK Census, spatial microsimulation

1. Introduction

In the light of the increasing complexity of the world that is also accompanied by the growth in the abundance of data, the field of geodemographics is facing an inevitable shift in ways it is materialized and put to use. Conventionally, two types of geodemographic classifications have been developed – open and closed classifications (Gale *et al.*, 2016). Whereas open segmentations make use of open-source census statistics, closed classifications typically revolve around exclusive consumer data (supplemented by census outputs). Importantly, reliance of open classifications upon Census data means that they are of a high quality and are of known provenance but have poor temporal granularity. Contrarily, closed classifications have a much higher spatio-temporal granularity but may not provide full coverage of the population and have low transparency and reproducibility.

2. Hybrid geodemographics

Recognising an increasing cost of conducting censuses and a need for much frequent yet still comprehensive information about the population, we are experiencing "the twilight of the census" (Coleman, 2013). Consequently, much of current research on geodemographics focuses on alternative ways of deriving reliable and frequent population estimates, that can offer information on socio-economic characteristics invaluable for geodemographic segmentations. Appropriately, this study evaluates the future of population estimates and geodemographic classifications, by exploring ways of data concatenation and conflation that preserves strengths of each data source. The importance of this work is exemplified by the disruptions to the 2021 Census, with the Scottish enumeration postponed by a year. Thus, the aim of the research is both to develop the next Output Area Classification and address the issue of Census de-harmonisation.

2.1. Data

Appreciating the strengths of different data sources, the hybrid geodemographics framework proposes

[*] jakub.wyszomierski.16@ucl.ac.uk
[†] p.longley@ucl.ac.uk
[‡] chris.gale@ons.gov.uk

a data combination in a way that all good features and strengths are retained. By acknowledging different ways in which each data source is collected and structured, the framework recognizes that a combination of different yet complementary data sources and analysis techniques should be implemented.

The research makes use of primarily two data sources – 2011 Census outputs and Linked Consumer Registers (LCR) data. LCR data have been developed by the Consumer Data Research Centre (CDRC) and consist of over 143 million linked records of individuals aged 17 or above who have been recorded in a register at least once between 1996 and 2016 (Lansley *et al.*, 2019). As LCR provide individual-level data that include name, surname and address of each individual, linking subsequent (primarily electoral) registers enables analysis of the population mobility and societal changes.

2.2. Methods

To derive basic demographic characteristics at the individual level, the research has made use of the Ethnicity Estimator (EE) algorithms. The method discussed by Kandt and Longley (2018) utilizes surname patterns to model ethnicity of an individual. The technique has shortcomings, particularly with regard to some ethnic groups, for which the surname patterns are under-developed, but it nevertheless provides a useful and rare source of population characteristics at the individual level. The analysis of individuals and their movements over the years means that the ethnicity estimates also reflect societal changes and shifts in the structure of ethnic groups.

Furthermore, to combine comprehensiveness and reliability of census statistics with much finer temporal and spatial granularity offered by LCR, it is suggested that spatial microsimulation techniques can be successfully implemented. Incomplete individual-level LCR can be complemented with the aggregated statistics on the population recorded in 2011 Census. By generating complete individual level data with synthetic observations resembling individuals missing from input data, a synthetic population is derived; such dataset maintains comprehensiveness of aggregated statistics, while providing high spatio-temporal granularity of micro-data (Lovelace and Dumont, 2016). This research employs Iterative Proportional Fitting (IPF), a technique proposed by Deming and Stephan (1940) that progressively adjusts cell totals of the input data so that they match up with the margin totals of the constraint data (Lovelace *et al.*, 2015). In the analysis reported here, the population was synthesised with the use of three demographic variables: age, sex and ethnicity. Whereas ethnicity was derived with the use of EE algorithms, sex and age group were modelled using a classifier that infers the age and sex of individuals based upon forenames (Lansley and Longley, 2016).

3. Empirical analysis

The following section presents results of the initial empirical analysis conducted as a part of the research. Due to the nature of LCR data, the comparison covers the estimates of the adult population only. The analysis focused on London and its boroughs.

3.1. Population and ethnicity estimates

According to the 2011 Census, there were over 6,362,000 adults living in London, while the LCR recorded 6,125,000 adults. When looking at different aggregation units, it is evident that the coverage rate is subject to spatial variability. **Figure 1** presents how the correspondence between 2011 Census and 2011 LCR estimates vary by boroughs. Interestingly, there is an over-estimation in the Central and East London. Such high coverage rates may be attributed to a high proportion of certain demographic groups that may have a high probability of successful registration on the Electoral Roll, which constitutes a major source of LCR data.

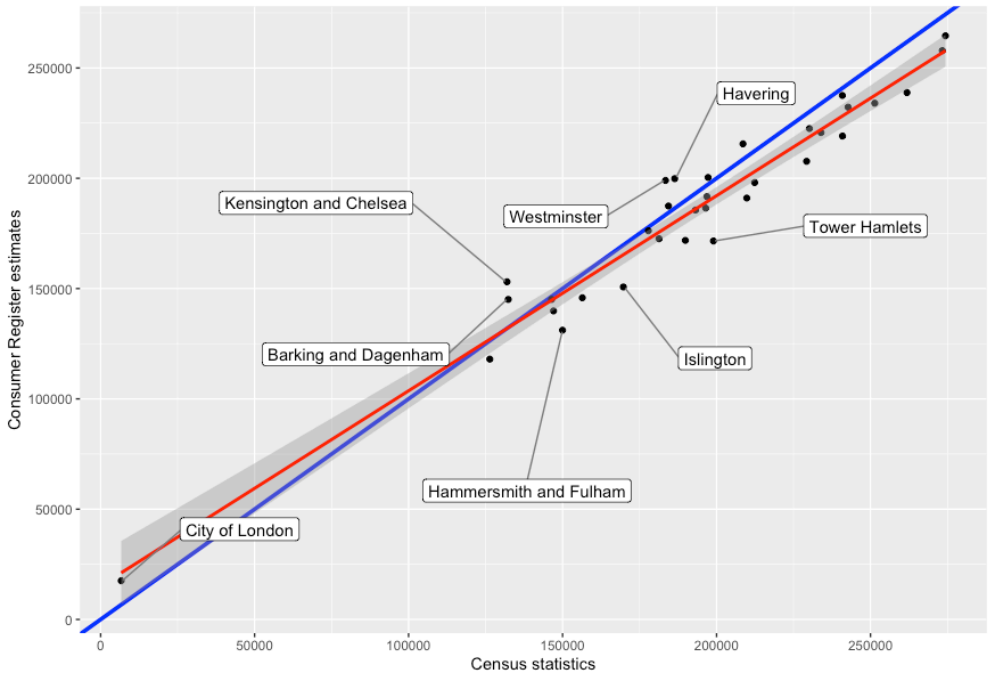

Figure 1 Adult population estimates at Local Authority level for 2011 Census (blue line) and 2011 LCR data (red line).

Importantly, the analysis shows that the estimates vary both by space and demographic group. **Table 1** presents LCR estimates for each ethnic group compared with those from the 2011 Census.

Table 1 Adult population estimates for each ethnic group in London

Ethnic group	Total 2011 LCR counts	Total 2011 Census counts for the adult population	Coverage rate
White British	3,401,430	3,001,048	113.34%
White Other	749,660	878,797	85.3%
White Irish	59,660	163,996	36.38%
Asian Indian	507,420	444,856	114.07%
Asian Pakistani	315,480	156,314	201.83%
Asian Other	126,535	301,149	42.02%
Asian Bangladeshi	125,190	141,946	88.2%
Asian Chinese	66,020	108,513	60.84%
Black African	329,840	379,453	86.93%
Black Caribbean	96,025	267,756	35.86%
Mixed Ethnicity	75,410	518,719	14.54%
No ethnicity found	272,355	0	–
Total	6,125,025	6,362,547	96.27%

As presented, the coverage rates vary significantly by ethnic groups – while the LCR seem to provide reliable estimates for White British and Asian Indian groups, Mixed Ethnicity and Black Caribbean groups are characterized by very low coverage rates.

3.2. Population synthesis

The population synthesis for two boroughs was conducted. Two models were developed – a baseline model and spatially enhanced model that borrows individuals from adjacent areas where an empty cells problem arises (i.e., an individual of certain set of characteristics is present in aggregated statistics, but is absent from the individual-level data). The individuals were reweighted and aggregated at the LSOA level to preserve confidentiality. To test the synthesis quality, Pearson's correlation and Root Mean Squared Error (RMSE) statistics were calculated. The zonal correlation coefficients between expected and observed estimates ranged from 0.998 to 1, indicating successful prediction.

The RMSE for the baseline model presented in **Figure 2** indicates that one LSOA is subject to a substantial empty cells problem. Implementation of the spatial model improves the synthesis and mitigates RMSE in all LSOAs, including the problematic one from 37.3 to 5.33.

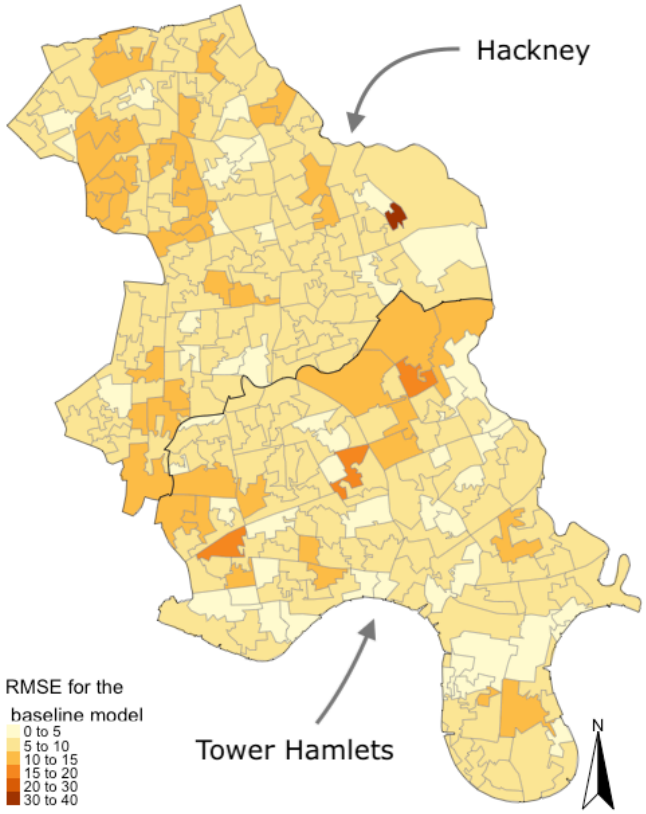

Figure 2 RMSE for the population synthesis baseline model.

4. Discussion

The proposed framework shows how data combination may benefit development of population estimates and geodemographic classifications. Spatial microsimulation techniques provide a useful approach for combining good aspects of different data, leading to an improvement in comprehensiveness and spatio-temporal granularity. Future research will focus on the improvement in the utilisation of EE and forename classification algorithms and providing updates to the synthetic population by inclusion subsequent LCR versions. Moreover, it is hoped that a potential provision of administrative data will advance depth of the hybrid segmentation.

5. Acknowledgements

This research was co-funded by the ESRC and Office for National Statistics. The Linked Consumer Registers data were provided by the Consumer Data Research Centre grant reference ES/L011840/1.

References

Coleman, D. (2013) The Twilight of the Census, *Population and Development Review*, 38, 334–351.

Deming, W. E. and Stephan, F. F. (1940) On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, *The Annals of Mathematical Statistics*, 11(4), 427–444.

Gale, C. G. *et al.* (2016) Creating the 2011 area classification for output areas (2011 OAC), *Journal of Spatial Information Science*, 12, 1–27.

Kandt, J. and Longley, P. A. (2018) Ethnicity estimation using family naming practices, *PLOS ONE*. Edited by F. Calafell, 13(8), 1-24.

Lansley, G., Li, W. and Longley, P. A. (2019) Creating a linked consumer register for granular demographic analysis, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4), 1587–1605.

Lansley, G. and Longley, P. (2016) Deriving age and gender from forenames for consumer analytics, *Journal of Retailing and Consumer Services*, 30, 271–278.

Lovelace, R. *et al.* (2015) Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique, *Journal of Artificial Societies and Social Simulation*, 18(2), 1-21.

Lovelace, R. and Dumont, M. (2016) *Spatial Microsimulation with R*. Boca Raton, FL: CRC Press, Taylor & Francis Group (The R series). Available at: https://spatial-microsim-book.robinlovelace.net (Accessed: 19 April 2020).

Biographies

Jakub Wyszomierski is a PhD student in Human Geography at University College London. In his research he studies ways in which combination of different data sources can facilitate development of national geodemographic classifications with a high temporal and spatial granularity.

Paul A. Longley is Professor of Geographic Information Science at UCL and Director of the UK Consumer Data Research Centre at UCL. His research focuses on the application of geographic information science, with a strong emphasis on the development and deployment of geo-temporal data infrastructures developed from Big or Open Data.

Christopher G. Gale is Head of Geospatial Analysis and Capability at the Office for National Statistics. He completed his PhD in Human Geography at University College London. His thesis focused on geodemographics and development of the 2011 Output Area Classification.