Developing a Data Infrastructure for Bespoke Demographic Analysis

Justin van Dijk [*1] and Paul A. Longley[†1]

[1]Department of Geography, University College London

February 12, 2021

Summary

This paper describes the steps involved in the creation of a UK-wide population register, covering the adult population from the start of 1997 to the end of 2020. We argue how this set of 'Linked Consumer Registers' will be the nucleus of a data infrastructure for bespoke demographic analysis that we are currently developing. We further appraise the applicability and value of this data infrastructure for empirical work within the social sciences, particularly in the context of modelling residential mobility and unpacking the relationship between socio-economic inequalities across ethnic groups and vulnerability to COVID-19.

KEYWORDS: consumer data; data linkage; electoral register; demographic analysis

1. Introduction

New big data offer large volumes of data at much greater spatial and temporal detail to what is currently available via traditionally sourced social datasets. Although of unknown provenance, administrative and consumer datasets capture a large share of the names and addresses of the UK adult population. By combining several of these data sources into a set of 'Linked Consumer Registers' (LCRs), a dataset can be created that can provide a detailed picture of the UK's population at a higher spatial and temporal granularity than available from any official source. The creation of a first version of the LCRs, covering the period from 1997 to 2016, was described in detail in Lansley *et al*. (2019). Here, we briefly describe the process of creating the second version of the LCRs, covering the period from 1997 to 2020. We subsequently argue how this new version will be the nucleus of a data infrastructure for bespoke demographic analysis that we are currently developing. We further appraise the applicability and value of this data infrastructure for empirical work within the social sciences, particularly in the context of modelling residential mobility and unpacking the relationship between socio-economic inequalities across ethnic groups and vulnerability to COVID-19.

2. Data and linkage

The data that comprise the LCRs are sourced from extensive databases of names at the address-level. The main source to feed into the LCRs are the public versions of the electoral register from 1997 until 2020. This is supplemented with consumer data from 2002 onwards to 2017 to capture those that opt-out or are not eligible to vote (see also Lansley *et al*. 2019). For the years 2018 to 2020, following the introduction of the General Data Protection Regulation (GDPR) in May 2018, the public version of the electoral register is the sole data source feeding into the LCRs.

The actual creation of the new LCRs involved two steps: address matching and register consolidation. In the first step, the address records of each of the component annual registers were cleaned and reformatted before being linked to the best available address frame (AddressBase Premium and the Royal Mail's Postcode Address File). Addresses within the same unit postcode were linked using newly designed and improved address matching procedure that used a combination of rule-based and fuzzy

[*] j.t.vandijk@ucl.ac.uk
[†] p.longley@ucl.ac.uk

matching. In the second step, all component registers were combined into one register and records further standardised, for instance, by trying to accommodate surname changes (e.g. following marriage or sex change) to filter out duplicated individuals. Apparent gaps in an individual's residence at an address were filled by using data from adjacent time periods. Similarly, records were imputed at addresses where no data were recorded by bringing forward records from previous years. The LCRs thus work as a moving average where the imputed records at the end of the time series are updated every time a new year of data is injected. The amalgamation of these 24 years of linked records presents a detailed individual-level dataset that captures the majority of the adult population in the United Kingdom. The final annual LCR counts are shown in **Table 1**.

Table 1 Total counts of the Linked Consumer Registers by year

Year	Input register [*raw data*]	LCRs [*filling gaps*]	LCRs [*record imputation*]
1997	45,466,638	45,031,283	48,646,768
1998	46,299,201	46,800,779	48,471,168
1999	46,616,530	47,290,286	48,612,622
2000	44,037,323	46,529,984	48,394,939
2001	43,713,671	45,614,907	47,460,072
2002	44,881,619	46,675,086	47,555,182
2003	42,733,269	45,373,626	46,496,694
2004	41,527,046	44,863,258	46,043,855
2005	37,573,888	43,528,123	44,477,186
2006	36,032,336	42,935,341	43,879,677
2007	36,556,222	43,795,832	44,421,423
2008	33,161,520	43,953,229	44,826,372
2009	42,203,205	45,769,026	46,885,489
2010	43,524,797	46,813,483	47,831,344
2011	41,235,002	44,483,423	46,595,641
2012	30,110,856	36,666,315	42,384,250
2013	31,794,812	35,001,341	41,642,885
2014	28,772,595	31,315,197	39,449,606
2015	23,839,713	26,338,953	36,650,078
2016	24,057,303	25,494,143	37,837,402
2017	2,497,930	17,572,466	35,140,147
2018	19,133,583	19,355,884	36,958,463
2019	18,552,832	18,585,659	36,838,354
2020	18,634,215	18,661,571	36,926,955

Due to a reduction in the volume of consumer data provided for 2011 and subsequent years, **Table 1** shows a decline in the number of raw data records post-2011. Comparison of the final LCR counts to official sources do suggest, however, that this decline in population coverage is consistent over space. **Figure 1**, for instance, shows the LCR population counts as the proportion of the number of individuals recorded in the UK Census (2011) and the ONS mid-year population estimates (2012-2019) (ONS 2020a). Please note that to improve legibility, the Local Authority District of the City of London has been excluded as it is a very clear outlier: where the overall mean ratio between the LCR counts and the mid-year population estimates for 2019 is 0.679, the ratio for the City of London is 1.3 (down from 2.5 in 2011). **Figure 2** shows the spatial distribution of these proportions for 2019 – the last year for which the ONS currently has released the mid-year population estimates.

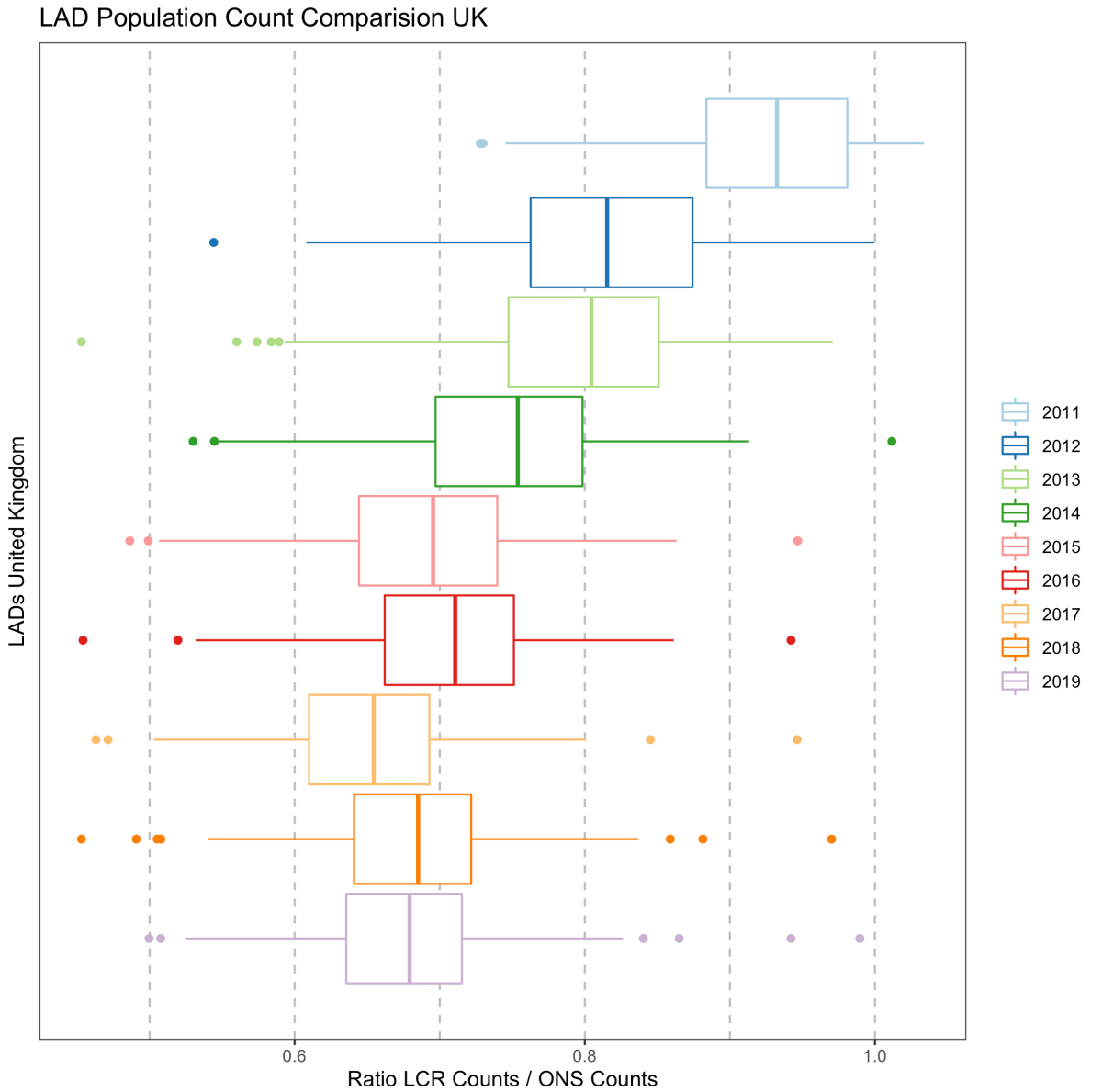

Figure 1 Ratio of the population counts captured in the Linked Consumer Registers and official population counts derived from the UK Census (2011) and mid-year population estimates (2012-2019) at Local Authority District level.

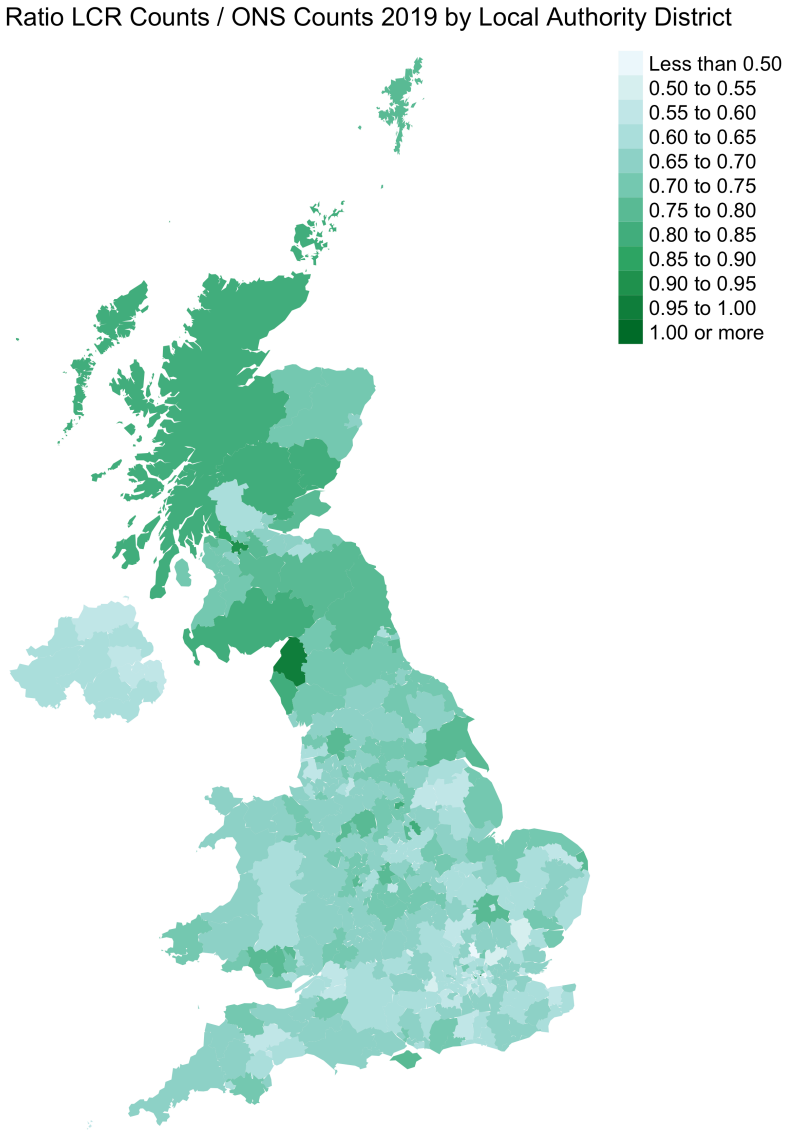

Figure 2 Ratio of the population counts captured in the Linked Consumer Registers and official mid-year population estimates at Local Authority District level for 2019.

3. Designing a Data Infrastructure

Although the LCRs contain only a few variables (i.e. forename, surname, geo-referenced address, and first and last year the individual was recorded at the address), the first version of the LCRs has been successfully used in a variety of studies on topics such as ethnic segregation (Lan *et al.* 2020) and, through indirect linkages to Historic Censuses of Population, intergenerational population change in Great Britain (Kandt *et al.* 2020). Following an improved address matching and register consolidation procedure, this second version of the LCRs will be used in a variety of projects too –mainly by creating a data infrastructure comprising of auxiliary datasets that are linked into the LCRs. Two current examples are: (a) modelling residential mobility; and (b) unpacking the relationship between socio-economic inequalities across ethnic groups vis-à-vis vulnerability to COVID-19.

1.1. Modelling Residential Mobility

The LCR data infrastructure has been used to successfully ascribe records pertaining to individuals that vacate a property to their most probable destination address (see Van Dijk *et al.*, Forthcoming). However, other available datasets are currently in the process of being linked into the LCRs infrastructure to further profile and characterise these movements. For instance, Zoopla rental listings data for individual properties will be used to apportion change to the rental market. Further linkage of Land Registry data (England and Wales) or Registers of Scotland price paid data will allow for the creation of a socio-spatial mobility index where transitions from the rental market into owner-occupation can be analysed. Other work involves enriching the LCRs by adding in housing characteristics from Domestic Energy Performance Certificates datasets.

1.2. Unpacking Inequalities along Ethnic Lines

The COVID-19 pandemic has painfully exposed extant socio-economic inequalities across ethnic groups across England and the United Kingdom as a whole: analysis carried out by the Office for National Statistics, for example, suggests that death rates for most ethnic minorities are higher compared to White ethnic group (ONS, 2020b). Since the 2011 Census of Population, however, the interaction between crowding, ethnicity and household structure has not been measured at neighbourhood scale, and there is no methodology for tracking changes over shorter time periods going forward from the 2021 Census. The LCRs allow, at least partially, to rectify these deficiencies through linkage of key administrative and consumer data sources in order to measure and chart changes in neighbourhood household structures and ethnicity concentrations. Notable in this context is that bespoke small-area ethnicity estimates, derived directly from the 1997-2020 LCRs, were supplied to the Joint Biosecurity Centre.

4. Conclusion

The ‘new’ version of the Linked Consumer Registers comprise a unique, scale-free, population-wide resource that will be of strategic importance to a number of policy concerns. The LCRs are therefore at the very core of a data infrastructure that is currently being developed and will allow the analysis of, amongst other things, (a) socio-spatial mobility; (b) small area characterisation of the origins and destinations of residential moves; and (c) the interaction between these two and household structure, measured in terms of household size and ethnic composition.

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’).

References

Kandt J, Van Dijk J and Longley P A (2020). Family name origins and inter-generational demographic change in Great Britain. *Annals of the American Association of Geographers*, 110(6), 1726-1742.

Lan T, Kandt J, and Longley P A (2020). Geographic scales of residential segregation in English cities. *Urban Geography*, 41(1), 103-123.

Lansley G, Wen L 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.

ONS (2020a). Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland. Accessed on: 10.02.2020.

ONS (2020b). Why have Black and South Asian people been hit hardest by COVID-19? Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/whyhaveblackandsouthasianpeoplebeenhithardestbycovid19/2020-12-14. Accessed on: 10/02/2020.

Van Dijk J, Lansley G and Longley P A (Forthcoming). Using linked consumer registers to estimate residential moves in the United Kingdom. Under review, copy available from the first author.

Biographies

Justin van Dijk is a Research Associate and Teaching Fellow in 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, socio-spatial inequalities, and geospatial data in general.

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.