Representing Population Dynamics from Administrative and Consumer Registers

Guy Lansley[*1], Wen Li[†1] and Paul Longley[‡1]

[1]Department of Geography, UCL

April 03, 2017

Summary

This research attempts to derive representative metrics of household dynamics and migration by analysing changes between two annual composite registers of the UK population. Through appropriate data cleaning and linkage techniques, it is possible to match addresses and record changes in their size and composition over a two year period. The paper also demonstrates that it is feasible to approximate migration trends by filtering and matching records of household units and individuals who are not recorded at the same address in both datasets.

KEYWORDS: migration, geodemographics, big data, population, consumer data

1. Introduction

There is an abundance of data on the population which are routinely collected by public and private organisations. However, the majority of such datasets are never analysed or repackaged as public datasets (Kitchen, 2014). Whilst the decennial census seeks to attain universal coverage, it provides very infrequent snapshots of population change. In this paper we show how data pooled from consumer and administrative sources may be used to create representative data on annual population changes at a small area level (Dugmore, 2010).

This paper summarises our efforts to date to model population dynamics from composite registers of the population. Two composite Consumer Registers were acquired from CACI Ltd (London, UK), and comprise the public version of the UK Electoral Register and individual records from a number of commercial organisations. The resulting databases contain the names and addresses for the vast majority of the adult population. By developing and applying appropriate heuristics, we demonstrate that it is possible to develop precise linkages at the levels of the individual and household, in order to estimate household change and even internal migration.

2. Data

Composite registers for 2013 and 2014 were acquired for this study. Both datasets represent a very high proportion of the adult population in the UK (Table 1). However, undoubtedly, the data will suffer from some uncertainty because the precise coverage achieved by the different data collection procedures is unknown. (e.g. see Bollier, 2010). Moreover, neither of the Consumer Registers are triangulated with a comprehensive address register, and therefore unique IDs do not span multiple data so it is not possible to easily establish longitudinal trends.

[*] g.lansley@ucl.ac.uk
[†] wen.li@ucl.ac.uk
[‡] p.longley@ucl.ac.uk

Table 1 Headline counts from the Consumer Registers

Year	Population	Households
2013	54,380,747	27,114,152
2014	55,397,463	27,371,755

In the absence of any additional variables, this study aims to model household change by linking addresses and recording changes in the composition of residents based on their names. We also explore the possibility of using registers as a means of estimating migration by identifying changes in unique compositions of names of residents arising when households change address.

Our first assumption is that an individual record which for two separate registers matches by address and full name, is the same individual and this person has not changed address between the two registers. Our second assumption is that if a composition of residents (as identified by their full names) are not recorded at the same address in the subsequent register, but that a household comprising identical named individuals has appeared at a single other address, then these records pertain to the same household.

3. Household change

To model changes in the adult population of the UK between 2013 and 2014 at the household level, addresses from both registers were joined using their full addresses and postcodes. 26,800,055 unique addresses appeared in both files, representing 98.8% of addresses recorded in the 2013 database.

Both the total number of members for each household and the individual matches between years have been recorded. Some households may maintain the same number of residents between years, but they could comprise entirely different people. Therefore, for each matched address, we recorded the number of persons who appeared in both years using the residents' full names. We also created a key to identify the minority of residents which may share full names within a household to ensure that each individual could be ascribed a unique ID for the name matching process.

Of course, this approach could not correctly account for individuals whose names may have changed or been misreported. The most common cause of name changes in households remaining at the same size will most likely be due to marriages, and we have therefore made efforts to identify recently married women from the registers. Whilst the data do include titles and the title "Mrs" does confirm that the female is married, the majority of females in the data are recorded as "Ms". Furthermore, a large number of titles are missing from the data and there are also gender neutral titles such as "Dr". Therefore, we used a forenames database outlined in Lansley and Longley (2016) to ascribe gender. We then created a filter which identified females whose forenames match within a household for both 2013 and 2014, but whose surnames do not. If the surname in 2014 matches the surname of a male occupant from the same household, we assume that this is the same individual and she has taken her husband's surname. 104,195 women met this criterion. This figure is reasonable given that roughly 120,000 women got married in 2013 (in England and Wales) and an unknown proportion of which did not change their surname or changed address (McLaren, 2013).

We also observed that persons with double-barrelled surnames were often inconsistently recorded within the data. To improve the match rate we also built a filter to identify persons may be recorded as having a double-barrelled surname in the database and just one of their singular surnames in the other year's data (using the joined household database). This model identified 23,075 individuals who were subsequently reassigned as stable residents.

The household matching analysis revealed that 73.7% of households in 2013 remained identical in composition in 2014. The results are presented in Table 2.

Table 2 Changing household characteristics, 2013-14

Household type	Number of households
Stable household	19,985,093
Complete change	3,202,241
Growth	1,552,735
Shrinkage	1,194,035
Unstable household[§]	865,951
Present in 2013 only	314,097
Present in 2014 only	571,700

The household level changes between 2013 and 2014 are visualised as an alluvial plot in Figure 1.

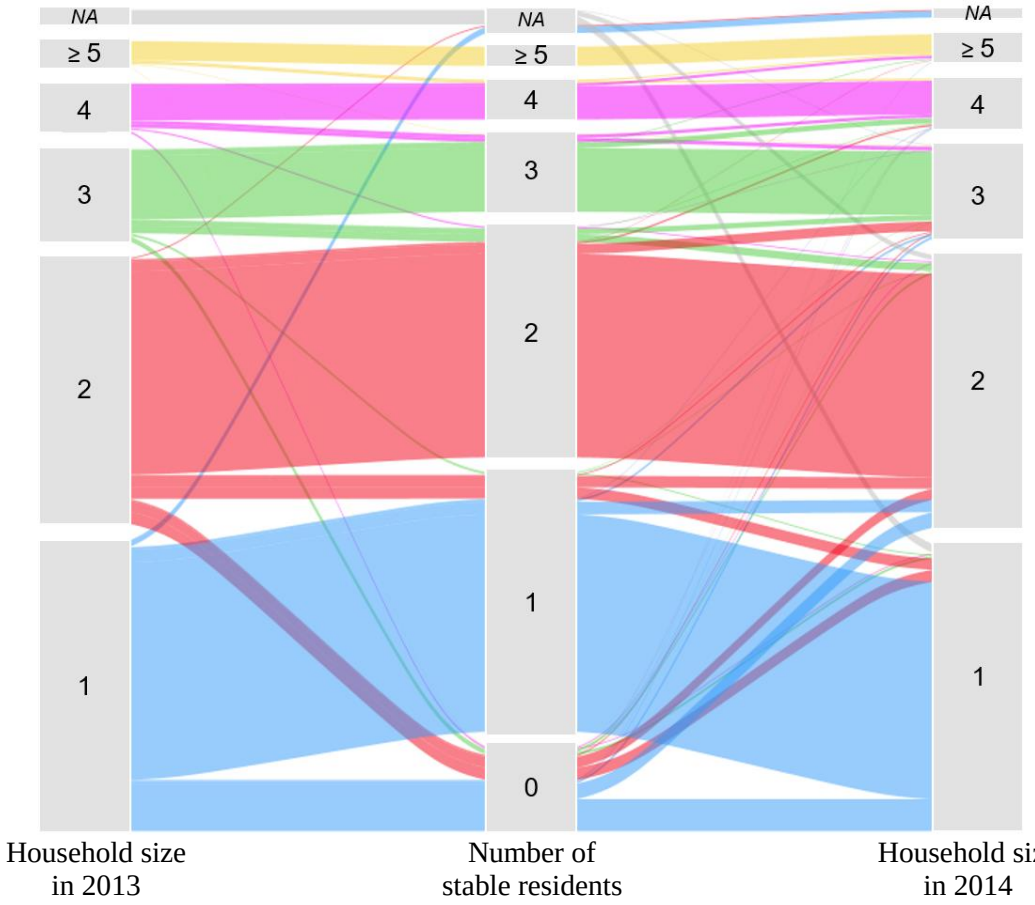

Figure 1 Alluvial plot of household changes between 2013 and 2014. *NA* refers to addresses that were not in one of the datasets

4. Estimating migration

Our second objective was to estimate migration flows from changes between the two registers. We first identified records which were not found at the same address in both registers. We were left with three databases which were labelled "Stable Residents", "2013 Movers" and "2014 Movers" (Table 3). Of course not all residents not defined by continued residence at the same address are movers: many may

[§] Unstable households refer to addresses which have remained the same household size, however, some of the residents have changed

have come of age or deceased, for example.

To provide some comparative context on annual population change, the 2011 Census confirmed that 7.5 million people changed address within the year prior to the Census, 6.8 million of whom had moved within the UK. It is also estimated that about 775,000 individuals turned 18 in 2013 (and would, therefore, enter the data). Pooled data from the ONS, NISRA and NRS revealed that in 2013 there were 576,968 deaths in 2013 in the UK. The Migration Observatory estimates that in 2013 526,000 persons moved to the UK, whilst 297,000 left (Vargas-Silva and Markaki, 2016). We would, therefore, expect between 8 and 9 million UK individual records to change each year.

Table 3 Residents that can be identified in both datasets

Description	Individuals
Stable residents	46,182,852
Present in 2013 and not 2014	8,197,895
Present in 2014 and not 2013	9,214,611

For each address within the 2013 movers database, we then created a unique key to represent the full composition of names within households in alphabetical order. This process was then repeated for the 2014 movers database. For this analysis, we only considered household compositions which were unique within both movers databases so that only definitive matches were possible. The vast majority of household name compositions were unique, yet only a proportion could be matched between 2013 and 2014. On-going research is aiming to assign duplicated household name keys into origin-destination pairs based on the heuristics identified from the modelled results from this paper and linkage to other data, such as price paid data from the Land Registry in England and Wales. We also searched for unique instances of full names within both movers datasets where household compositions could not be joined.

Despite only representing the first tier of data linkages of the wider research project, the basic model definitively identified the origin and destinations of over 769,000 individuals. In addition to these, over 80,000 households moved within the same postcode. We consider that many of these occurrences could have arisen due to properties changing name and therefore not address matching correctly.

Figure 2 visualises the key trends in internal migration in Great Britain at the local authority level as estimated by the consumer registers. The proportional symbols represent the number of moves which occurred within each district, whilst the lines represent moves between local authorities. Only popular flows of at least 40 migrants are shown. There are distinctive flows of migration between more populous districts and also many neighbouring local authorities. Most movers move relatively short distances, usually within the region. There was a median move distance of 21 miles.

Figure 2 A representation of internal migration within Great Britain. Migration within local authorities are displayed as proportional circles whilst flows of over 40 persons between districts are represented as lines.

5. Conclusions

This research has used advanced data mining and linkage techniques to estimate household dynamics and internal migration patterns from two annual consumer registers. Whilst this paper only presents the primary stages in a pipeline of matching procedures, future research will aim to increase the proportion of internal migration flows which can be successfully identified by considering data from additional registers and applying simulation techniques. We also seek to validate the representation of both the data and the modelled outputs in order to create uncertainty measures. In the absence of detailed and frequent indicators of population dynamics for intercensal years, it is hoped that the research could provide a framework for the integration and validation of consumer data products.

6. Acknowledgements

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1.

7. Biography

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

Wen Li is a Data Scientist at the UK Consumer Data Research Centre and the Department of Geography at UCL. His main research focuses on data integration by applying methodologies from information retrieval and distributed computing.

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 14 books and more than 150 refereed journal articles and book chapters. He is a former co-editor of the journal Environment and Planning B and a member of four other editorial boards.

References

Bollier, D. (2010). *The promise and peril of big data (p. 1)*. Washington, DC: Aspen Institute, Communications and Society Program.

Dugmore, K. (2010). *Information collected by Commercial Companies: What might be of value to Official Statistics? The case of the UK Office for National Statistics*. London: Demographic Decisions Ltd.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, London

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

McLaren, E. (2013) *Marriages in England and Wales: 2013*. Office for National Statistics, Statistical Bulletin.

Vargas-Silva, C. and Markaki, Y. (2016) *Briefing: Long-Term International Migration Flows to and from the UK*. The Migration Observatory, The University of Oxford. Online: http://www.migrationobservatory.ox.ac.uk/resources/briefings/long-term-international-migration-flows-to-and-from-the-uk/ (accessed: 05/01/2017)