Modelling Small Area Level Population Change from Administrative and Consumer Data

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

[1]Department of Geography, UCL

January 07, 2018

Summary

This research presents a longitudinal database of the UK adult population at the address level through linkage of administrative and consumer datasets released from 1998 to 2017. The analysis first devised heuristics to maximise the linkage of addresses between different annual datasets; then secondly, linked residents that occurred at the same addresses between years. In doing so, it was also possible to determine the duration of time that households have resided at given addresses. With the additional contribution of address-level open datasets, it was possible to build population churn estimates that could be released at a small area level.

KEYWORDS: big data, population, population change, geodemographics, data linkage

1. Introduction

This paper presents a unique highly granular longitudinal database of the adult population compiled from annual population registers released from 1998 to 2017. All of the registers include public versions of the electoral register and most of them have been supplemented by consumer datasets. Due to their high coverage, electoral data have historically been used to guide sampling in social sciences research (Hoinville et al., 1978). However, as the registers do not collect personal information beyond names and addresses, they have been overlooked as a source of useful geodemographic information.

The hypothesis for this study is that through bespoke data linkage techniques, it will be possible to determine the duration that households have resided at their addresses. In addition to being a useful social sciences dataset, the database could be used to clean alternative big datasets on the population or to provide a spine through which consumer data are linked. Indeed, with the withdrawal of the long-form census approaching, it is important that researchers maximise the opportunities presented by big datasets that are routinely collected (Dugmore, 2010; Anderson et al., 2016). However, as this research demonstrates, harnessing Big Data fundamentally transformed how representations of the population are devised.

2. Data

This research acquired the 20 registers from three different sources. Firstly, public Electoral Register records from 1998 to 2002. The electoral registers usually come into operation in February although the bulk of the data are collected in the preceding October. The Representation of the People Act 2000 introduced the ‘edited register’, which excludes those who requested to opt-out. In 2002 an opt-out option was provided in electoral registration forms so the proportion of adults that chose to omit themselves from the edited register may have risen drastically then (White and Horne, 2014). The

[*] g.lansley@ucl.ac.uk

[†] wen.li@ucl.ac.uk

[‡] p.longley@ucl.ac.uk

registers from 2003 onwards have been supplemented by consumer data.

The data for 2003 to 2012 were acquired from DataTalk Ltd. The data included a flag to indicate if each record was obtained from the edited electoral register or from anonymous commercial sources. It is understood that all of the commercial records included in each register were updated (or still considered present) within the previous 18 months of the data release (February). The registers for 2013 to 2017 were provided by CACI UK Ltd and also flag records that were not obtained from the edited electoral registers. These registers have been bolstered with legacy records (records that were collected in earlier years), and last seen dates are also provided.

Unfortunately, the data were not collected for population analytics, they, therefore, may not represent every adult accurately. While the edited electoral registers exclude certain groups (e.g. those not eligible to vote due to nationality), they also do not record those that have failed to register or have asked to opt-out (Electoral Commission, 2016). They also only include those of the voting age (18 in England), or due to become of age before the release of the subsequent register. Furthermore, no information on the supplementary data sources from 2003 onwards were disclosed due to commercial sensitivities.

Table 1 shows the population counts in each register and the portion of adults that were obtained from the electoral register. It also compares the number of records to the mid-year population estimates (MYPE) of persons aged 17 and over.

Table 1 The number of records in the electoral registers (1998-2002) and consumer registers (2003-17), the percentage of records from the electoral register and comparisons to mid-year population estimates (persons aged 17 and over).

Year	Individual Records	% Electoral Register	% of MYPE
1998	45,466,638	100.00	99.40
1999	46,299,201	100.00	100.76
2000	46,616,530	100.00	100.90
2001	44,037,323	100.00	94.73
2002	43,713,671	100.00	93.39
2003	44,881,619	76.04	95.26
2004	42,733,269	73.69	90.05
2005	41,527,046	72.50	86.61
2006	37,573,888	77.30	77.68
2007	36,032,336	76.69	73.79
2008	36,556,222	72.12	74.13
2009	33,161,520	75.04	66.70
2010	42,203,205	57.00	84.14
2011	43,524,797	55.78	85.96
2012	41,235,002	63.97	80.93
2013	54,380,747	41.48[§]	106.06
2014	55,397,463	55.78	106.33
2015	55,456,742	50.70	107.29
2016	54,969,038	42.55	104.65
2017	53,711,052	39.82	*NA*

Spatial distribution of records has been mapped in Figure 1. While the data may be reflective of the adult population to a large extent, variable data collection techniques between local authorities may have contributed to the uneven coverage between regions.

[§] There was no source flag in the data, therefore, the electoral role proportion has been estimated by acquiring all of the data that was entered in the October of the previous year.

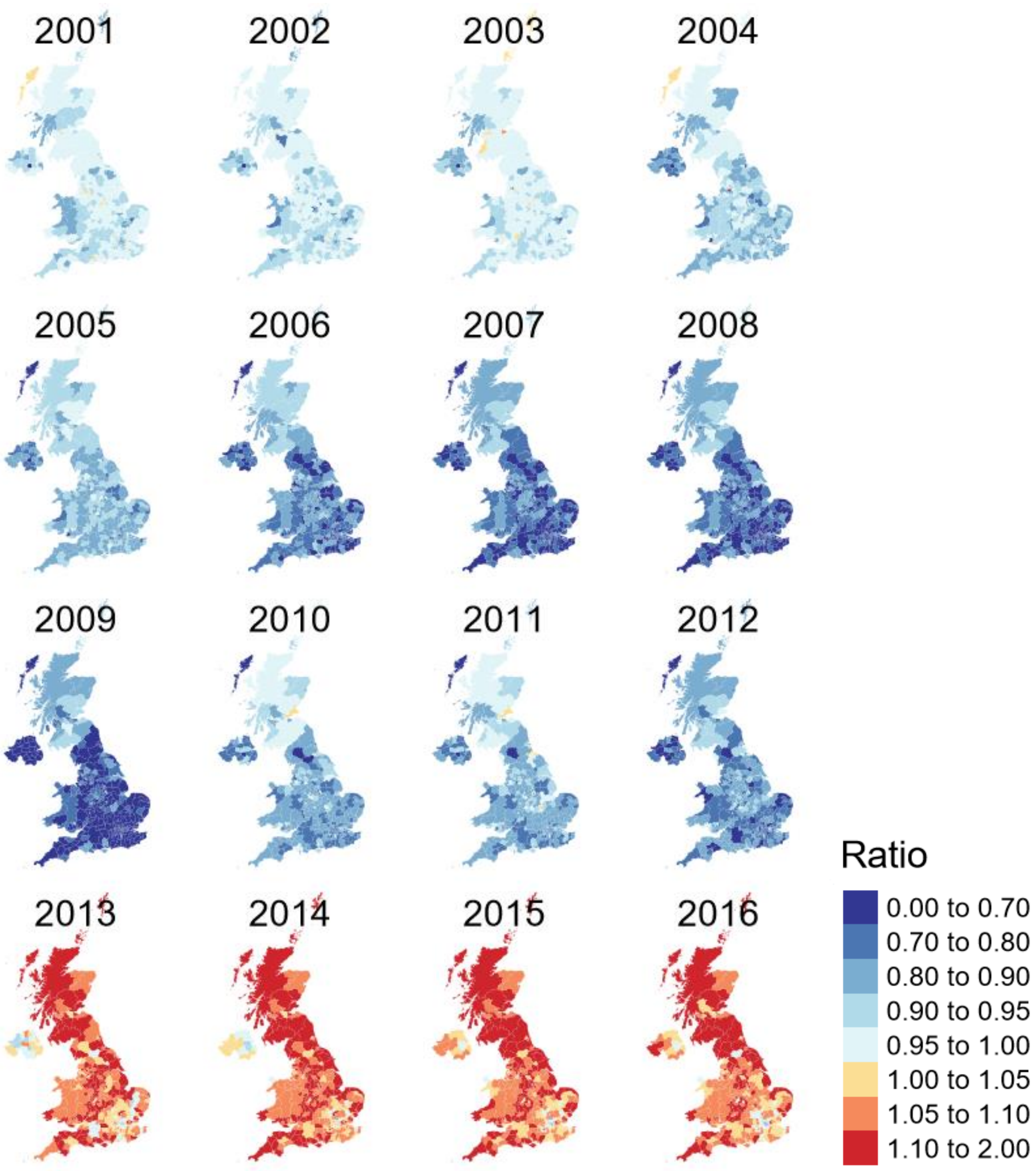


Figure 1 The ratio of records individual population registers (2001 to 2016) by the mid-year population estimates for each district

3. Address matching

The first challenge was to link all of the addresses into a consistent framework so that households could be analysed over time. Whilst UK addresses and the postcode system were developed in order to ensure that each address could be individually specified. Inconsistencies in formatting mean that many addresses may be recorded slightly differently between alternative sources. Therefore, we devised a novel address matching algorithm which attempted to link every address in the registers to the UK AddressBase (by Ordnance Survey). The 2016 AddressBase contains records for 28,581,702 residential addresses.

The aim of the algorithm was to match as many addresses to AddressBase as possible. Following a string match, three similarity functions were used to assign addresses that failed to match within each postcode respectively. The first one considered numbers within address strings. The second is based on the word difference between two addresses, where less common words had a higher weighting. The third approach is a variant of Levenshtein Distance (Edit Distance) which is a measure of the difference between two strings at the character level and emphasises the differences at the beginning of the address strings. Based on the three approaches, each address from each register was assigned to their most likely Unique Reference Number (URN) from AddressBase. The matching processes are demonstrated in Table 2. Over 26.7 million addresses could be matched, the remainder were given temporary URNs.

Table 2 A demonstration of the string matching process

Match type	Before	After
String	27 farm lane	27 farm lane
Number Based	2-21 queens road	flat 2 21 queens road
Number Based	flat d 79 forthbridge street	79d forthbridge street
Character-level Edit Distance	oaktree bishop road	oak tree bishop road
Word based distance	the farm cottage ham street	the farm ham street

4. Individual matching

Having established a means to link addresses to a common reference system it was then feasible to build a longitudinal database which recorded the presence of individuals at each address across all 20 registers. Following the removal of duplicates, 154,741,203 unique occurrences of persons at addresses were identified across all of the registers.

Indeed, a limitation of working with big data in social sciences is veracity. As such, an unknown proportion of adults are misrecorded or not recorded at all in each register. For instance, there are over 30 million individuals who were recorded as absent in years that occurred between registers where they were present (as demonstrated in Figure 2). In response, a data cleaning algorithm was applied to fill in the gaps for these cases.

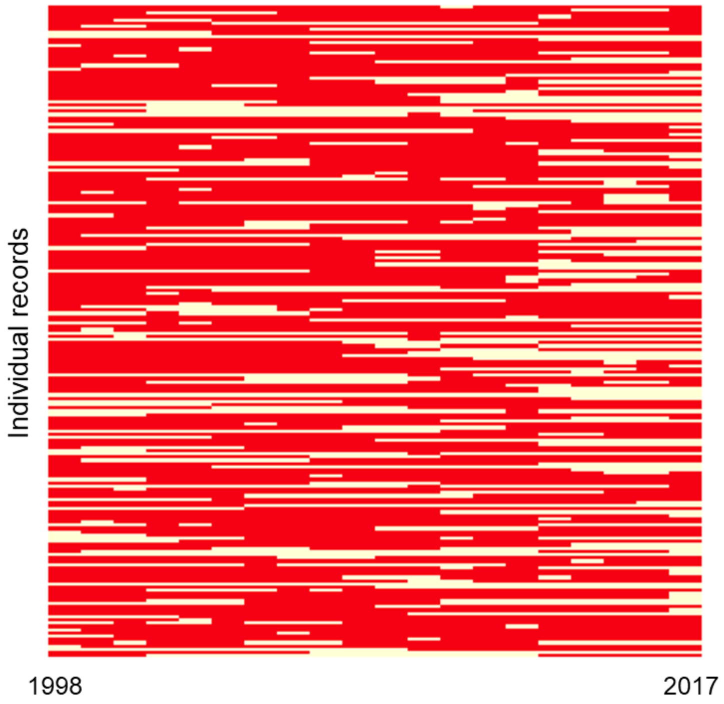


Figure 2 The occurrence of 200 randomly selected records across all of the registers

5. Churn

Having cleaned the longitudinal data, it was then feasible to identify when households moved in. This information was aggregated to create a churn index at the Lower Super Output Area (LSOA) level. The data were reassigned to the years of their data collection rather than data release. This entailed shifting the data collected from the electoral registers from 1998 to 2012 to the previous years to coincide with the October canvas.

Following this, two address level datasets for England and Wales were also considered: the occurrence of property sales from Land Registry price paid data (1995 onwards) and the occurrence of new rental transactions from Energy Performance Certificate (EPC) data (2008 onwards). Although unfortunately, the EPC data are of partial coverage. In addition, a filter was applied to identify active properties, these are addresses that matched address records from the 2016 AddressBase or had been observed at least once in the population registers since 2013. It identified 28,589,817 active addresses. The methodology is summarised in Figure 3.

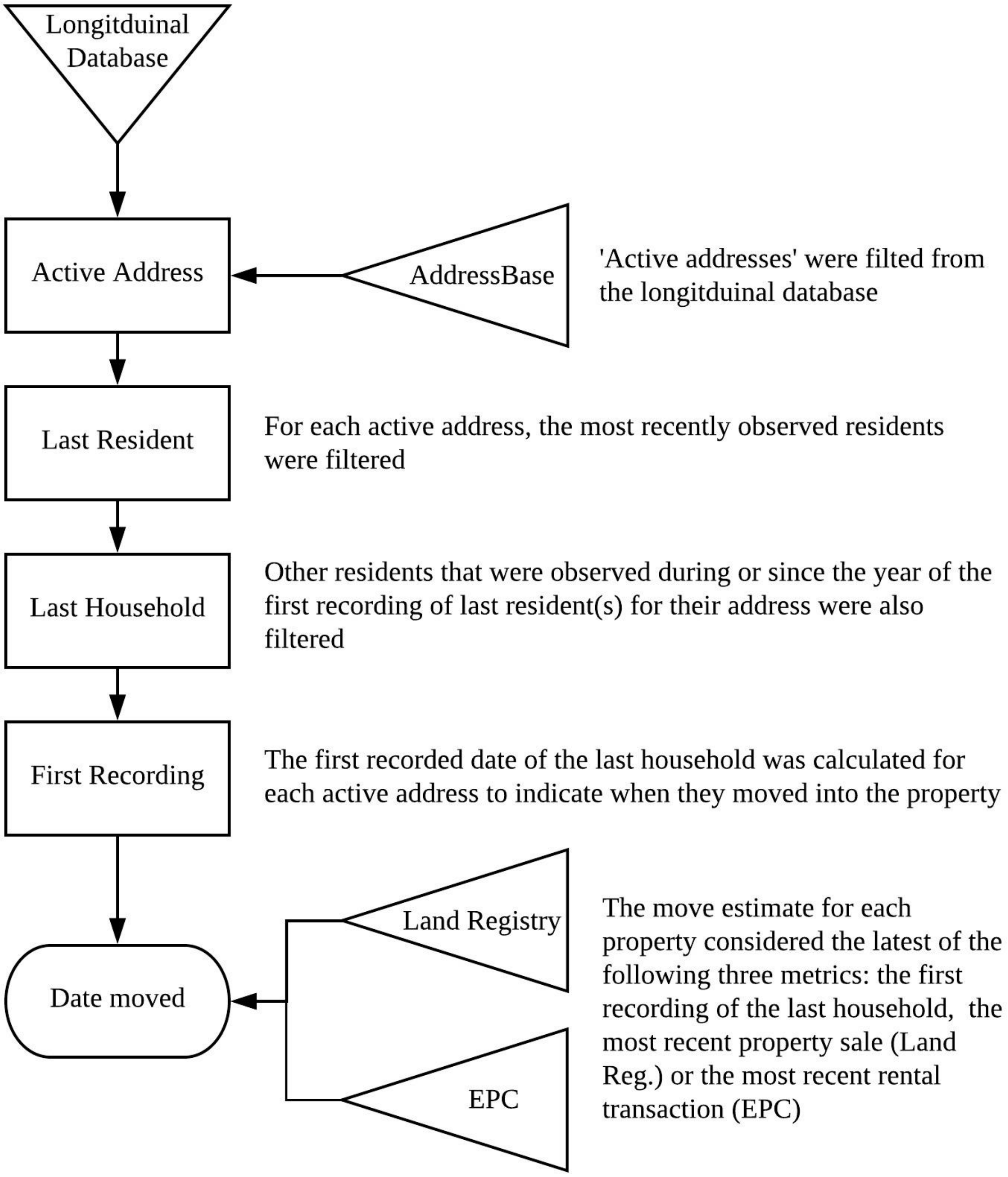

Longitduinal Database
Active Address
AddressBase
'Active addresses' were filted from the longitduinal database
Last Resident
For each active address, the most recently observed residents were filtered
Last Household
Other residents that were observed during or since the year of the first recording of last resident(s) for their address were also filtered
First Recording
The first recorded date of the last household was calculated for each active address to indicate when they moved into the property
Land Registry
EPC
Date moved
The move estimate for each property considered the latest of the following three metrics: the first recording of the last household, the most recent property sale (Land Reg.) or the most recent rental transaction (EPC)

Figure 3 Flow diagram of the methodology implemented to create the churn index

In addition, to the analysis described above, if only one household was ever detected at an address, then land registry data were tested to see if they could identify a property exchange before their first occurrence.

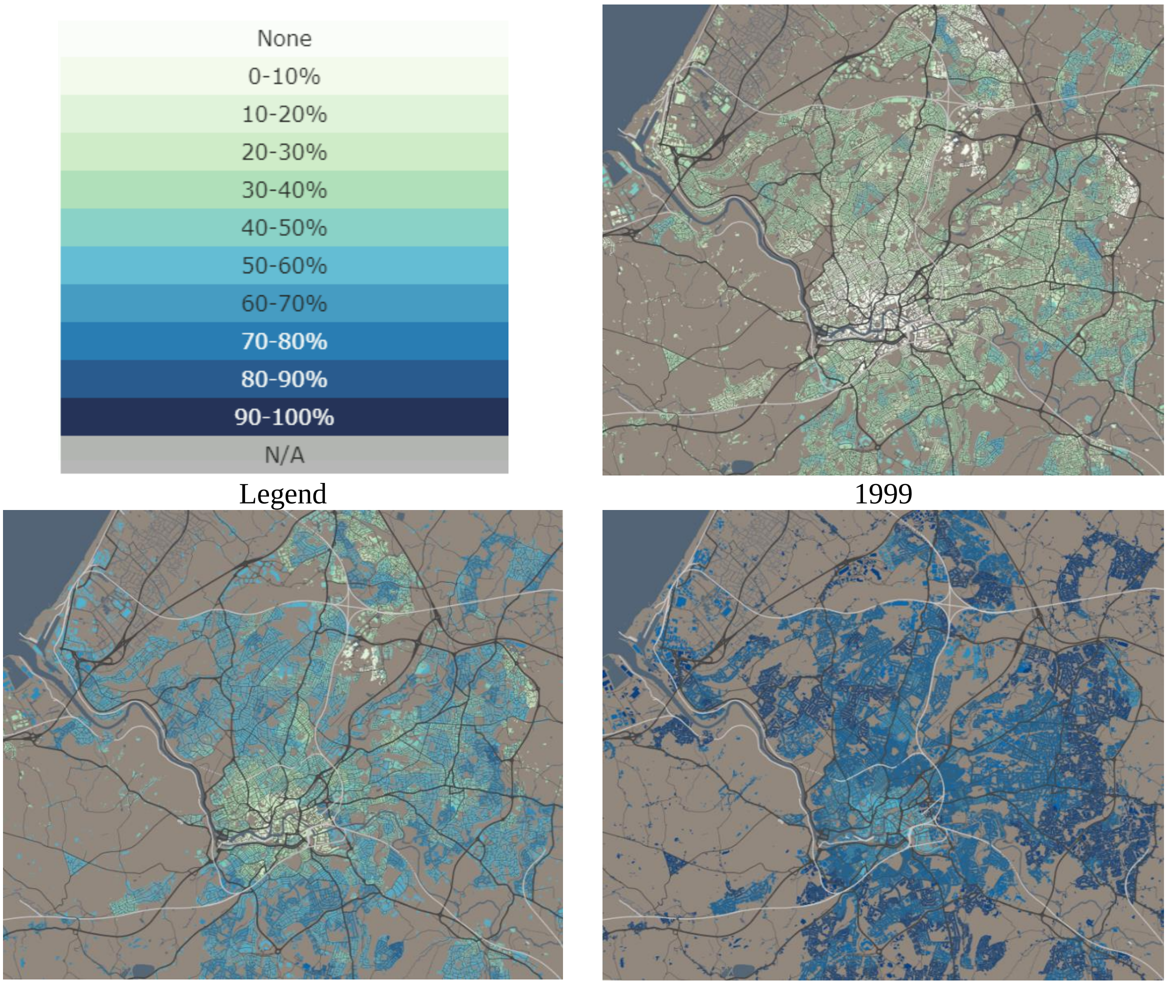
The frequencies of the first seen dates for the households in the churn database are displayed in Table 3.

Table 3 The frequency of household records by the year they were first observed

Year first seen	Number of households	Cumulative percentage
Before 1998	7,182,869	25.1%
1998	831,862	28.0%
1999	739,192	30.6%
2000	895,976	33.8%
2001	631,544	36.0%
2002	933,999	39.2%
2003	808,451	42.1%
2004	833,751	45.0%
2005	820,691	47.8%
2006	1,060,987	51.6%
2007	1,087,169	55.4%
2008	696,078	57.8%
2009	1,001,719	61.3%
2010	1,052,655	65.0%
2011	1,033,382	68.6%
2012	1,439,983	73.6%
2013	1,910,722	80.3%
2014	1,536,743	85.7%
2015	1,800,046	92.0%
2016/17	2,291,998	100.0%

The churn index estimates that roughly 25% of households had at least one member that had resided at the same address for at least 19 years. In addition, over 2.2 million households were first identified at addresses in the most recent population register. This figure could be indicative of the private rental sector where short tenancy contracts are common. There was a dip in the frequency of new households moving into addresses in 2008. This could be partly due to the recession which saw a considerable decrease in the number of properties sold. Although some of the fluxes could also be due to discrepancies in data collection over the years.

Figure 4 compares the spatial distribution of the proportion of households that had not changed address since 2014, 2009 and 1999 across the City of Bristol. Unsurprisingly areas in the centre of the City were found to have the most rapid turnover. Neighbourhoods nearest to the centres of large cities typically have high concentrations of young adults in rented accommodation. In contrast, the outer suburbs had the least household changes over 20 years. This is especially true for areas where home ownership is high.



Legend

1999

2009

2014

Figure 4 The 2016/17 churn index in Bristol for 1999, 2010 and 2015. This index represents the portion of present households in each Lower Super Output Area (LSOA) that have remained at the same address between 2017 and each of the aforementioned years.

6. Conclusions

This research has highlighted the importance of data linkage in order to maximise the value of Big Data for social sciences research. The great strength of this analysis is that it retains the individual person and household as the units of analysis, making it possible to devise scale-free representations of population trends such as household formation and dissolution that are otherwise unobservable. Furthermore, other research has found it possible to infer additional trends from population registers, including migration (Lansley et al., 2017) and ethnic segregation (see Mateos et al., 2011).

7. Acknowledgements

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

8. 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.

9. References

Anderson, B., Lin, S., Newing, A., Bahaj, A. and James, P. (2017). Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. *Computers, Environment and Urban Systems*, 63, 58-67.

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.

The Electoral Commission (2016). *The December 2015 electoral registers in Great Britain, Accuracy and completeness of the registers in Great Britain and the transition to Individual Electoral Registration*. The Electoral Commission Report, July 2016.

Hoinville, G. and Jowell, R. (1978). *Survey Research Practice*. Heinemann Educational Books, London

Lansley, G., Li, W. and Longley, P. (2017). Representing Population Dynamics from Administrative and Consumer Registers. *Proceedings of the 25th Conference on GIS Research UK (GISRUK)*, Manchester

Mateos P., Longley P. A. and O'Sullivan, D. (2011). Ethnicity and population structure in personal naming networks. *PLoS ONE*, 6(9) e22943; 1-12

White, I. and Horne, A. (2014). *Supply and sale of the electoral register.* House of Commons Library, SN/PC/01020.