Creating a Linked Consumer Register for Granular Demographic Analysis

Summary

A very large share of the adult population frequently assent to provide data on their place of residence to local governments and businesses when registering for or acquiring goods and services. When linked together, such data can provide highly granular inventories of local populations and their characteristics on far faster refresh cycles than conventional statistical sources. However, each of the constituent data sources is of largely unknown provenance. In this paper, we describe how careful curation, linkage and analysis of consumer and administrative data sources can resolve many questions of content and coverage; resulting in comprehensive, highly disaggregate and frequently updateable representations of population structure, along with reliable estimates of incompleteness and possible bias. We link 20 consecutive annual public UK Registers of Electors to a range of consumer data sources in order to create annual updates to a longitudinal profile of the adult residents of almost every domestic property. We illustrate the applicability and value of the resulting unique data resource through the derivation of an annual small area household change index. We also assess the prospects of other, related, data linkage projects.

Keywords: Consumer data; Electoral Register; Data linkage; Census; Residential mobility, population

1. Introduction

A large and increasing share of the Big Data collected about citizens in recent years has arisen through transactions between consumers and the (private and public sector) organisations which provide them with goods and services. Collectively, they can be referred to as consumer data, although they comprise a range of different forms and originate from a wide variety of sources (Longley et al 2018). They are best thought of as digital ‘exhaust’, in that they are essentially by-products of business or service delivery (Harford, 2014). These data can also be interpreted as digital footprints that can be repurposed to create precise indicators of population statistics to supplement those traditionally collected by government agencies. In other instances, such data can provide entirely new and novel insights on population activities and characteristics.

The wide penetration of Big Data collection procedures is today enticing researchers and statistics agencies to repurpose said data to describe the population at large. Indeed, the future of conventional long-form based Censuses is uncertain and several countries are considering ways in which conventional statistics might be supplemented using administrative records (ONS, 2018) or even commercial data (ONS, 2017a). The core limitation of traditional sources of population statistics is that datasets that aim to achieve near-complete coverage are costly to produce and are infrequently collected. New forms of data are attractive due to their volume, velocity and (often) readily availability, despite their disconnection from scientific sampling procedures or quality controls. Seen from this perspective, data driven approaches are motivated by the richer **content** and faster **refresh** of nascent Big Data sources, albeit at the expense of full population **coverage** issues or the basis to generalisation, inference and scientific replicability (Hand, 2018; Norman et al, 2017).

Thus, new forms of data are fundamentally changing empirical analysis in statistics, and indeed the practice of social science. Concerns have arisen that the epistemology of ‘data driven’ approaches to representing populations are unclear (Miller and Goodchild, 2015), and that data driven analytical methods may be unable to accommodate the poorly understood sources and operation of bias in Big Data. Few, if any, consumer data sources can approach completeness of coverage, not least because no consumer organisation has a monopoly of market share, and few if any goods or services are consumed by every member of any crisply-defined population (Lansley and Cheshire, 2018). Conventional survey research requires prespecification of the probability of selection of any member of a known and clearly defined population, a condition that is less easy to fulfil with many administrative sources, where some sub-groups may be under- or over-enumerated (ONS, 2017b). Furthermore, many administrative datasets are difficult to reconcile with one another in the absence of an over-arching address frame (Goerge and Lee, 2002). Further complications arise when individuals change address. However, none of these problems are insurmountable and the spirit of our own research is to develop and extend work that has been undertaken using administrative data in the context of consumer data research. We believe that this offers the prospects of improving the range of characteristics that can be assigned to individuals, and of improving the spatial and temporal granularity for which such data may be harvested (see also ONS, 2018).

Our specific goal is to re-use underexploited consumer data to construct an annually updated linked database of the residences of the individuals and households that make up the entire UK adult population. While the source and operation of bias in the component datasets is largely unknown, we develop and apply address matching and data linkage procedures to develop a consistent inventory of individual names and addresses for the period 1997-2016. Our motivation is to facilitate reliable annual estimation of the changing attributes of neighbourhoods and the characteristics of households that reside within them, and to better understand the social and spatial consequences of residential mobility. Most of our data sources are from commercial organisations, but we also include public versions of the Electoral Roll – which, although fulfilling administrative functions are also considered ‘consumer data’ as they facilitate choice of elected representatives and (since 2003) indicate consent by the named individuals to contact for unrelated purposes such as marketing.

Here we describe the creation of an individual-level Linked Consumer Register (LCR) which traces the residences of individuals between 1997 until 2016. This process required the initial assembly of multiple consumer data sources, as defined above, their reconciliation with an assured address framework for each year, and their subsequent linkage over time at the level of the individual. Procedures were developed to establish the provenance of the different sources. Through the amalgamation of 20 years of linked records, we present a detailed individual-level product and demonstrate how it can be used to infer a longitudinal perspective on the changing characteristics of the adult population. We conclude by speculating upon the implications of this work for empiricist approaches to social science in the Big Data era.

2. Consumer Data Sources

No comprehensive population register is collected for the UK population, although local authorities have a statutory obligation to make available annual lists of electors who have not opted out of inclusion, according to a published schedule of charges. Our constituent datasets each comprised lists of adult names and addresses, obtained with appropriate consents, along with dates upon which individuals were ‘last seen’ by each data collection agency. These dates were bunched around the deadline for filing voter registrations in the case of the public Electoral Rolls. The data were structured into annual time intervals for the 20-year period.

The full Electoral Register includes eligible electors for both parliamentary and local government elections. Prior to electors being given the option to opt out of inclusion from 2003 onwards, Electoral Registers were frequently used to frame social surveys and investigations (Hoinville and Jowell, 1978). The bulk of electoral registrations are compiled during a canvasing period in October each year and the public versions are usually made available by individual local authorities after the following February. Thus, the compiled registers generally represent the population of the preceding year.

The coverage of electorates in the public ‘edited’ register has gradually decreased since its introduction (White and Horne, 2014). By 2014, 14.5 million electors in England and Wales opted to be excluded from the edited register. This trend accelerated following the introduction of individual electoral registration in place of registration by a self-nominated head of household in 2014, when the number of opt-outs stood at 25 million, or 59% of electors. The opt-out rate varies considerably by local authority (see Figure 1), at least in part because of differences in the layout of voter registration materials.

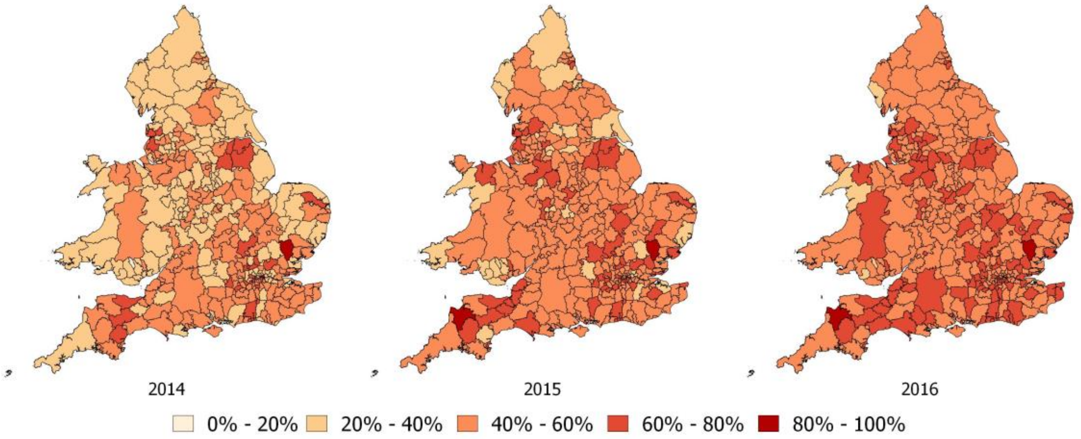


Figure 1. The proportion of electors that opted out of the Edited Electoral Register in 2014, 2015 and 2016 in England and Wales.

It has become the practice of value added data resellers to supplement the public Electoral Roll with additional consumer data, in order to enhance its value in marketing applications. The data in the research reported here were sourced from the composite ‘Consumer Registers’ for 2003-12 from DataTalk Ltd. (St Ives, UK), and for 2013-17 from CACI Ltd (London, UK). The identities of the providers of the consumer data enhancements are not revealed for commercial reasons, but each are identified by a separate flag in the files. Generally, the proportion of records in the Consumer Registers that were acquired from the consumer data files fluctuates between 20% and 40% for each dataset. The total numbers of records, and associated proportion obtained from the contemporaneous electoral register, are shown in Table 1.

Year	**Individual Records**	**% Electoral Register**
1998	45,466,638	100
1999	46,299,201	100
2000	46,616,530	100
2001	44,037,323	100

2002	43,713,671	100
2003	44,881,619	76.04
2004	42,733,269	73.69
2005	41,527,046	72.50
2006	37,573,888	77.30
2007	36,032,336	76.69
2008	36,556,222	72.12
2009	33,161,520	75.04
2010	42,203,205	57.00
2011	43,524,797	55.78
2012	41,235,002	63.97
2013	48,370,910	46.47 +
2014	54,283,557	57.19
2015	51,820,247	54.49
2016	51,387,463	45.98
2017	53,711,052	39.82

Table 1: The number of records in each Electoral Register (1998-2002) and each Consumer Register (2003-17), and the proportion of records which are derived from the most recent public version of the electoral register. (+ The percentage for 2013 is a minimum figure, because of some ambiguity on the flags provided for this year's data.)

Compared to DataTalk Ltd., the registers from CACI Ltd. include more individuals because records are carried forward to later years if no new information on the individuals resident at an address was collected. CACI flags indicate that the proportion of records collected within 12 months of each annual release incrementally but cumulatively declined in successive registers. The Consumer Registers do not conform to any address standard and so official address products from the Royal Mail, Ordnance Survey, and Land Registry were used to establish consistency. Each of the datasets are described in Table 2. The data on resident individuals and households is less reliable than that on addresses (Lynn and Taylor, 1995). AddressBase and address-level house sale data were not available for Northern Ireland.

Data Source	**Collection purpose**	**Likely Strengths**	**Likely Limitations**
Full electoral register 1998-2002	Enumeration of all named voters (age 17+) for all elections. This includes attainers, i.e. teenagers due to become eligible voters during the currency of the registers.	Legal requirement for completion (with minor caveats); includes Old and New Commonwealth citizens; includes Irish citizens and other EU citizens.	Underrepresentation of Commonwealth and EU citizens; double or undercounting of students and recent movers; possible double counting of some second homeowners. No non-voters.
Public version of Electoral Register 2003 - 2017	Enumeration of all named voters and those coming of	As above. Although the EU enlarged during this time	As above plus exclusion of 'opt out' individuals.

	voting age for any elections. In Scotland, attainers from the local government register are not included (as these can be as young as 15).	period. Scotland lowered the legal voting age to 16 in 2013 for local elections.	Variability in opt out rates from 24-60% over period 2003-17 (see Table 1).
Consumer files (2003-2017)	Provision and promotion of consumer goods and services.	Fills in many of those that ‘opt out’ of public version of electoral register and those ineligible to vote.	Unknown motivations for inclusion and consent; possible systematic bias for inclusion; non-standard address fields.
Land Registry records of domestic property transactions in England and Wales (1995-2017)	Payment of Stamp Duty and title registration.	All transactions recorded; very high correspondence with residential moves in owner-occupied sector; precise transaction dates.	Hard to differentiate minority of landlord transactions from majority of owner-occupier residential moves.
Registers of Scotland Sales for Considerations Data. Scotland (2003 – 2016)	As above.	As above.	As above.
Ordnance Survey AddressBase Premium 2018 (ABP)	Enumeration and location of residential addresses (including historic records dating back to 1990).	Near complete coverage of residential addresses, all address names have been consistently formatted and include a unique reference number (URN) which is used in other official products.	Not entirely complete or accurate. Great Britain only.
Postcode Address File 2016 (PAF)	Enumeration and approximate locations of residential addresses.	As above, but extends to the rest of the United Kingdom.	Not entirely complete or accurate. Contain some non-domestic records. Only made available for a single snapshot in 2016.

Table 2. A summary of the data components of the Linked Consumer Register (LCR).

Past research has suggested that the full electoral register under enumerates young adults and ethnic minorities (Lynn and Taylor, 1995). Private rental tenants and recent movers are also known to be under-enumerated (Electoral Commission, 2016). Unfortunately, less research has focused on the provenance of the edited version of the electoral registers, beyond rudimentary geographical analysis at local authority scale (Figure 1). Little is known about the quality of the consumer data files, beyond that they are supplied by four different suppliers in most Consumer Registers (2013-2017) and we anticipate that their compilation may be prone to errors (for example, data linkage errors) which could lead to duplications or removal of records. Prima facie, it is reasonable to anticipate similar lack of coverage of recent movers and migrants as the data involve address-based registrations, although non-voters are eligible for inclusion. Indeed, we anticipate that in blending multiple datasets of unknown provenance we may encounter issues of under-coverage (of hard to reach groups) and over-coverage (of those that might be duplicated because of changes of address or 2nd home ownership). There is a need to investigate such issues in future research, using methods promulgated in the Census Coverage Survey (Abbott, 2009).

3. Constructing the Linked Consumer Register (LCR)

A barrier to our core objective of establishing a reliable linked data product is that individuals, businesses or local authorities may use differing conventions for recording names and addresses. As such, it is often difficult to reconcile individual records, requiring the development of bespoke heuristics. The construction of the LCR required two core linkage exercises: the construction of a common address spine; and attribution of household composition (including assignment of houses in multiple occupation) to each address. Below we first describe the steps used to link records pertaining to the same address and de-duplify individual records using linkage to the address frames and fuzzy matching procedures. Second, we link individuals through matching names at each address and a series of steps which attempt to identify instances where individuals may have changed name or recorded a part of their name in a different way.

Given the personal nature of the data ethical review was sought and approved subject to conduct of the research in a safe researcher environment. This research only considers public and private sector datasets for which appropriate consents have been obtained by third party organisations. Our processing of the data falls under the public interest derogation for research under Article 89 of General Data Protection Regulation (GDPR). While formed from proprietary component data sources, the resulting LCR are available for bona fide research purposes upon successful application by accredited safe researchers to the UK Economic and Social Research Council Consumer Data Research Centre (CDRC: cdrc.ac.uk). Such access enables access to the code (written in Scala and SQL) that has been used to link the registers for different years. Furthermore, aggregated data products which have been run through disclosure controls, will be made available to the research community and public institutions to improve the availability of statistics for further research and end uses in providing public services.

3.1. Address matching

Across all of the registers for 1998-2017, 67.6 million unique address strings were recorded, more than twice the expected number of addresses. An initial exploration suggested that some unique addresses were composed in any of eight major variants. Table 3 identifies the nature of the address matching task over the 1998-2017 period.

Data sources	**Number of unique address strings**
Consumer and Electoral Registers, 1998-2017	67,582,896
AddressBase Premium 2018 (includes demolished addresses from 1990 and non-domestic addresses)	45,967,398
Postcode Address Files 2016	30,063,575
Land Registry (England and Wales), 1995-2017, cumulative total, property sales only	16,115,514
Registers of Scotland, 2003-2016, cumulative total, property sales only	1,562,488

Table 3. The number of unique address strings in each data source (2014 DCLG dwelling estimate: 28.1 million).

It was therefore necessary to standardise and consolidate the list of addresses from the diverse sources using AddressBase Premium (ABP) and the Postcode Address Files (PAF). These datasets each contain individual address records for Great Britain and the UK respectively, and both were used to establish consistent content, format and complete UK coverage.

Prior to matching, the addresses in the Consumer Registers needed to be cleaned and reformatted to remove inconsistencies. Common abbreviations (such as 'st.' or 'rd') were expanded to their full forms, and commonly used property partitions (such as 'gff', ground floor flat) were similarly expanded using a standardisation procedure. Changes in postcodes were accommodated using a Royal Mail update lookup table of 272,240 postcodes that changed between 1992 and 2006. Other possible duplicates were identified by filtering out multiple unit postcodes that shared precisely identical reference coordinates in the ONS Postcode Database (ONSPD).

Following this, we utilised three different approaches to address matching. At each stage, we attempted to reduce the number of unique addresses in the Consumer Registers. The procedures were designed to minimize false matches in favour of non-matches, as the latter could be picked up in a subsequent stage. They are briefly summarised below.

Rule-based Matching

Given that the Consumer Registers share no common standardised address format, and that any component may be inconsistently ordered or configured, we successively rearranged the address components in the ABP and PAF framework datasets to ascertain whether any would then directly match to the registers. Only matches within the same unit postcode identifier were considered. However, it is possible that selecting certain components of an address may incorrectly match some addresses. Therefore, matches that linked to multiple records in ABP or PAF were not amended.

Occupier-based Matching

We also took advantage of the data on residents to reduce the number of unique addresses in our database. Our assumption was that it is very unlikely that two properties within a postcode will share an identical composition of residents' names. Thus, we concatenated occupant names from addresses and searched for identical occurrences within the same postcode for each source register, and repeated the procedure for immediately succeeding years. Where duplicates were identified, they were merged into a single address (favouring the string that matches ABP or occurred most commonly).

Fuzzy Matching
It is also feasible that addresses may not match because of typographic errors. Therefore, we implemented a fuzzy matching procedure which was based upon three separate techniques:

1. Comparison of flat and address numbers to give an indication of the likelihood that they pertain to the same address. If the numbers did not match, the pairs were considered further.
2. Comparison of text strings using a word-bag approach to consider the difference in unique words used in the addresses. Common address words, such as 'road' and 'street' were assigned low weights in inverse proportion to their frequency of occurrence. This step also took into account use of common abbreviations.
3. Use of a variant of Levenshtein Distance (Edit Distance) of the difference between successive two character strings, with stronger weighting upon differences detected at the beginning of each address string – because the first address elements typically pertain to unique addresses and the later strings relate to aggregations such as districts or towns.

The three parts of the similarity function were linearly combined with tunable parameters to reduce false matches. The parameters were manually tuned following testing on small samples of the data.

Matching Stages
Each stage of the matching processes condensed the total number of addresses in our Consumer Registers by eliminating possible duplicates (see Table 4). However, it is of course extremely difficult to validate matching processes on data so vast, and some domestic residences may not be included on the PAF or ABP. The existence of inconsistencies between ABP and PAF highlights the difficulties in attaining universal coverage.

We estimate that c. 30 million residential addresses have existed in the UK over the 1998-2017 period. This estimate derives from the number active addresses in the PAF and ABP, recent dwelling stock estimates, the number of demolitions (1998-2016) and the number of conversions between 1998 and 2016. Our final list of addresses that occur in the consumer and electoral registers stands at just over 32 million entries (see Table 4). This overestimation is possibly a consequence of the fact that our databases include postal addresses that may have been altered overtime and are thus been duplicated, in addition, our data also include a very small proportion of non-domestic addresses. It is also feasible that some addresses may appear more than once because of the different formatting of individual records.

Steps	**Address Identities**
1. Rules based	37,976,018
2. Occupier matching	36,704,969

3. Fuzzy Matching	32,034,661

Table 4. Cumulative reduction in addresses at successive stages of the analysis.

Table 5 shows how each of the final 32 million unique addresses were identified and assigned a unique reference number (URN). The table also shows how each unique address in the property sales data for England and Wales and for Scotland may be linked to the final URNs. 94.9% of unique addresses where sales occurred could be linked to the unique addresses in the Consumer Registers. Almost all of these were linked to ABP, indicating that Land Registry transaction data are generally of better quality than those assembled for electoral registration or marketing to consumers.

Linkage	**Consumer Registers and Electoral Registers**		**Property Sales Data**	
	Unique Addresses	**Percentage**	**Unique Addresses**	**Percentage**
Link to ABP	28,019,531	87.47	13,872,557	99.78
Link to PAF but not ABP	842,479	2.63	3,063	0.02
Linked only to other CRs/ERs	2,927,970	9.14	24,371	0.18
Unmatched/Unique	244,681	0.76	3,428	0.02

Table 5. The reference frames for the URNs in the consolidated 1998-2017 database.

3.2. Resident matching

Having established a universal address spine for all of the registers, we were then able to link residents across the 20-year period. Individuals' names may differ between registers because of issues of marriage, name changes, alternative variants of spellings and misspellings. Therefore, an additional pipeline method was developed to improve the match rate of residents. In each step, the occurrence of each unique name at each unique address by year was recorded. It is rare in the UK, but conceivable, that a household may include multiple individuals that share the same name (e.g. junior and senior), although the Consumer Registers do not include minors. Implausibly high duplication of names occurred within addresses each year, averaging c. 440,000: duplicate names were thus flagged and merged. Prior to the analysis, empty spaces and punctuation were removed from the names (excluding hyphens which were used in subsequent steps).

3.2.1 Alternative versions of forenames

Apparent inconsistencies arise out of use of shortened or informal versions of forenames. We therefore developed a database of nicknames and their common name equivalents by recording the co-occurrence of forenames which commonly share both addresses and surnames. The assumption is that many individuals will record their different name variants over-time and therefore, within addresses, the two monikers they volunteer may share higher than expected rates of co-occurrence.

The most frequently co-occurring forenames were combinations of the most common, yet distinctive, names in the database – for example, almost 80,000 Margarets and Johns were observed to share both addresses and surnames, and were not of interest to this analysis. Instead, Table 6 shows the pairs of names that had the highest co-occurrence ratios, i.e. the frequency of a co-occurrence relative to the total frequency of the less common name of a pair. For instance, 80% of occurrences of the name “Stpehen” appear in the same household as an occurrence of the name “Stephen”. In this case, it is likely the former name is a misspelling.

Moniker	Forename	Pair Count	Moniker Count	Co-occurrence Ratio
stpehen	stephen	1,087	1,364	0.80
rober	robert	1,338	1,865	0.72
wiliam	william	2,406	3,513	0.68
gilian	gillian	1,370	2,193	0.62
magaret	margaret	1,131	1,829	0.62
patrica	patricia	3,750	6,762	0.55
valarie	valerie	1,834	3,563	0.51
malcom	malcolm	1,334	2,719	0.49
shiela	sheila	2,702	6,337	0.43
hillary	hilary	1,654	4,149	0.40

Table 6. The ten most frequent moniker-forename pairs with a frequency of 1000 or more.

In addition to common alternative spellings we also considered co-occurrences that differ in length by two or more characters to demonstrate shortened name variants (Table 7).

Moniker	Forename	Pair Count	Moniker count	Co-occurrence Ratio
liz	elizabeth	4,445	14,358	0.31
tasha	natasha	1,208	4,114	0.29
pat	patricia	5,039	19,397	0.26
val	valerie	1,074	4,158	0.26
pam	pamela	2,117	8,210	0.26
les	leslie	1,478	5,766	0.26
gill	gillian	2,857	12,066	0.24
sue	susan	9,084	38,961	0.23
jacqui	jacqueline	3,248	14,276	0.23
mick	michael	1,380	6,281	0.22

Table 7. The ten most frequent shortened moniker-forename pairs with a frequency of 1000 or more.

Thus, a moniker lookup table was produced for name-pairs with co-occurrence ratios above 0.05. This table was manually inspected to insure that no erroneous pairings were generated. Some monikers appeared to match multiple forenames and in such cases only the pairing with the highest score was retained. In addition, a handful of moniker-forenames were reversed to account for shortened names that could match two distinctive forenames, e.g. matching of ‘Steve’ matched both ‘Stephen’ and ‘Steven’. Some pairs were removed if they were clearly not variants of the same name, e.g. Kehinde and Taiwo had a co-occurrence ratio of 13.6%. Interestingly, these names originate from West Africa and are typically given to twins. In total, 1,253 unique monikers remained in the lookup table and over 680,000 records were subsequently cleaned.

3.2.2 Initialisms

In total, almost 3.65 million records in the database provided initials instead of forenames. As they could hamper linkage when used inconsistently, we sought to link initials to other forenames that shared the same surname and commenced with the same letter. Where an initial could be linked to two or more other records on this basis, the flag identifying the source of the data was used to prioritise linkage of data from different providers, after this priority was given to pairings that occurred across the most similar time period. A total, 1.68 million duplicates were identified and merged in this step.

3.2.3 Double-barrelled names

We also anticipated that some individuals may use both double-barrelled and their single surname components. We created a filter to identify cases where a forename occurred twice within a household: once with a double-barrelled surname, and once with just one of the components of the double-barrelled name. These records were then merged and the shorter name was retained. In total, 1.76 million records contained hyphens. Although only approximately 200,000 of them could be linked via the described method, as the majority of double-barrelled name bearers reported their surnames consistently. Following this stage, hyphens were removed and the matches were rerun.

3.2.4 Surname changes

Although name changes are generally rare, many women take their husband's surname upon marriage. It is estimated that between 1998-2015 there were almost 290,000 marriages a year in the UK. An algorithm was developed to identify probable female surname changes following marriage. These records were de-duplicated, and flagged with both the maiden and married names in all databases as they might be useful for future linkage work.

The following additional steps were undertaken.

1) Gender was ascribed through linkage to a forename database of probable genders (see Lansley and Longley, 2016). The source database was compiled from over 10 million records from birth certificates and consumer data files. 94% of records in the LCR were assigned a probable gender at this stage of the analysis.
2) A flag was created to identify female forenames that appeared multiple times (but with different surnames) at an address in our linked database.
3) A second search identified if one of the female surnames was shared with a male at the same address.
4) Married women were then identified where a probable female with a duplicated forename bore a family surname unless:
    a. The female without the family name was first recorded after the first recording of the individual with the family name.
    b. The address contained a large (over 35) number of individuals.

In total, 1,969,411 probable married women with changed names were identified in the database.

3.2.5 Fuzzy matching

Despite the above steps, many misspelled names may be retained in the database so a fuzzy matching procedure was implemented. The Soundex fuzzy matching technique is based on phonetics as pronounced in English and was devised for matching names (Stanier, 1990). It produces Soundex codes based on homophones which can be used to group words that sound the same but are spelt slightly differently. A soundex code was assigned to each name that remained in the database, although no changes were made to names that were matched at an address but nevertheless probably had different genders (e.g. Jean and John, Michelle and Michael). Where a match occurred, the most common name was retained. The process was run separately for forenames and surnames.

A summary of the number of unique records in the data following each step is shown in table 8. For each individual, we retained a flag to indicate what stage of the analysis they were linked as a measure of uncertainty.

Process	**Number of unique records**
Joining all registers	154,514,095
Text cleaning	150,031,561
Monikers	149,349,785
Initialisms	147,666,541
Double-barrelled names	147,485,472
Marriages	145,516,061
Fuzzy matching	143,789,049

Table 8. The cumulative reduction in the number of unique names at unique addresses, at successive stages of the analysis.

3.3. Identifying missing records

The amalgamation of data from numerous different sources for each year may cause many persons to appear in and drop out of address records in successive registers. Thus, the final section of the data cleaning attempted to impute data where records were thought to be missing. The primary means of doing this was by identifying gaps in an individual's apparent residence at an address, and then using data from adjacent time periods to fill in the gaps. While it is possible that some people may indeed vacate a property and then return (e.g. university students), inspection of the datasets suggested that the vast majority of the gaps were from incomplete temporal records. When blending the registers, we aligned the records to the years when they were most probably collected. Thus most of the registers were timestamped as pertaining to the immediate previous year to account for their autumn collection dates unless specific 'last seen' dates were provided.

The populations that were attributed to the registers for each year following the linkage exercise are included in Table 9. By monitoring residence over the entire study period, we are able to boost the number of individuals allocated to addresses throughout the intermediate years of our study. Only very small numbers of adults were supplemented to the earlier years largely because the coverage from the full electoral register data was very high. However, unfortunately, this approach is less effective at supplementing records during the later years as the number of new records diminishes. Other consumer data providers are available, however, and in future work we plan to address this issue.

Efforts were made to simulate the missing records at addresses where no data were collected. This was achieved by bringing forward records from previous registers for active properties that were missing data in 2016. Where active properties are considered as those identified as in use in the 2016 AddressBase. This approach is viable as the vast majority of adults recorded at a property in a particular year are likely to remain at their address during the following year. Unadjusted LCR records suggest that 95% of residents spend more than one year at their address. Between 2001 and 2011, an average of 11% of individual LCR records terminated in each year, indicating that the adult has either changed address or deceased. This amounts to a very modest apparent under-enumeration relative to the figures from the 2011 ONS estimates of 12.2%. The ONS figure includes all international emigrants, internal migration and death statistics. We seek to accommodate vacancies between known residences by assuming that the most of such instances arise from failures in data capture. The Electoral Commission (2016) identified that a minority of elector records are correctly updated within a year of a change of address. Thus for properties that appear to be vacant, we backdated incoming households by up to two years and, in a small number of instances also roll forward the outgoing household to fill the remaining void. Although gaps of over 2 years occurred only for a very small minority of properties. These records were flagged to signal a measure of uncertainty about properties that may well have been vacant.

We also attempted to identify changes at properties that occurred since the last evidence (if any) that a property had been occupied. Property sales data were used to identify households that have probably vacated and anonymous residents were imputed for vacant dwellings. The specific number of anonymous residents for each property was based on the median number of residents per year as recorded in the earlier data. Finally, there were some new build properties recorded as in-use in AddressBase but had no recorded occupants in the LCR. These properties were allocated two notional adult residents. Transitions in the rental sector can be modelled using historic lettings records from companies such as Zoopla or the Tenancy Deposit Scheme, though this was not available for this analysis. This is unfortunate, given that these households have a higher residential churn rate than owner-occupied households.

The final counts in the LCR are shown in Table 9.

Year	**Frequency of records seen following record linkage**	**Number of records from enhanced households**	**Number of anonymised records imputed from house sales data**	**Final counts in the LCR**
1997	45,128,532	0	0	45,128,532
1998	46,973,618	8,446	411	46,982,475
1999	47,365,943	15,023	2,754	47,383,720
2000	47,172,883	37,320	14,505	47,224,708
2001	45,717,253	368,915	24,900	46,111,068
2002	47,167,988	81,713	34,605	47,284,306
2003	46,157,405	112,111	57,307	46,326,823
2004	45,634,333	862,727	77,875	46,574,935
2005	44,604,330	1,546,450	96,549	46,247,329
2006	44,131,069	2,355,005	128,712	46,614,786
2007	44,835,431	2,302,915	162,632	47,300,978
2008	45,008,081	2,603,161	187,080	47,798,322
2009	46,875,665	2,175,449	217,888	49,269,002
2010	47,909,994	2,506,732	270,567	50,687,293
2011	45,908,413	3,830,296	396,688	50,135,397
2012	36,714,366	12,512,141	609,189	49,835,696
2013	35,619,824	14,709,334	937,793	51,266,951
2014	31,260,608	19,552,928	1,433,729	52,247,265

2015	27,319,898	23,081,685	2,212,712	52,614,295
2016	25,732,822	24,630,985	3,156,526	53,520,333

Table 9. The number of records in the final version of the UK LCR by year.

The final LCR data represent the vast majority of the adult population for every year between 1997 and 2016. It comprises data on almost every single active property for every year, although the size of the overall counts waned over the 20 year period. This is largely a consequence of incomplete households being recorded at many addresses. While it was possible to enhance the data for many households where there were some relevant data, we did not have a means of imputing residents that were completely missing. In addition, some records were also missing where properties did not appear in our data for a large number of years. Figure 2 compares the number of records in our analysis with UK mid-year population estimates. On average, the total number of imputed records in the LCR for each year is only 1.8% different to the mid-year estimates from 1997 to 2016. However, it was observed that following the analysis the final years of the LCR very slightly overestimate the adult population. This is probably because of the unknown size of new build properties (which were simply ascribed the rounded mean household size for the UK), and because some duplicated addresses may have been retained despite our best efforts. The slight overestimation of the adult population between 1998 and 2000 likely arises because of those registered at multiple addresses, or possible lags in data entries. Nevertheless, it is also worth remembering that official mid-year population estimates are approximate calculations (Rees et al, 2004).

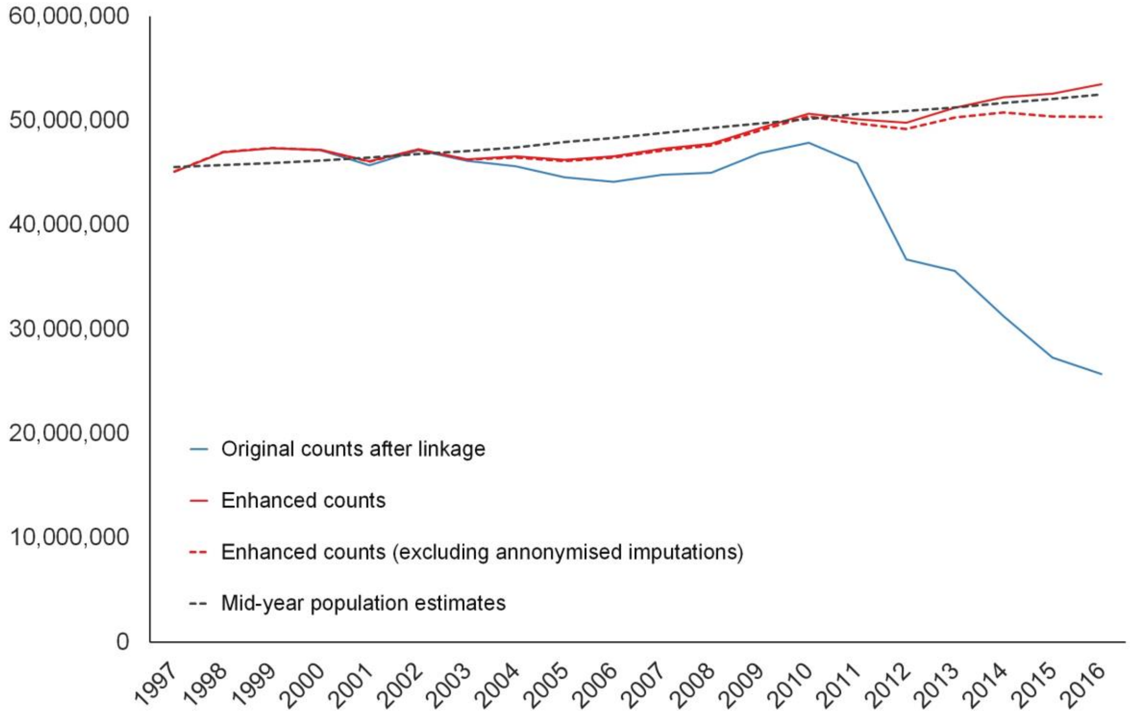


Figure 2. The number of individual records in the UK LCR compared to mid-year population estimates.

We also observed that for each year, the counts in the LCR corresponded very closely with mid-year population estimates at the district level. The correlation coefficients for every year were over 0.99. The coefficients in the most recent three years were the lowest; however, mid-year population estimates are believed to deviate further from the true totals as the elapsed time since the most recent census increases.

4. Application: Residential mobility and neighbourhood change

The LCR offer several opportunities for the investigation of neighbourhood characteristics at any convenient spatial scale, and with annual temporal refresh. A core limitation of censuses is that their infrequent collection makes it impossible to monitor rapid population changes. Thus, for example, Short (1978) argues that Electoral Registers could be an invaluable tool for understanding population turnover because of their annual refresh and high coverage. With the advent of opt out provisions for electoral registers, and with the advent of data handling technologies that allow integration of other consumer data sources for which consents have been obtained, the LCR can be considered the natural successor to full public UK Electoral Registers. Following Short (1978) and Marshall (1971), we use recurrence of adults at addresses as a means of estimating population turnover across space (see also Clark and Coulter, 2015). Using the LCR series, we are able to develop estimates of annual neighbourhood turnover (or ‘churn’) for the entire UK settlement system, and identify individual addresses where the occupants have changed on an annual basis. Thus we are able to identify areas that have undergone considerable change.

In our analysis, we have attempted to pinpoint the year in which each household joined and vacated an address. In this application we have investigated neighbourhood change using the years in which the most recently identified households at each address first joined their properties. If all household members did not join an address in the same year, then we considered the first seen date of the earliest household member. Household members were defined as all residents estimated to be present in 2016. The specific dates were refined using property sales records and aggregated to Lower Super Output Areas (LSOAs) of between 400 and 1,200 households. The equivalent small area units were used for Scotland (Data Zones) and Northern Ireland (Super Output Areas). The resulting Household Change Index (HCI) records the proportion of active addresses that have changed in occupation completely between 2016 and each of the preceding years. Active addresses were identified as properties that were recorded as in use in ABP, were recorded in the 2016 PAF, or were recorded the Consumer Registers after 2010.

This approach enables us to hone the dates for when an entire set of house members ‘refresh’. As such, the analysis is limited to addresses rather than household units. Houses in multiple occupation (HMOs) may exhibit partial rather than collective transitions, and in such instances, our approach records the year in which the first 2016 resident moved into the property. The cumulative frequency of the ‘first seen’ dates for the households in the HCI are shown in Table 12.

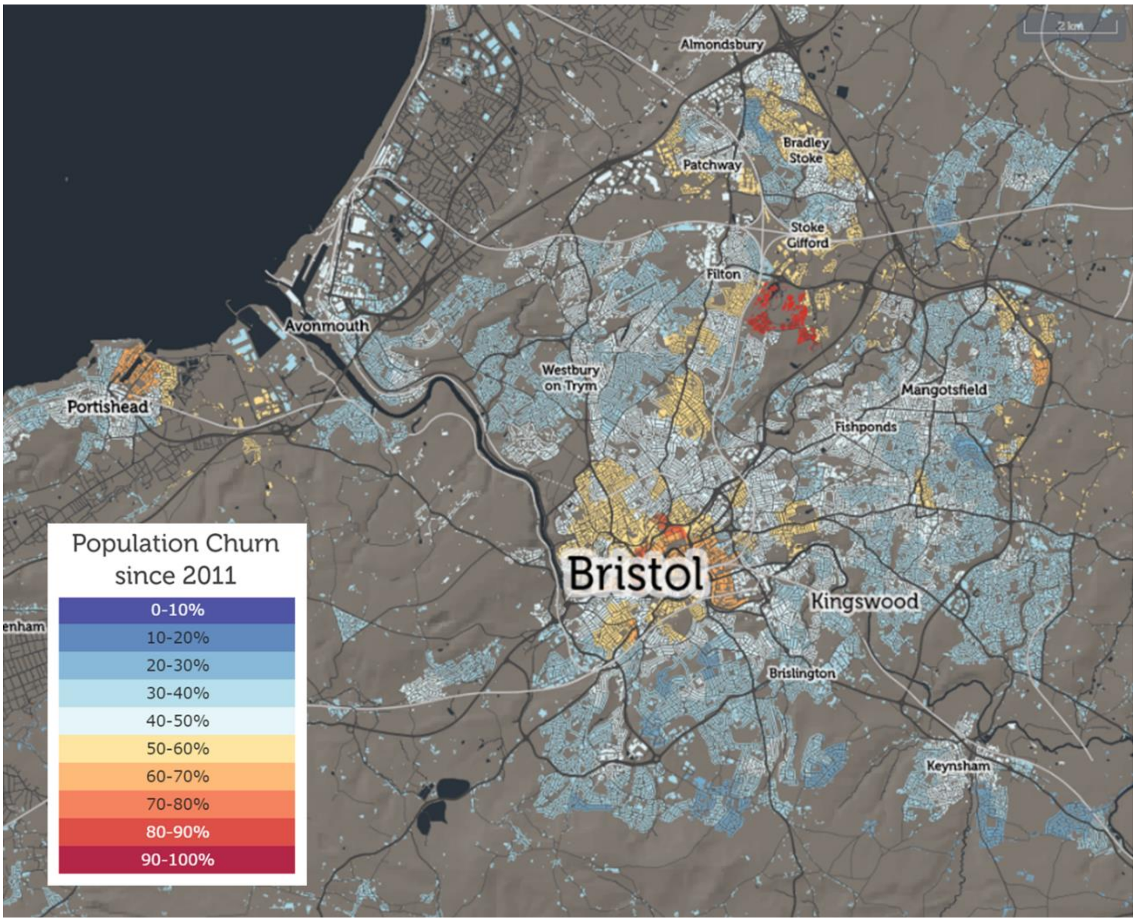
Year first seen	Frequency of households	Cumulative Percentage
1997 and before	7,839,962	100.00%
1998	795,695	73.53%
1999	639,182	70.85%
2000	710,924	68.69%
2001	517,490	66.29%

2002	859,346	64.54%
2003	672,635	61.64%
2004	645,499	59.37%
2005	673,778	57.19%
2006	793,510	54.92%
2007	823,548	52.24%
2008	621,289	49.46%
2009	1,223,087	47.36%
2010	1,424,420	43.23%
2011	1,256,909	38.42%
2012	1,656,489	34.18%
2013	2,367,456	28.59%
2014	1,394,258	20.60%
2015	2,010,633	15.89%
2016 and after	2,695,846	9.10%

Table 12: The years in which the last recorded households in properties extant in 2016 joined their present address.

The change index estimates that 34.18% of households in 2016 had moved to their current address in the period since the last available (2011) Census data on residential moves. In addition, just over 38% of properties have changed since 2011 (when the most recent Census was recorded). A large share of these households are likely to be from the private rental sector where short-term tenancy agreements are common. There is a dip in the frequency of households joining addresses in 2008. This might reflect the effects of the financial crisis in that year which resulted in a sharp decrease in property sales.

The LSOA level index reveals that, in general, central urban areas have experienced the greatest population turnover, especially during the last five years. Neighbourhoods that are known to have young and cosmopolitan populations with high proportions of properties in the private rental sector have experienced the greatest rate of change. For example, Figure 3A below shows the proportion of change since 2011 across Bristol. The central parts of the city experienced the highest rates of change; although areas where there have been extensive new residential developments also obviously experienced change. A large area of developments has occurred near the University of the West of England campus at Filton; here 84% of households have moved in since 2011. In addition, Figure 3B shows that substantial changes have occurred in Portishead since 2001, following the redevelopment of the marina and the construction of a large number of properties to the east of the town.


2 km
Almondsbury
Bradley Stoke
Patchway
Stoke Gifford
Filton
Avonmouth
Portishead
Westbury on Trym
Mangotsfield
Fishponds
Bristol
Kingswood
Brislington
Keynsham
enham
Population Churn since 2011
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
90-100%

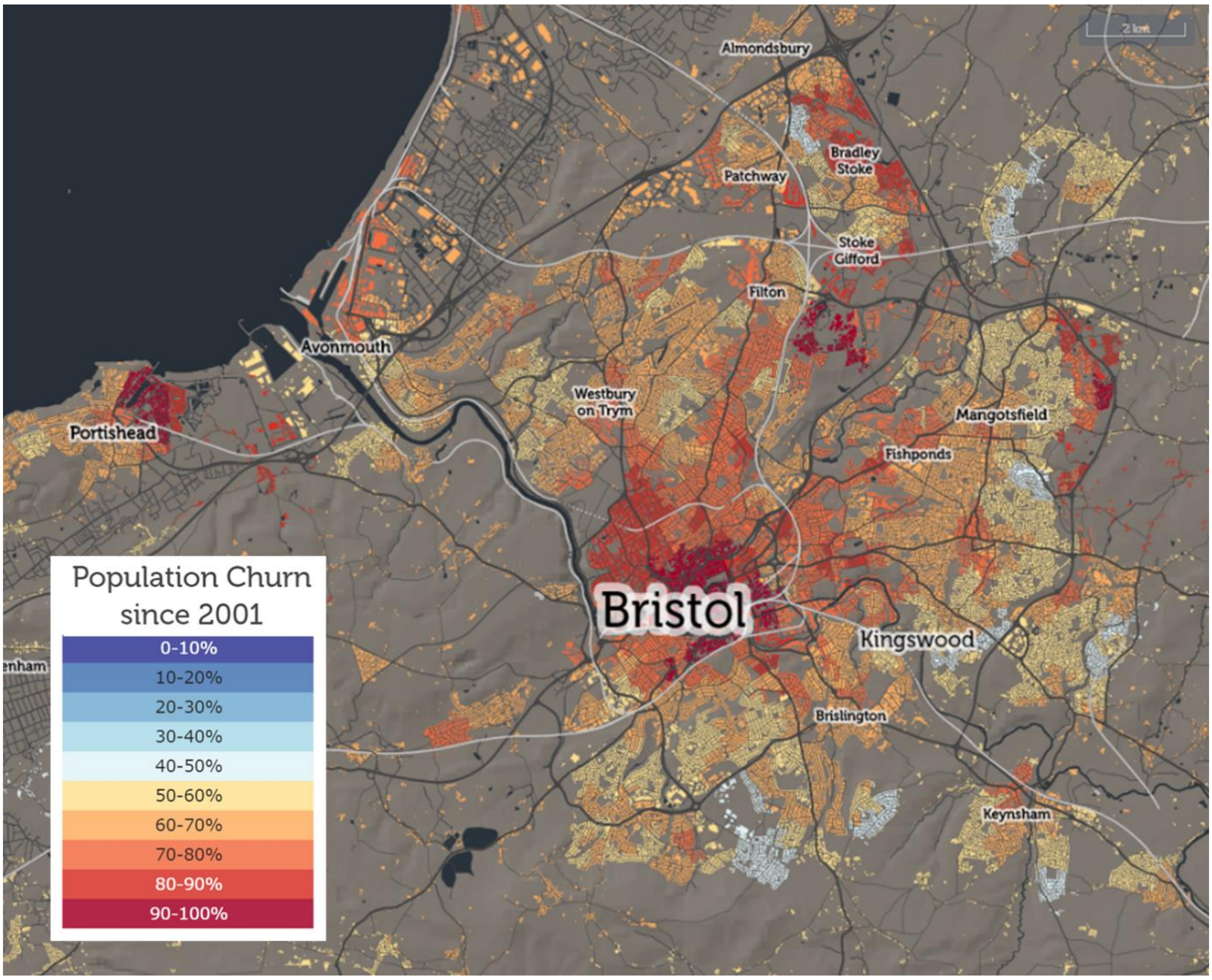


Figure 3. (A) The incidence of household change 2011-16 across Bristol and surrounding areas; and (B) the same measure for 2001-16. Source: https://maps.cdrc.ac.uk/#/indicators/churn/

The analysis of population turnover is just one of many useful applications that can be developed using the LCR. Table 13 identifies some other potential applications that might be useful for understanding local population composition and changes over time.

Variables	**Comment**
Gender and age group	Using a forenames database built from birth certificates and consumer data files it is possible to ascribe the probable demographic statistics to individual-level names data (Lansley and Longley, 2016).
Ethnicity	Annual updates of neighbourhood ethnicity profiles can be developed using Ethnicity Estimator to ascribe probable ethnic group using names (see Kandt and Longley, 2018).
Household compositions	It is possible to detect the number of adults per address and to use surnames as indicators of family membership in order to detect rates of shared households (e.g. see Samuel et al, 2019).
Internal migration	Use of patterns in names within households to investigate the nature of residential transitions (as demonstrated in Lansley and Li, 2018). The granularity of the data enables us to measure trends such as social mobility through linkage to small area data on deprivation or socio-economics.

Table 13. A selection of potential research applications that might use the LCR.

5. Evaluation

Ever increasing amounts of data are collected about citizens today, and an increasing real share of these data are collected by customer-facing organisations. An important contribution of this work is our demonstration that such data can be blended with administrative data and refashioned into comprehensive, timely and granular datasets that can be used for the social good – for example by facilitating better understanding of neighbourhood dynamics and the socio-spatial implications that follow from them. We thus conclude that such data linkage exercises present a pivotal opportunity to potentially broaden our conceptions of population statistics to include indicators of activities and processes. The Linked Consumer Register (LCR) which we have described in this article presents an important underpinning to more granular and frequently updated demographic statistics, which are especially timely given the developing interests in non-traditional social data sources (Hand, 2018) and data-led approaches to robust small area estimations (Tzavidis et al, 2018).

This analysis presents a means of supplementing conventional and new methods of estimating local population size and composition. UK Office for National Statistics Mid-Year Population Estimates are blended from geographically referenced administrative data (such as births and deaths), national-level data (such as international migration statistics), and decennial Census data. This process involves amalgamating data produced at different time periods and at different scales and then attempting to deduce trends for small areas. In contrast, the LCR is built up from large assemblages of public Electoral Register and consumer data sources that remain grounded at the level of the individual. It can therefore offer fresh insights at a highly granular scale and can assist with honing aggregate statistics.

However, the Consumer Registers do not have full and accurate adult population coverage, and their provenance is unknown. This makes it necessary to 'harden' the raw data by anchoring them to conventional statistical sources, where and when these are available. There are thus challenges arising from repurposing data that are not collected for research purposes, not least because the sources of bias in the recorded data may be systematic, and may operate to exclude certain groups in society. Our analysis considered both ABP and PAF as address frames, but additional addresses were added where their occurrences in the Consumer Register could not be reconciled with these frames. Address-based frames are imperfect and also create data issues where they do not perfectly correspond with household definitions or property use categories.

The research described here is consistent with heightened interest in the use of hybrid Big Data sources to supplement or even replace some conventional official statistics (Hand, 2018). In particular, it draws on methods to enhance record linkage at the individual-level, as is often required with administrative and consumer datasets. Our innovative approach to harnessing extensive lists of residents could be adopted by institutions that have access to more complete datasets that are otherwise unobtainable elsewhere. However, the sensitivity of the data dictates that such datasets should only be accessed at highly granular scales within safe research environments. Crucially, in addition to a novel source of demographic data, the names-based individual-level data also provide a spine through which additional administrative and consumer data can be linked and, through this, the provenance of said data can be investigated and new, pertinent and current geodemographic insights may be gleaned.

6. Acknowledgements

This work was funded by the UK Economic and Social Research Council Consumer Data Research Centre (CDRC) grant reference ES/L011840/1. The raw data and aggregated outputs from this project may be obtained upon successful application through the CDRC Data Service. Access will also be granted to specialised software to assist with data linkage and processing.

7. References

Abbott, O. 2009. 2011 UK Census coverage assessment and adjustment methodology. *Population trends*, 137(1), 25-32.

Clark, W.A. and Coulter, R. 2015. Who wants to move? The role of neighbourhood change. Environment and Planning A, 47(12), 2683-2709.

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.

Goerge, R.M. and Lee, B.J. 2002. Matching and cleaning administrative data. New Zealand Economic Papers, 36(1), 63-64.

Hand, D.J. 2018. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 555-605.

Harford, T. 2014. Big data: A big mistake? *Significance*, 11(5), 14-19.

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

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

Lansley, G. and Cheshire, J. 2018. Challenges to representing the population from new forms of consumer data. Geography Compass, e12374.

Lansley, G. and Li, W. (2018). Consumer Registers as Spatial Data Infrastructure and their Use in Migration and Residential Mobility Research. Longley, P., Cheshire, J. and Singleton, A. (Eds.) *Consumer Data Research*. UCL Press. 15-27

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

Longley, P.A., Singleton, A. and Cheshire, J. (Eds.) (2018). Consumer Data Research (pp. 15-27). UCL Press, London.

Lynn, P. and Taylor, B., 1995. On the bias and variance of samples of individuals: a comparison of the Electoral Registers and Postcode Address File as sampling frames. *The Statistician*, 173-194.

Marshall, M.L. 1971. The use of probability distributions for comparing the turnover of families in a residential area, in A.G. Wilson (Ed.) *London Papers in Regional Science 2. Urban and Regional Planning*. Pion, London. 171 -193

Miller, H.J. and Goodchild, M.F. 2015. Data-driven geography. *GeoJournal*, 80(4), 449-461.

Norman, P., Marshall, A. and Lomax, N. 2017. Data analytics: on the cusp of using new sources? *Radical Statistics*, 116, 19-30

ONS. 2017a. Research Outputs: Using mobile phone data to estimate commuting flows. Online: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/populationcharacteristics/researchoutputsusingmobilephonedatatoestimatecommutingflows (Accessed 19.02.19)

ONS. 2017b. Research Outputs: Estimating the size of the population in England and Wales, 2017 release. Online: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatace nsusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/researchoutputsestimatingthesizeofthepopulationinenglandandwales2017release?platform=hootsuite (Accessed 19.02.19)

ONS. 2018. Annual assessment of ONS's progress on the Administrative Data Census: July 2018. Online: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments/annualassessmentofonssprogressontheadministrativedatacensusjuly2018 (Accessed 19.02.19)

Rees, P., Norman, P. and Brown, D. 2004. A framework for progressively improving small area population estimates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(1), 5-36.

Samuel, A., Lansley, G. and Coulter, R. 2019. Estimating the Prevalence of Shared Accommodation across the UK from Big Data. *Geographical Information Science Research UK (GISRUK) 2019*.

Short, J. R. 1978. Population turnover: problems in analysis and an alternative method. *Area*, 10, 231-5.

Stanier, A. 1990. How accurate is Soundex matching. Computers in Genealogy, 3(7), 286-288.

Tzavidis, N., Zhang, L.C., Luna, A., Schmid, T. and Rojas-Perilla, N. 2018. From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4), 927-979.

White, I. and Horne, A. 2014. Supply and sale of the electoral register. House of Commons Library, SN/PC/01020. Online: http://researchbriefings.files.parliament.uk/documents/SN01020/SN01020.pdf (Accessed 01.12.17).