ORIGINAL ARTICLE

Using linked consumer registers to estimate residential moves in the United Kingdom

Justin T. van Dijk (ID) | **Guy Lansley** (ID) | **Paul A. Longley** (ID)

Department of Geography, University College London, London, UK

Correspondence
Paul A. Longley, Department of Geography, University College London, Gower Street, London WC1E 6BT, UK.
Email: p.longley@ucl.ac.uk

Abstract

This paper argues that frequently updated data on the nature of residential moves and the circumstances of movers in the United Kingdom are insufficient for many research purposes. Accordingly, we develop previous research reported in this Journal to re-purpose consumer and administrative data in order to develop annual estimates of residential mobility between all UK neighbourhoods. We use a unique digital corpus of linked individual and household-level consumer registers compiled by the UK Consumer Data Research Centre, comprising over 143 million unique address records pertaining to the entire UK adult population over the period 1997–2016. We describe how records pertaining to individuals vacating a property can be assigned to their most probable residential destination, based on novel methods of matching names, assessing household composition, and using information on the date and probable distance of residential moves. We believe that the results of this analysis contribute highly granular, frequently updated estimates of residential moves that can be used to chart population-wide outcomes of residential mobility and migration behaviour, as well as the socio-spatial characteristics of the sedentary population.

KEYWORDS
consumer data, data linkage, internal migration, linked consumer registers, residential mobility

1 | INTRODUCTION AND OVERVIEW

Every year, in the process of registering the right to vote, most of the UK's adult resident population assents to inclusion on the public version of the Electoral Roll, and others consent to inclusion in contact lists in the course of acquiring goods or services. Subject to appropriate consents being given, these lists are then used by local governments and businesses in further aspects of business and service planning. When concatenated within annual time periods, these data can provide highly granular inventories of local populations and their characteristics, on faster refresh cycles and higher spatial granularity than many conventional statistical sources. In a previous paper published in this Journal, we described the linkage and analysis of consumer and administrative data sources and the provenance of these 'linked consumer registers' (LCRs: Lansley et al., 2019). This paper described how, for each year over the period 1997–2016, registers provide comprehensive, highly disaggregate and frequently updateable representations of population size and structure, along with reliable estimates of incompleteness and possible bias. These registers were linked only in cross section from numerous sources, and models were developed to impute gaps in them when sources failed to detect continuity of residence because data for particular years were missing. The paper appraised the applicability and value of the resulting unique data resource through the derivation of an annual small area household change index.

Our research agenda is to further develop and evaluate these registers along with other new data sources to create timely and pertinent nationally representative datasets for policy analysis (see Longley et al., 2018 for a consolidated statement). Our endeavours using conventional statistics alongside consumer data suggest very high levels of population coverage, and triangulation with conventional statistical sources makes it possible to investigate potential bias and other data quality issues (see Hand, 2018). These innovations can be seen as part of a wider movement to re-purpose new Big Data sources in order to supplement conventional statistics and permit richer analysis of known populations of interest. The specific motivation for this paper is to build on the data infrastructure of the linked consumer registers in order to explicitly link all individual records throughout the 20-year period that they cover in order to better understand the outcomes of intra-national migration and residential mobility.

The task of constructing any geographically extensive longitudinal dataset can be hugely challenging and time-consuming: in the United Kingdom it is additionally complex because population data are captured by three separate statistical agencies covering England and Wales, Scotland and Northern Ireland, while periodic boundary changes can limit direct comparison of aggregated data (Lomax & Stillwell, 2017). Detailed and reliable origin–destination figures on moves between small administrative geographies are available through decennial UK Censuses of Population, although truly population-wide updates have not been collected since the 2011 Census, with subsequent local level estimates derived from administrative sources, principally National Health Service (NHS) records. Greater granularity and specification of origin–destination flows in estimating internal migration is highly desirable in order to provide appropriate levels of services, especially where precise geographic targeting is an issue (Travers et al., 2007). In this context, internal migration plays an important role in shaping the current population structure at the local level (Lomax et al., 2014).

Data on population movements can contribute to understanding of related processes of internal migration (typically motivated by employment opportunities) and other residential mobility (typically motivated by adjustment of housing requirements to changed household circumstances): see Coulter et al., 2016. In response, the social composition of neighbourhood areas may be considered to endure or change over time, depending on whether the limited numbers of movers are replaced with others of similar ilk (Timms, 1975) or whether structural changes in residential composition manifest

neighbourhood gentrification or deterioration (Hamnett, 1991; Harvey, 1973; Park & Burgess, 1925). In both cases, population movement (or the lack of it) is the engine of stasis or change (see Smith & Denholm, 2006). At coarser regional scales, the net effects of neighbourhood change may both drive and be driven by local labour markets and may be associated with changing patterns of social inequalities (Fielding, 1992) as well as wider economic, political, cultural and environmental contexts (Smith et al., 2015).

A primary motivation for residential moves is the desire of households to improve or adapt to changing living circumstances, rendering residential mobility an indicator of social mobility (Fielding, 1992) as well as family life cycle change (Stapleton, 1980) or life course transitions (Tyrrell & Kraftl, 2015). Where some places offer greater opportunities than others, as manifest through high employment levels, they typically attract young migrants (Fielding, 1992; Rees et al., 1996), resulting in positive net migration flows (as first postulated by Ravenstein, 1885). Young adults are typically more geographically mobile as they tend to move independently rather than as a household: Bell et al., (2002) report that young adults have the highest propensities to move, and that they subsequently become increasingly sedentary with age until eventual retirement. This said, Fielding (2012) also suggests that the relationship between life course and migration is intricate and that spatial outcomes can differ between places as a result of some young adults postponing their entry into the labour market or delaying family formation. Taken together, understanding issues of labour market differentiation, kinship and identity are all manifest in residential moves (see Clark & Moore, 1980; Fielding, 2012; Finney & Simpson, 2008; Smith et al., 2015), and highly granular measurement of household structure and origin–destination attributes is essential to improve understanding of motivation and process.

In the absence of other comprehensive and frequently updated data on the nature of local residential moves and the circumstances of movers, this paper proposes a highly disaggregate framework for measuring residential moves in the United Kingdom. Building on the LCRs, we develop explicit longitudinal linkage of all records throughout the 20-year period covered by the registers in order to uncover the outcomes of intra-national migration and residential mobility decisions. Ascertaining whether moves are motivated by domestic circumstances (typically described as residential mobility) or household economics (typically defined as migration), is difficult (see Coulter et al., 2016) and we consider that either can result in address transitions between annual updates of the consumer registers. We describe how records pertaining to individuals that vacate a property are assigned to their most probable destination address, based on novel methods of name matching, assessment of household composition and use of information on the date and probable distance of residential moves. The individual level results of this migration model are aggregated and, in the case of the 12 months prior to the 2011 Census, are compared with official statistics. The results of this analysis contribute highly granular, frequently updated estimates of residential moves and can be used to chart population-wide outcomes of residential mobility behaviour, and to examine the characteristics of individuals and households that do or do not move.

2 | HOUSEHOLD MOVES AND THE LINKED CONSUMER REGISTERS

The temporal and spatial granularity of conventional statistical sources is widely understood to be insufficient to understand the outcomes of a number of residential mobility or migration processes (Lomax et al., 2013; Lomax & Stillwell, 2017). While cross-sectional Census data are only available at 10 yearly intervals, longitudinal census records link successive censuses, albeit for just 1%

of the population (Champion & Shuttleworth, 2017). Small sample sizes also limit the geographic granularity of other Office for National Statistics (ONS) surveys such as the Labour Force Survey and the General Household Survey, with deleterious consequences for understanding the detail of neighbourhood change. As a consequence, researchers and practitioners have sought alternative data sources which might reliably inform about residential mobility occurrences on a more regular basis.

One of the most popular alternative sources of data on migration, used by the ONS to compile Mid-Year Population Estimates, has been the NHS Central Register (NHSCR), outputs of which record the moves of patients between, but not within, health authority areas (Lomax & Stillwell, 2017). Unfortunately, this source was discontinued in February 2016 (ONS, 2020a) and the subsequent Patient Register Data Service (PRDS) for England and Wales was left as the main dataset feeding into the annual population estimates (Lomax & Stillwell, 2017; ONS, 2020a). Where the Patient Register does not cover within-year moves, the Personal Demographic Service (PDS), which replaced the NHSCR in mid-2017, is also now used in calculating Mid-Year Population Estimates by the ONS. Similar to the NHSCR, the PDS records weekly updates on the movements of patients and, together with the PDS, is used to estimate residential moves between local authorities. The PDS also makes it possible to estimate cross-border flows between England and Wales, Scotland and Northern Ireland (ONS, 2020b). However, even where service organizations have a mandate for universal coverage, as with the NHS, address records are patchy in quality and some transient groups (especially young adults) are heavily under-recorded and, more generally, short distance moves are underrepresented (Lomax et al., 2013; Lomax & Stillwell, 2017). Moreover, these data are only made available for research purposes at relatively coarse spatial aggregations (Stillwell & Thomas, 2016).

Unlike NHS data sources, the LCRs are compiled from administrative and consumer sources and are created using a blend of both deterministic and fuzzy procedures in order to link public versions of the UK Electoral Register and consumer files over the period 1997–2016. The first 6 years of the series are drawn from the full Electoral Register prior to the introduction of opt-out provisions in 2003, and numerous consumer sources are used for subsequent years in order to supplement the public (post-opt-out) registers. The component registers were obtained in annual releases from a range of industry value-added resellers, although the identities of the different private sector providers are not known. The combined coverage of the registers wanes over time, principally because consumer data sources do not fully compensate for increasing rates of opt-out from the public Electoral Register. Additionally, the early years of the LCRs have known 'non-voter' bias (see Electoral Commission, 2016; Hoinville & Jowell, 1978), but the completeness and provenance of the post 2003 registers is largely undocumented. A full discussion of remedial steps to address these issues is provided in Lansley et al. (2019).

The annual component LCRs were each created through linkage of each of the component annual registers to the best available address frame, comprising Ordnance Survey AddressBase Premium (Ordnance Survey, 2020) and the Royal Mail's Postcode Address File (PAF: Royal Mail, 2020). AddressBase Premium is the most comprehensive geographic dataset of addresses available in the United Kingdom. The Postcode Address File is a list of all postal address in the United Kingdom, owned and maintained by Royal Mail. In addition to addresses, the LCRs record adults' given and family names and the first and last time they were observed at any given recorded address. The total numbers of adults recorded in each year of the LCRs were within 2% of official ONS Mid-Year Population Estimates, after including the very small number of imputations used when house sales were known to have occurred but no new residents were found in any later time period.

The LCRs have many useful applications that can be tailored to bespoke temporal and geographic aggregations, subject to strict disclosure controls. Estimates of neighbourhood population turnover have been created by detecting when new household units join an address (Lansley et al., 2019). Elsewhere, name-based tools have been used to infer ethnicity (Kandt & Longley, 2018) and hence create neighbourhood estimates of changing ethnic composition (Lan et al., 2019); similar tools have been used to chart intergenerational population change (Kandt et al., 2020). In important respects, the benefits of the LCRs accrue from their granularity and their regular annual update cycle.

3 | RECORD MATCHING PROCEDURES AND THE SPACE–TIME GEOGRAPHY OF RECORDED MOVES

The LCRs comprise only names, addresses and timestamps indicating probable start and end dates of residence. In this paper, we present a novel methodology to link those leaving an address to their most probable destinations, using timestamps, relative locations and combinations of names within households recorded at addresses. The methodology essentially entails linking apparent disappearances of named individuals at one address and reappearance at a different one within predefined timeframes. This procedure works very effectively where combinations of individuals identify a unique household, or where an individual's given and surname combination is rare (or even unique). In other cases, it is necessary to use a probabilistic method to assign names that could link multiple origins and destinations, incorporating time and distance factors. Once completed, our results are compared with official statistics from the 2011 Census and other aggregate migration estimates. Here we set out the procedures used to construct internal migration estimates from the LCRs and the empirical findings arising from the matching process.

3.1 | Matching of forename–surname pairings

Names are tokens that provide reliable ways of tracing individual and household movements because they are usually retained throughout their bearers' life courses, unless changed upon marriage or gender reassignment. While most forename–surname pairs are not unique, they are usually very uncommon: see Figure 1.

While it is thus not always possible conclusively to identify a unique individual from their full name, household combinations at unique addresses are very likely to be distinctive. For example, LCR data identify the most common full name bearer combination at a single address in 2016 as John Smith and Margaret Smith living in the same household, occurring at 170 addresses: the component names John Smith and Margaret Smith occurred individually 9,853 and 6,122 times respectively. Viewed in each of the 20 years for which LCR data are available, an average of 62.6 % of households comprised unique name compositions, a figure that rises to 85 % for households comprising two or more persons. The proportion of unique household combinations has gradually increased since 1997, and only 7% of households share name combinations with more than 50 other households, the majority being lone adults. The LCRs document an apparent decrease in average household size with lone adults increasing from 36 % to 46 % over the 20-year period: this figure is slightly higher than the 41 % recorded by the 2011 Census of Population, and may manifest replacement of (self-assigned) head of household registration for the Electoral Roll with individual registration in 2014. Failure to match addresses drawn from different LCR sources may also contribute to the higher LCRs figures. Against the background of these characteristics, individuals' names and their household aggregations were
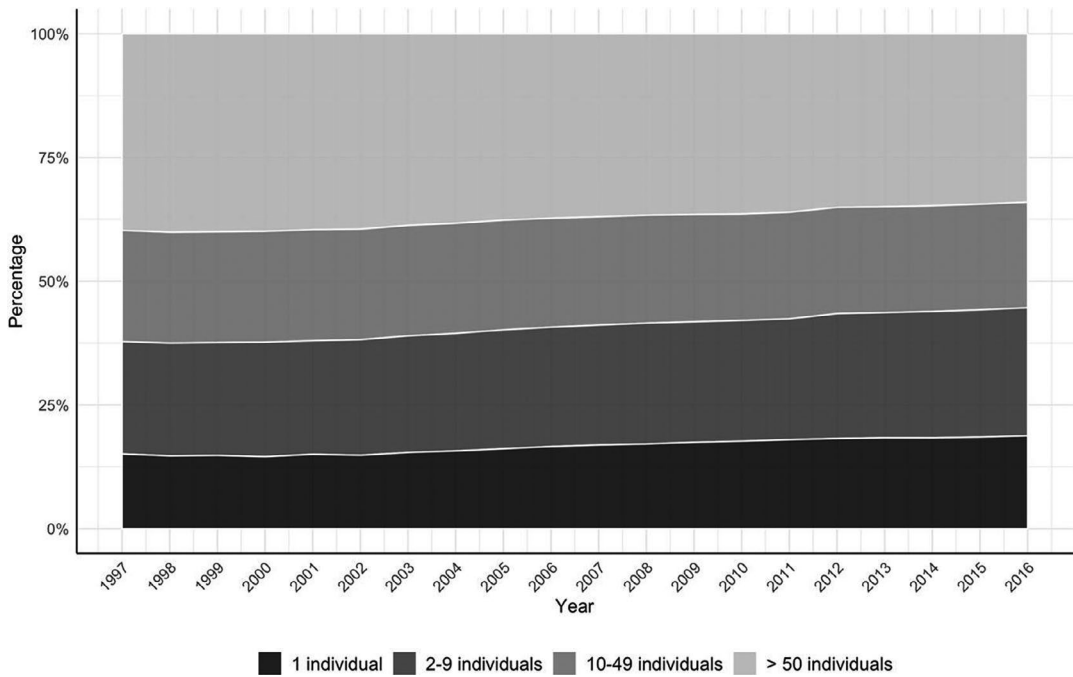
FIGURE 1 Percentages of the adult population captured in the linked consumer registers bearing forename–surname pairings that occur 1, 2–9, 10–49 and 50 or more times over the period 1997–2016 (source: Author calculations)

used to trace likely movements between addresses across the 20-year LCR period, resulting in either unique matches or a shortlist of candidates for further consideration.

3.2 | Temporal lags in recording moves

The format and precision with which residence at addresses is recorded varies between the constituent LCR data sources. Land Registry data record changes in property ownership, albeit not always the precise dates on which sales take effect. Many residential moves take place in 'housing chains', wherein house finance requires coordination of multiple moves on the same date, which can render linkage of household movements straightforward. Address-level house sales data from the Land Registry (Open Data for England and Wales) and Registers for Scotland (obtained from the Urban Big Data Centre) were linked to the LCRs in order to cross validate some of the apparent moves and to hone estimates of the dates upon which house sales in the owner-occupier sector triggered moves. Just under 14 million entries in the LCRs were linked to at least one property sale using Land Registry (England and Wales) or Registers of Scotland price paid data. Voter registration data are of more mixed quality, but their provenance is quite well understood: the Electoral Commission (2016) has identified that the majority of electors re-register within 2 years of changing address, although lags of 2 to 3 years are not uncommon, particularly between General Elections. However, it is also estimated that 17 % of eligible voters (9.4 million adults) in Great Britain are not correctly registered at their current address, and that 11 % of the full register entries are inaccurate, affecting up to 5.6 million adults (Electoral Commission, 2019). Comprehensive accuracy assessment is therefore difficult, since the timestamps from the various consumer data sources are also of largely unknown provenance.

3.3 | Probable distances of moves

The attenuating effect of distance upon numbers of residential moves has been broadly understood for over 125 years (Lee, 1966; Ravenstein, 1885). Today it remains the case that the majority of moves occur over short distances: the 2011 Census records that 57.1 % of the individuals aged 16 and over that changed address within the preceding 12 months moved within the same Local Authority District (LAD). Previous use of conventional aggregate statistics has not established detailed distributions of distances moved (Stillwell & Thomas, 2016), although evidence suggests that residential mobility over longer distances has become less common since the 1970s (Champion & Shuttleworth, 2017). In recent years the growth of the private rental sector (where short-term lets are common) is also understood to have increased the frequency of moves, although such moves often remain focused upon the same employment centres (House of Commons, 2013).

4 | IMPLEMENTATION OF MATCHING PROCEDURES

We developed a two-stage model in order to estimate the origin–destination pairings of as many apparent movers as possible. The tension in this Big Data problem, applied to the circa 143 million individual name and address records in the LCRs, was to capture as many actual moves as possible, while remaining cognizant of the risk of assigning false positives in an incomplete dataset of variable data quality. Processes of household formation and dissolution needed also to be considered, as well as the effects of international migration. In the first stage of the procedure, deterministic assignment was used to link uniquely named lone individuals and their household members between origin–destination pairs in each time period. The second stage used a probabilistic approach to match remaining individuals who appeared to have moved, by developing a highly computationally intensive procedure in which every possible interaction within a 3-year window was identified and the most probable pairs were matched.

Lansley et al. (2019) describe how the LCRs are engineered as a smoothed time series in which any gaps in annual records of a named individual at an address between their dates of moving in and out are simply filled using that individual's name. They also describe how household characteristics are imputed in later years where individuals are known to have moved out but the names of replacement residents are not known (e.g. because they have not yet registered as voters). In such instances, the number of replacement residents is noted but the records not used in the residential mobility analysis reported here: the number of cases increases in the final years of the time series because there is only a short run of records over which gaps may be plugged. A consequence of more frequent resort to imputation in later years is that the residential mobility results likely become less complete, pending inclusion of post 2016 data that plug gaps by picking up new residents, for example through lagged voter registration.

Our initial premise was that the start and end dates of every residence history recorded in the LCR could be used to identify a move within the United Kingdom—without addressing additions or losses to the system arising from comings of age, marriages (if associated with a name change), deaths and international migration. Intra-UK migration indeed accounts for the majority of transitions, as evidenced by 2011 Census estimates that identify that over five million adults changed UK address in the preceding 12 months. As mentioned above, Land Registry property sales data were used, where possible, to calibrate the start and end dates of LCR residential histories where single or multiple residential moves were identifiable from financial transactions. In other cases, the 'first seen' and 'last seen' dates recorded in the LCRs were used. The frequency distribution of the recoded residential histories that result is shown in Table 1.

TABLE 1 Frequencies of 'last seen' and 'first seen' occurrences in each year of the linked consumer registers (LCRs)

Year	Last observation	First observation
1997	6,209,906	45,342,404
1998	6,065,950	8,123,676
1999	8,142,354	6,528,756
2000	4,045,262	8,437,748
2001	6,288,883	3,055,384
2002	5,744,780	7,433,461
2003	5,001,650	5,099,172
2004	5,578,480	4,553,431
2005	5,108,975	4,665,936
2006	3,979,450	4,793,242
2007	3,542,251	4,698,893
2008	4,438,117	3,654,831
2009	5,325,528	6,003,444
2010	6,727,176	6,203,975
2011	13,744,674	4,779,294
2012	7,575,166	5,016,716
2013	6,415,383	6,292,786
2014	6,586,815	2,111,463
2015	6,814,220	2,409,062
2016	26,454,029	4,585,375

In this Table, the 'last observation' figure relates to all individual records that are not recorded at the same address in any subsequent year, while all other records are defined as 'first observations'. Thus the 'last observation' figure for 1997 indicates that there were 6,209,906 individual records that are not recorded at the same address in any of the subsequent years and are therefore entered into the linking procedure. A feature of each annual consumer register is that different sources contribute towards them and under the terms of data supply, detailed metadata of non-Electoral Register sources are not provided. Reduction in the volume of consumer data provided for 2011 has the consequence of inflating the 'last observation' figure, and the 'first observation' figure is reduced. This largely accounts for the different sizes of the 'first observation' and 'last observation' figures across the years.

4.1 | Stage 1

The 'first seen' and 'last seen' data, verified where possible using Land Registry property sales data, were used to bound residences for all forename–surname pairings with precisely two occurrences within the LCRs using the following steps:

1. If a forename–surname pairing disappeared from an origin and subsequently reappeared at a destination and remained there for more than a year, the records were linked as a move.

All such linkages were considered as definitive and were therefore excluded from the second stage of the matching procedure.

2. For each uniquely matched forename–surname pairing, the non-unique names of any associated household members were retrieved and added if they matched the same origin–destination pairing. Household members were defined as any individual that shared the same address at any time during the residence of the uniquely named individual at the address. These individuals were also removed from the second stage of the matching procedure.

In total, Stage 1 identified 3,045,108 moves, an average of 152,255 moves per year. Figure 2a shows that a large share of matches—accounting for 43.1 % of the total number of all matched records—were made over consecutive years, consistent with Electoral Commission findings that most individuals re-register within 2 years of changing address, albeit that only a minority re-register within 12 months of moving (Electoral Commission, 2016). Figure 2b confirms a strong distance attenuation effect upon mobility, despite there being no distance weighting included in the matching procedure. Over the 20-year LCR period, 62.3 % of individual moves capture in this stage of the model occurred within the same local authority: the equivalent figure from the 2011 Census for individuals aged 16 and over is 57.1 %.

4.2 | Stage 2

The Stage 2 model was devised to match as many unmatched individuals leaving an address as possible with their most probable name-matched pairings. This entailed attempted linkage of all remaining names observed to depart from an address to every bearer of the same name that joined any other address within a 3-year window, including the same year. There were c. 7.8 billion candidate non-unique pairings over the 20-year period, including multiple moves by the same individuals. For example, 1,243 John Smiths were last observed at addresses in 2013, making potential matches for any of the 14,238 John Smiths that were first recorded at other addresses during a subsequent 3-year window.

In addressing this huge combinatorial problem, we draw upon Gale and Shapley's (1962) 'stable marriage problem', which develops an algorithm to allocate men and women into suitable marriages using the ranked preferences of each individual. Our adaptation of this approach allocates individuals believed to have left addresses to the most probable vacated properties. Arranged by order of importance, our score was based on the number of residents' names that matched, whether both properties were sold on the same day, the distance between the two properties, and the time lag between the last observation at the origin address and the first observation at the new one, taking into account normal time lags in detection of new voter registrations, etc. The precise score comprised:

a. A count of the number of full names that matched at each candidate address;
b. whether linked Land Registry or Register for Scotland data indicated that the origin and destination properties were (weighted 1) or were not (weighted 0) sold on the same day or sold in the same year (weighted 0.1);
c. the distance separating the origin and destination, measured as the inverse of the straight-line distance connecting origin and destination, range standardized from a theoretical maximum distance of 1,200 km to a value between 0 and 1; and
d. the time elapsed between the 'last seen' date at the origin address and the earliest 'first seen' occurrence at the destination, rescaled to values between 0 and 0.1 and with moves within the same year weighted the same as moves matched across 2 years.
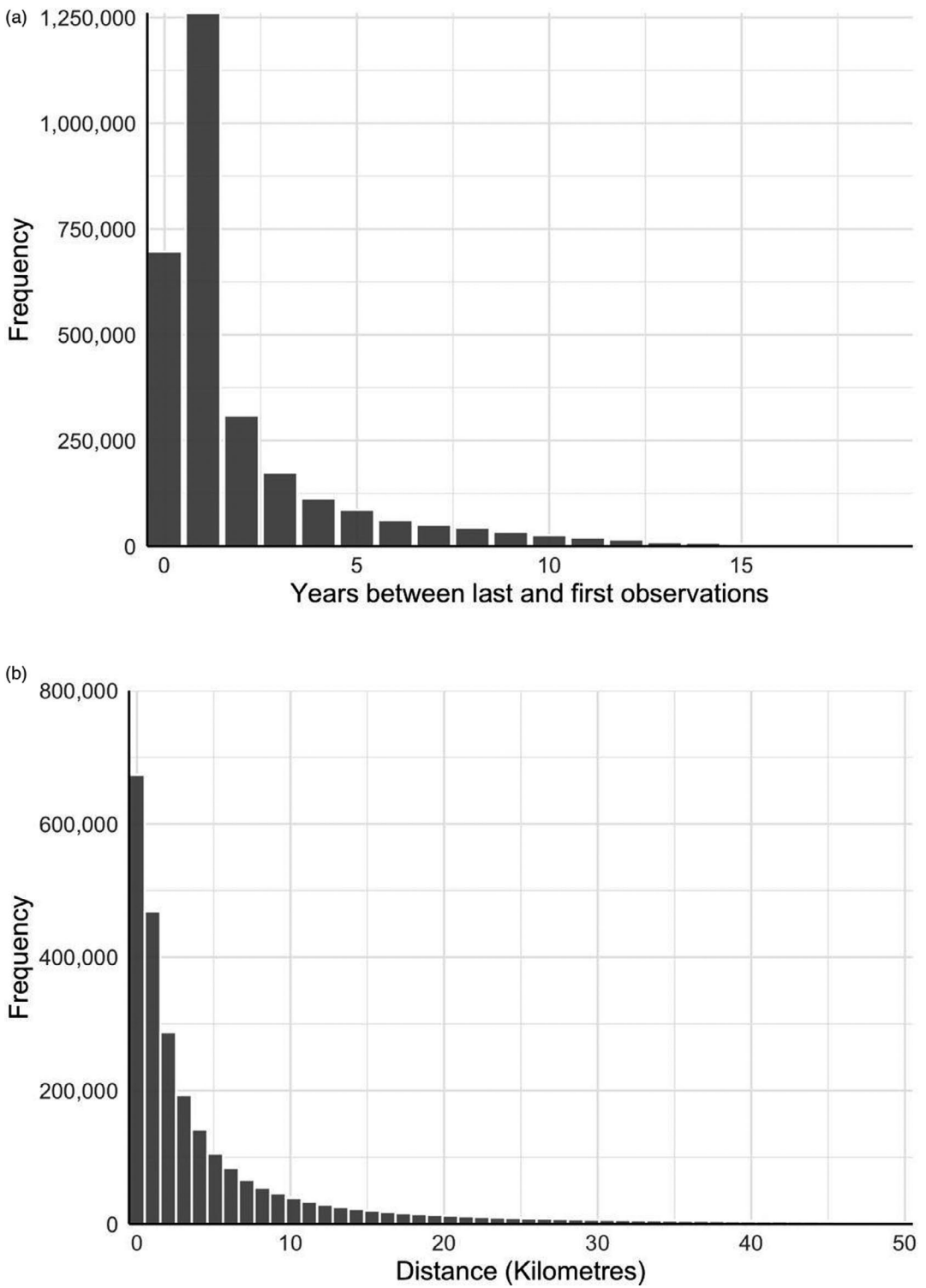
FIGURE 2 (a) Lag time (years) required in order to match moves using names; and (b) distribution of straight-line distances moved by matched individuals in Stage 1

4.3 | Additional data cleaning and record selection

The combination of Stages 1 and 2 returned c. 47.5 million moves across the 20-year period covered by the LCR. Analysis of origin–destination pairs in Stages 1 and 2, however, identified many apparent moves over very short distances, including c. 3.5 million moves within the same unit postcode. Exploratory analysis indicated that some of these addresses appeared in multiple formats in successive LCR entries, despite the best efforts to standardize them. Three diagnostic checks were therefore devised in order to remove probable duplicate addresses in the combined results of Stage 1 and Stage 2:

1. full string matching of the first line of the address (irrespective of postcodes or other inconsistencies);
2. use of Soundex fuzzy matching (Stanier, 1990) to filter out addresses that were very similar when read aloud (and may have been captured incorrectly where consumer addresses were entered through dictation); and
3. identification of addresses that differed by just one or two characters, as identified using the Levenshtein Distance measure.

Together, these additional checks matched 6.7 % of the moves originally identified in Stage 1 and 2, which were reclassified as pertaining to the same address. The probability density distribution of distances over which these moves took place is shown in Figure 3. The majority of probable duplicate addresses as identified by the address matching, Soundex matching or Levenshtein matching are moves over <100 metres—suggesting that these moves were indeed a result of formatting differences in the LCRs. We then took out all moves that arrived at a destination in 1997 and left an origin in 2016: these are, respectively, the beginning and end of the time series, for which we do not know whether arrivals at a destinations or departures from origins are the result of a move simply an artefact of the time series—and so might be recorded as false positives. After removing the probable
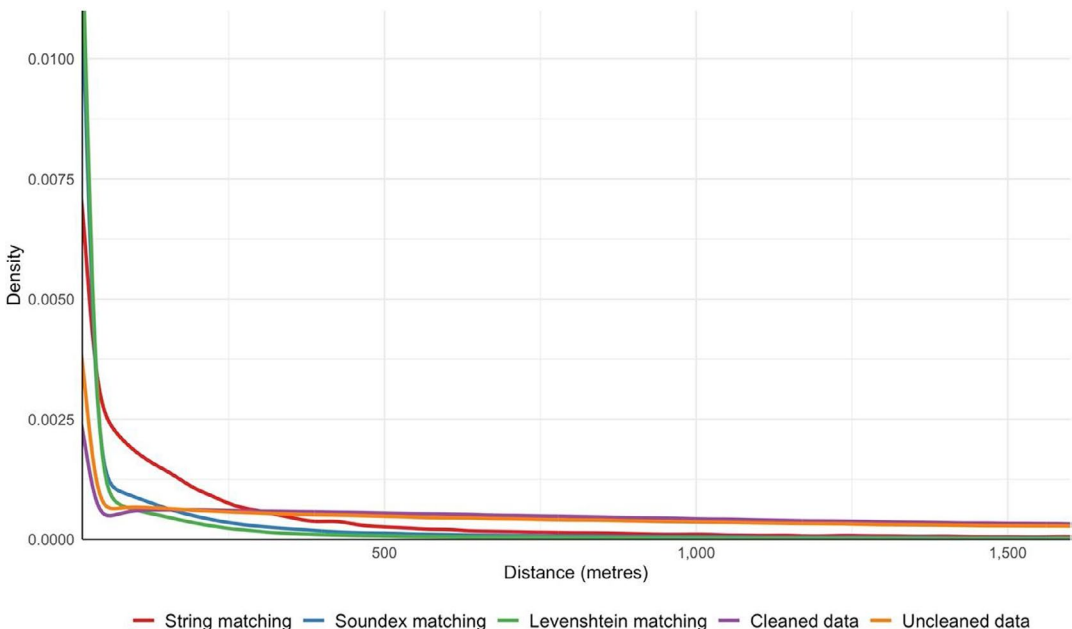

FIGURE 3 Density distributions of assumed similar addresses by distance of apparent move

duplicate addresses and updating the records to exclude moves ending in 1997 or beginning in 2016, 41,658,922 moves were identified over the 20-year period covered by the LCRs—yielding an average of 2,082,946 moves per annum.

5 | RESULTS

5.1 | Aggregate estimates

Although the procedures set out here were applied consistently to all of the LCRs, the results inevitably reflect (a) the completeness of the raw source data and (b) the extent to which it is possible to replace gaps in the time series, which is greater in earlier time periods. The apparent fluctuation over time in LCR estimates of median distances moved grouped by year of arrival (Figure 4) should be viewed in this context. The early years (1997–2002) of the LCRs derive from a single administrative source (albeit collected by multiple local authorities) and return the smallest median distances of moves for these years. (Note that 1997 is excluded from Figure 4 as 1997 is not used as a year of first arrival because it is the start of the time series). Subsequent 'opt out' provisions for the Electoral Roll source and use of multiple consumer data sources to top up the registers creates greater uncertainty about the completeness and provenance of the data, and this is associated with increased estimated median values of distances moved, particularly in 2014 and 2015. Gaps in the register in any year can
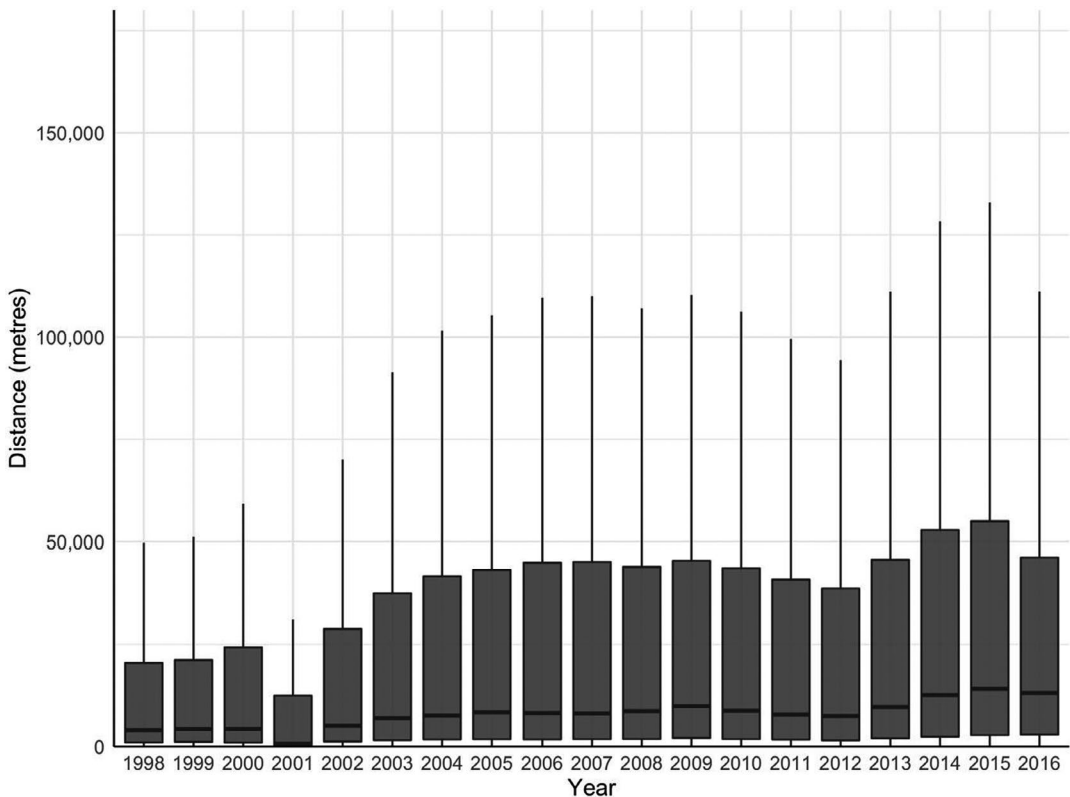

FIGURE 4 Median distance moved estimates (boxes) and interquartile ranges (whiskers) grouped by year of first arrival, 1998–2016

be filled by carrying forward observations from previous years or carrying back observations from subsequent years. Post 2013 the 'carry back' window is increasingly truncated by the impending end of the time series and so the numbers of unfilled gaps progressively increase the probabilities of mismatching individual records. A manifestation of this is likely to be at least part of the increased distance of move estimates in the later years, although of course it may also arise because of increases in the actual distances moved. Accordingly, and as we discuss below, it may be appropriate to consider the later distance statistics as provisional, in much the same way as some conventional ONS statistics are badged as provisional.

2011 Census estimates (ONS, 2014) of residential moves within and between local authorities of residents aged 16+ are broadly comparable: the UK-wide Pearson correlation coefficient comparing the origin–destination counts recorded in the 2011 Census with the assigned LCR origin–destination counts for the corresponding year is 0.97. This value is calculated using the pairwise interaction matrix of 380 harmonized LADs and accounts for the flows in both directions between every pair of LADs. In terms of coverage, the LCR matches are equivalent to 63.4 % of the total number of moves recorded in the 2011 Census: 52.3 % of the total number of recorded intra-Local Authority moves and 78.4 % of the total number of recorded inter-Local Authority moves. For 2011, the median distance of moves between all origins and destinations using the LCRs is 8.14 km, which is almost twice the 4.2 km captured in the 2011 Census flow data (ONS, 2015): this is likely to be because the LCR captures a larger share of inter-Local Authority moves. The 2011 LCR median intra-Local Authority distance moved is 1.7 km, while that captured by the 2011 Census is 1.7 km. For the inter-Local Authority moves these median distance values are 37.3 and 30.6 km respectively.

No UK-wide origin–destination tables are available for inter-Local Authority moves post 2011, although pairwise annual mid-year estimates of moves derived from NHS patient registers provide data on moves between LADs in England and Wales are available (ONS, 2020b). The ONS also releases annual local area estimates of the total number of internal migrants flowing into and out of each UK District (ONS, 2019b). In Figure 5 we present estimates of moves into and out of each District, expressed as ratios between the LCR and ONS estimates. Here, a ratio of unity identifies perfect correspondence between the estimates. LCR under-predictions outnumber over-predictions, but where outward moves are over-estimated so too are inward moves, and vice-versa. ONS UK-wide District Mid-Year Population Estimates pertain to all individuals (ONS, 2019a) rather than adults, and as such are expected to be larger than their LCR counterparts. The Pearson correlation coefficients between these total inflows and total outflows for the post-Census years of the LCRs remain stable at 0.93 ± 0.01. Figure 5 also shows that the correspondence between estimates endures over time: the mean ratios for inflows and outflows in 2011 are 0.72 ± 0.18 and 0.71 ± 0.17 respectively, and the corresponding figures for 2016 are 0.65 ± 0.15 and 0.65 ± 0.16. The slight decrease in the mean figures over time can most likely be ascribed to the LCR recording a lower fraction of the total adult population in the last years covered by the Registers.

Figure 6 confirms that the 2011 patterns of residential mobility between LADs estimated using the LCRs are broadly consistent with those recorded by the Census, although not all longer distance moves from and to London are captured—most likely because the migration model trades off distance and household composition when evaluating text string matches in order to reduce the likelihood of false positive matches. However, shorter distance moves from Northern Ireland to mainland United Kingdom to destinations close to Liverpool and Manchester are more common than recorded in the Census. The left-hand map shows the 2011 flows reported by ONS (2014) with a minimum two-way flow of 200 and the right-hand map shows the LCR estimates with a lower minimum two-way flow of 150—in recognition that the LCRs pick up 63.4 % of all moves and 78.4 % of all inter-Local Authority flows.
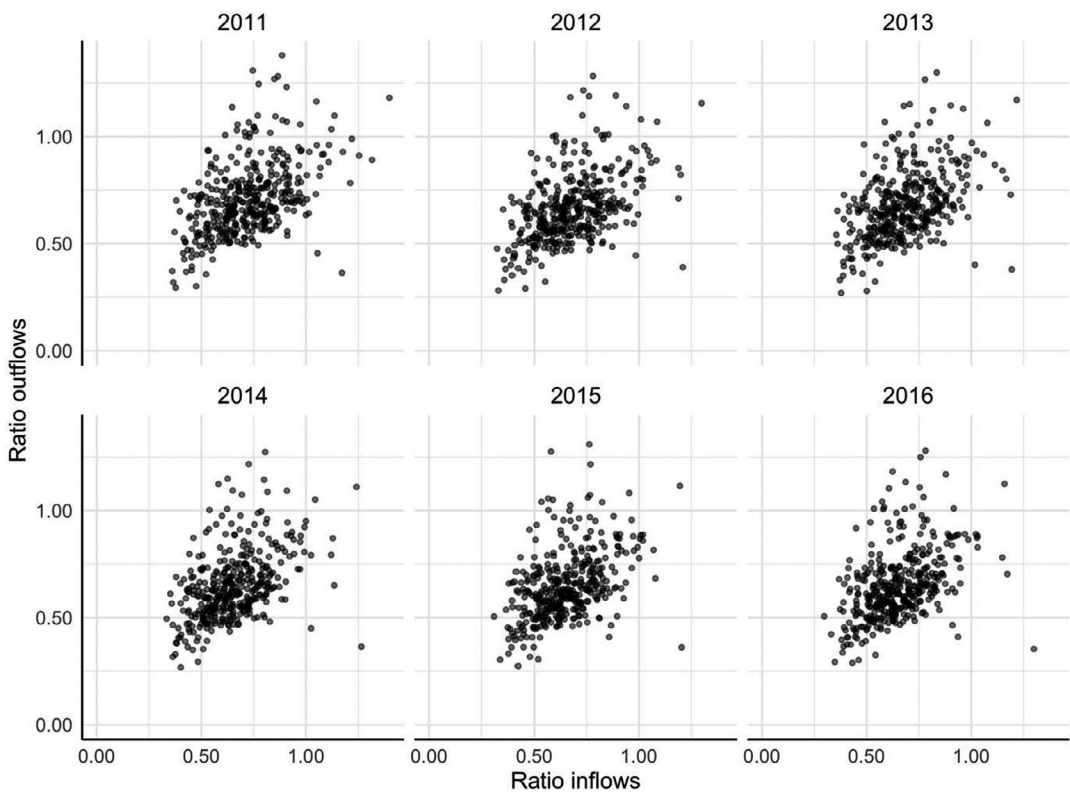
FIGURE 5   Estimates of ratios between the linked consumer registers and office for national statistics. Local Area estimates of total inflows and total outflows, for selected years

A great strength of these estimates is their granularity and local specificity, but the lack of these qualities in most frequently updated statistics frustrates many direct comparisons. Figure 7 nevertheless shows two social aggregations. Figure 7a illustrates that the 2011 Consumer Register has a consistent coverage of all Output Area Classification (OAC: Gale et al., 2016) Super Groups when compared to the Census flow data (ONS, 2015). Estimated in- and out-migration rates are balanced and consistent with general processes of household formation, progression through the housing market and dissolution. Figure 7b shows similar consistency when moves are broken down by Index of Multiple Deprivation (IMD) deciles for Great Britain (2019 IMD for England and Wales, 2020 IMD for Scotland), and changes in the balance between in-migration and out-migration suggest general upward filtering of households through the housing market. The over-all message is thus of consistency of LCR estimates with the 2011 Census. These findings provide a platform for further use of the data for locality studies, to which we now turn.

5.2 | Disaggregate estimates

A major advantage of the grounding of the LCRs at the level of the individual is that, subject to disclosure control, estimates of residential mobility may be generated for any convenient aggregation. To this end, Figure 8 presents mobility patterns between Middle Super Output Areas (MSOAs) in Greater London, subject to minimum two-way flows of 10 or more individuals. These results identify that many Outer London Boroughs (e.g. Kingston-upon-Thames and Sutton) host large numbers of moves
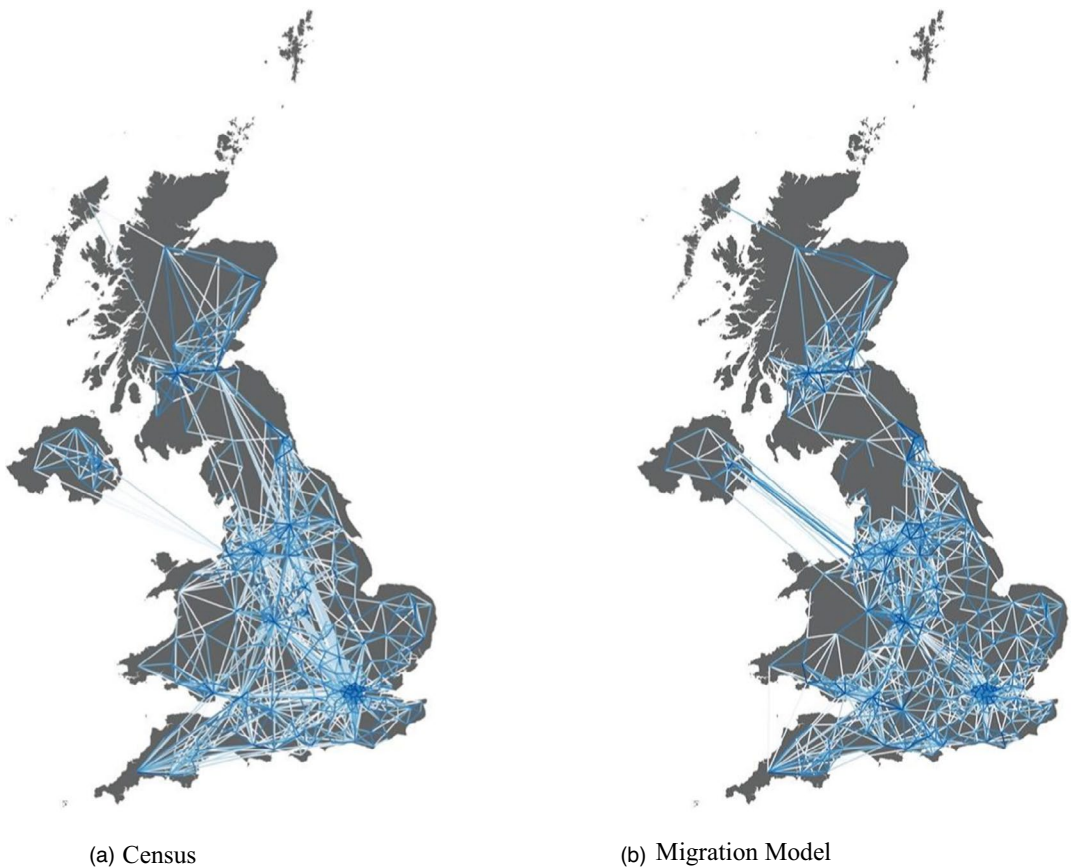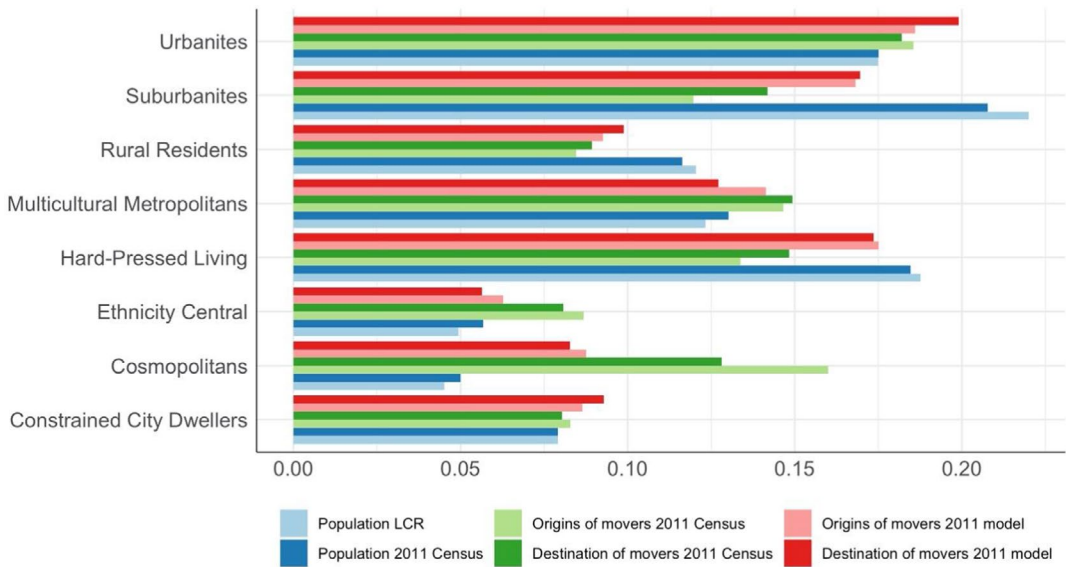
(a) Census

(b) Migration Model

FIGURE 6 UK-wide principal migration flows between Local Authority origin–destination pairs in 2011: (a) As reported by ONS (2014); and (b) as estimated using the LCRs. Darker lines indicate larger flows

within their boundaries and that these outnumber moves further afield. Physical or psychological barriers remain associated with diminished mobility, such as moves across the Thames or across the Lea Valley Park separating Waltham Forest from Enfield, Haringey and Hackney.

In a similar vein, over-all trends may be disaggregated at local level, in order to envision the precise ways in which residential filtering occurs. Knowledge of the precise origins and destinations of moves makes it possible to characterize the ways in which local housing markets function across a full range of scales. For example, the innermost ring of Figure 9a details how the gentrifying area of Spitalfields and Banglatown in East London attracts an estimated 24.3 % of movers from UK origins outside London, shown as the sectioned arc across the top of the innermost ring. Those moving out, shown in Figure 9b are less likely (20.6 %) to select destinations outside of London, shown by the shorter sectioned arc in a similar position on the innermost ring. Over the full period, the migration model captured 7,847 moves going into Spitalfields and Banglatown and 9,202 moves going out.

Selected popular origins and destinations are labelled in this Figure, subject to disclosure control thresholds, but all are identifiable using the LCRs. In Figure 10, origin (Figure 10a) and destination (Figure 10b) locations of movers into and out of this area are characterized as more or less deprived as measured using the 2019 IMD for England and Wales. Although net population change is quite

(a) 2011 Output Area Classification Super Groups
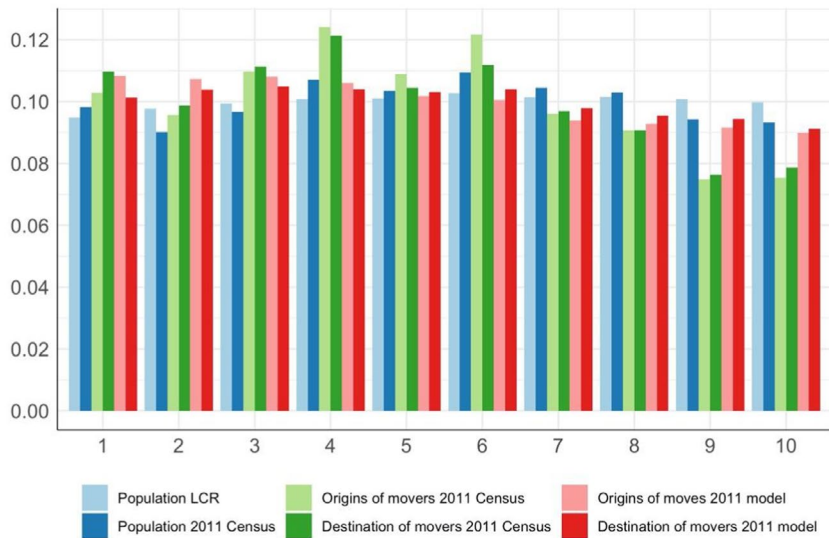
(b) Index of Multiple Deprivation

FIGURE 7 (a) 2011 Census and linked consumer register (LCR) adult population estimates for each of the 2011 Output Area Classification Supergroups, 2011 LCR estimates of adult individual moves into and out of these categories, and 2011 Census counts of individual moves into and out of these categories; and (b) proportions of the 2011 LCR falling into each Index of Multiple Deprivation (IMD) decile and associated estimates of moves into and out of each decile

small (Table 2), it is quite clear that incomers to Spitalfields are somewhat more likely to come from areas in the least deprived IMD deciles, while the destinations of out-movers over the period 1997–2016 are less salubrious. Such measures make it possible to profile changes in the characteristics of neighbourhood residents and to relate these to the trajectories of their neighbourhoods (cf. Rabe & Taylor, 2010).
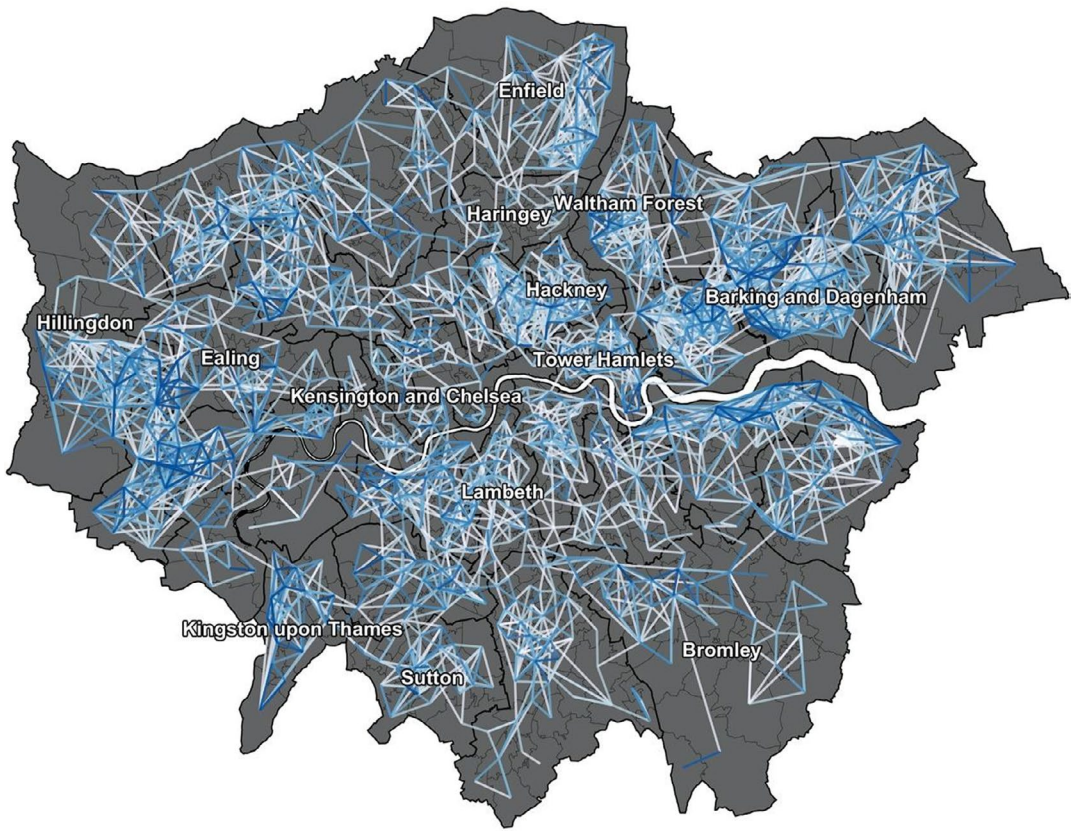
FIGURE 8 Aggregated post-2011 intra-London movers between MSOAs, subject to a minimum disclosure threshold of 10

6 | DISCUSSION AND CONCLUSIONS

The LCRs enable a new UK-wide exploration of local outcomes arising from internal migration and residential mobility over the period 1997–2016, in unprecedented spatio-temporal detail. In particular, they make possible analysis of the origins and destinations of a large proportion of residential moves and facilitate detailed analysis of the processes of residential filtering consequent upon residential mobility.

The innovation of this research is to use UK nationwide lists of names and addresses, compiled from consumer and administrative sources, to link the successive addresses occupied by individuals and households over a 20-year period, with annual updates. The source data are individual level and recorded at very high geographic and temporal granularity, and triangulation with periodic Census sources and Mid-Year Population Estimates suggests that they are representative of underlying population movements and the structure of the internal migration system. The main caveat to this is that the period for which 'carry back' observations to plug gaps in the Registers is reduced for later years, with the implication that estimated distances of residential moves slightly increase if local matches are undetectable. The diversity of extant surnames and of naming practices otherwise renders full names an effective means of linking records over time and space, albeit that the procedures of linkage are necessarily complex and computationally intensive. This methodology might be adopted for different time periods and in different parts of the world in order to devise generalized and timely snapshots of
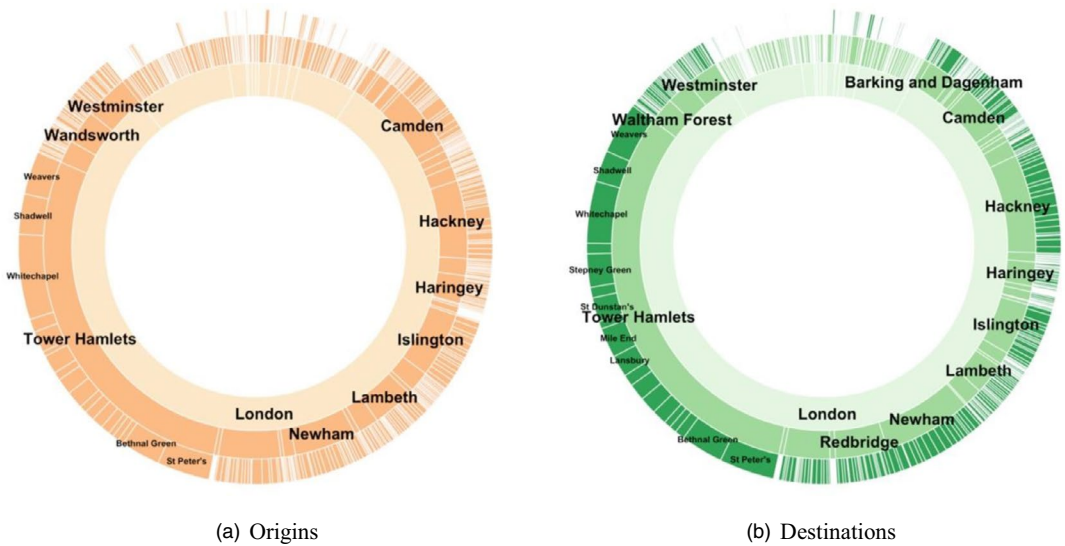
(a) Origins   (b) Destinations

FIGURE 9 (a) Origins of moves to Spitalfields and Banglatown Ward in East London; and (b) destinations of outward moves, 1997–2016. In each case the innermost ring (in lightest hue) identifies regions, the middle ring identifies constituent Districts and the outermost rings identifies constituent Electoral Wards
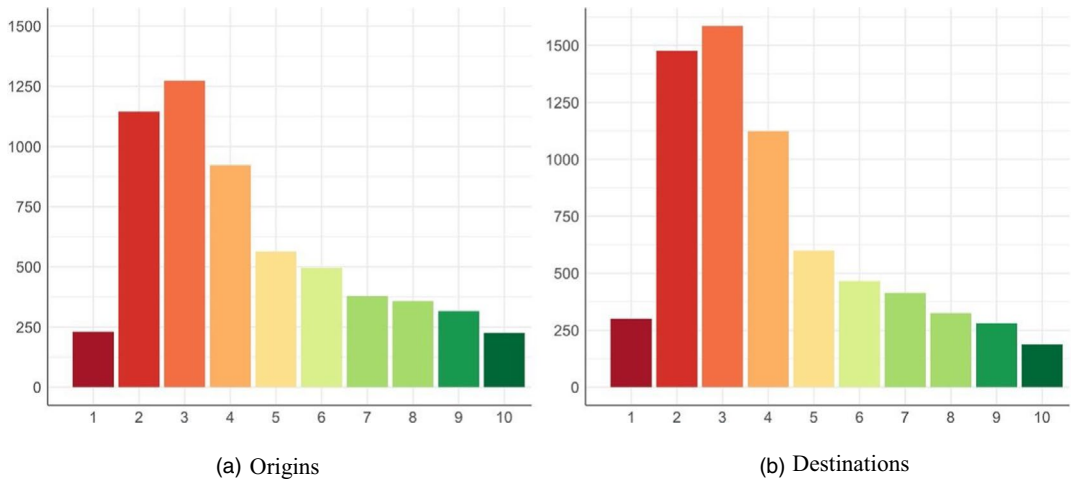

(a) Origins   (b) Destinations

FIGURE 10 (a) Origins of movers to Spitalfields and Banglatown Ward in East London by 2019 English Index of Multiple Deprivation (IMD) deciles; and (b) destinations of outward movers by 2019 English IMD deciles, 1997–2016. IMD deciles 1–10 range from the most to the least deprived IMD deciles

the outcomes of residential mobility. This approach is especially timely, in view of ever-growing interests in repurposing administrative and consumer data sources to supplement conventional population statistics derived from censuses and social surveys.

Conventional Census statistics provide comprehensive coverage of residential mobility throughout the United Kingdom, albeit only every 10 years. Patient Register Data Service and Personal

TABLE 2 Origins and destinations of movers to and from Spitalfields and Banglatown Ward in East London by 2019 English IMD deciles. Index of Multiple Deprivation (IMD) deciles 1–10 range from the most to the least deprived

IMD Decile	Percentage of movers moving into the area	Percentage of movers moving out of the area
1	3.9	4.6
2	18.8	20.5
3	20.4	22.8
4	14.9	16.1
5	9.6	9.2
6	8.6	7.2
7	7.1	6.2
8	6.4	5.0
9	5.8	4.8
10	4.5	3.6

Demographic data do much to fill this gap, but are only available to researchers at much coarser granularity. Subject to disclosure control procedures, the LCRs offer annual updates at any convenient level of spatial granularity, yet using the assumptions made in this paper only identify 63.4 % of the total number of adult residential movers when compared with the figures from the 2011 Census. Clearly, the analysis reported here does not detect all adult movers, but our results nevertheless break new ground by providing disaggregate annual updates that cannot be gleaned from any other statistical source. The relative importance of individual name text string matching, household composition and distance of move might also be changed to create versions of the register that require higher matches for particular types of moves, such as those over long distances, in applications where the risk of false positives can be managed. Further research, and the results of extending the run of the LCRs, should be conducted in order to examine the degree to which estimates for the later years can be refined, and our own research is presently seeking to extend the series to 2020 and beyond. Further research will also examine the degrees to which our matching procedures should be relaxed or strengthened in order to represent the residential mobility and migration behaviour of specific groups, such as students or the elderly.

These initial findings suggest a number of other directions for future research. First, the changes in data collection in recent years have potential implications for the generality of our migration model and we continue to undertake sensitivity analyses in order to ascertain whether bias is amplified or reduced over time. This work includes investigation of whether Zoopla rental listings data for individual properties can be used to apportion change to the rental market. Second, more sophisticated distance measures might be used in order to better represent the attenuating effect of distance upon moves, using 2011 Census data to calibrate the matching procedure (cf. Stillwell & Thomas, 2016). Third, processes of household formation and dissolution might be modelled in order to fill the gaps in origin and destination data that become increasingly apparent in the later years of the LCRs, and post 2016 updates might be used to assist in this process. Fourth, our own research is investigating the identifiable correspondences between names and ethnicity, age and gender, as well as related issues of household structure: we see this as a promising way of evaluating the accuracy of our models and of extending their usefulness into a number of important domains. For example, the materials developed here might be used to model changes in local incidence of multi-generation households from ethnic minorities, viewed in the context of the Covid-19 pandemic.

The last of these challenges speaks to a shift in modelling focus from aggregations to the human individual, extracting information from given and family names. Research using other data sources such as consolidated name and date of birth data alongside published records of baby names has modelled the age and sex characteristics of individuals (Lansley & Longley, 2016), while given and family name pairings have also been associated with census data in order to predict ethnicity (Kandt & Longley, 2018). This focus on prediction at the individual level offers the prospect of adding a range of probable individual and household covariates that are available only from censuses or (in more aggregate form) from health registers.

For now, the first full version of the LCR migration estimates ('CDRC Migration Model 1.0') provides new opportunities to better understand residential mobility patterns and their associated social implications. In developing data on residential movements in this paper we have restricted ourselves to providing brief illustrations of potential applications through linkage to conventional small area deprivation and geodemographic indicators. These illustrative examples are indicative of the potential full range of geodemographic indicators that might be used alongside LCR estimates. Subject to disclosure control procedures, the high spatial granularity of the LCRs makes it possible to contribute to analysis framed using any convenient geography. The data source makes possible comprehensive analysis of population flows and neighbourhood changes. In our own future research, we propose to investigate residential moves within and between different niche housing markets, in order to better understand the local, regional and national patterns of housing market differentiation across the United Kingdom today. In this paper, we have focused only upon identifying the timing and location of residential moves and have not addressed the ways in which the LCRs enhance understanding of the nature and timing of household moves, or the characteristics of households that remain sedentary over the life course. As such, we also propose to use individuals' names as tokens of identity that can be used to describe household structure as well as residential mobility history, and thus to develop scale free geographic representations of social mobility outcomes.

CODE DISCLOSURE

Our processing of the data falls under the public interest derogation for research under Article 89 of the General Data Protection Regulation. Although formed from proprietary component data sources, the resulting LCRs are available for *bona fide* research purposes on successful application by accredited safe researchers to the UK Economic and Social Research Council Consumer Data Research Centre (cdrc.ac.uk). This enables access to the code that has been used to link the individuals over for different years. 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.

ORCID

Justin T. van Dijk ⓘ https://orcid.org/0000-0001-5496-425X
Guy Lansley ⓘ https://orcid.org/0000-0002-3406-178X
Paul A. Longley ⓘ https://orcid.org/0000-0002-4727-6384

REFERENCES

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J. et al. (2002) Cross-national comparison of internal migration: Issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(3), 435–464. https://doi.org/10.1111/1467-985X.t01-1-00247

Champion, A.G. & Shuttleworth, I. (2017) Are people moving address less? An analysis of migration within England and Wales, 1971–2011, by distance of move. *Population, Space and Place*, 23(3), e2026. https://doi.org/10.1002/psp.2026

Clark, W.A. & Moore, E.G. (1980) *Residential mobility and public policy*. Beverly Hills: Sage Publications.

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*. London: The Electoral Commission. Available from: https://www.electoralcommission.org.uk/sites/default/files/pdf_file/The-December-2015-electoral-registers-in-Great-Britain-REPORT.pdf (accessed date September 26, 2019).

Electoral Commission. (2019) *Accuracy and completeness of the 2018 electoral registers in Great Britain*. London: The Electoral Commission. Available from: https://www.electoralcommission.org.uk/who-we-are-and-what-we-do/our-views-and-research/our-research/accuracy-and-completeness-electoral-registers/2019-report-accuracy-and-completeness-2018-electoral-registers-great-britain (accessed date September 26, 2019).

Coulter, R., Van Ham, M. & Findlay, A.M. (2016) Re-thinking residential mobility: Linking lives through time and space. *Progress in Human Geography*, 40, 352–374. https://doi.org/10.1177/0309132515575417

Fielding, A.J. (1992) Migration and social mobility: South East England as an escalator region. *Regional Studies*, 26(1), 1–15. https://doi.org/10.1080/00343409212331346741

Fielding, A.J. (2012) *Migration in Britain, paradoxes of the present, prospects for the future*. Cheltenham: Edward Elgar.

Finney, N. & Simpson, L. (2008) Internal migration and ethnic groups: Evidence for Britain from the 2001 Census. *Population, Space and Place*, 14(2), 63–83. https://doi.org/10.1002/psp.481

Gale, C.G., Singleton, A.D., Bates, A.G. & Longley, P.A. (2016) Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, 12, 1–27. https://doi.org/10.5311/JOSIS.2016.12.232

Gale, D. & Shapley, L.S. (1962) College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1), 9–15.

Hamnett, C. (1991) The blind men and the elephant: The explanation of gentrification. *Transactions of the institute of British Geographers*, 16(2), 173–189.

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. https://doi.org/10.1111/rssa.12315

Harvey, D. (1973) *Social justice and the city*. Athens: University of Georgia Press.

Hoinville, G. & Jowell, R. (1978) *Survey research practice*. London: Heinemann Educational Books.

House of Commons (2013) *The Private Rented Sector. First Report of Session 2013–14*. London: The Stationery Office Limited.

Kandt, J. & Longley, P.A. (2018) Ethnicity estimation using family naming practices. *PLoS One*, 13(8), e0201774. https://doi.org/10.1371/journal.pone.0201774

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

Lan, T., Kandt, J. & Longley, P.A. (2019) Geographic scales of residential segregation in English cities. *Urban Geography*, 41(1), 103–123. https://doi.org/10.1080/02723638.2019.1645554

Lansley, G., Li, W. & 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. https://doi.org/10.1111/rssa.12476

Lansley, G. & Longley, P.A. (2016) Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services*, 30, 271–278. (G Lansley, P A Longley) https://doi.org/10.1016/j.jretconser.2016.02.007

Lee, E.S. (1966) A theory of migration. *Demography*, 3(1), 47–57.

Lomax, N., Norman, P., Rees, P. & Stillwell, J. (2013) Subnational migration in the United Kingdom: Producing a consistent time series using a combination of available data and estimates. *Journal of Population Research*, 30, 265–288. https://doi.org/10.1007/s12546-013-9115-z

Lomax, N. & Stillwell, J. (2017) United Kingdom: Temporal change in internal migration. In: Champion, T., Cooke, T. & Shuttleworth, I. (Eds.) *Internal migration in the Developed World: Are we becoming less mobile?*. London: Routledge, pp. 120–146.

Lomax, N., Stillwell, J., Norman, P. & Rees, P. (2014) Internal migration in the United Kingdom: Analysis of an estimated inter-district time series, 2001–2011. *Applied Spatial Analysis*, 7, 25–45. https://doi.org/10.1007/s12061-013-9098-3

Longley, P.A., Cheshire, J. & Singleton, A. (2018) *Consumer Data Research*. London: UCL Press.

ONS. (2015). 2011 Census: Flow Data. [data collection]. UK Data Service. SN: 7713-2011 Special Migration Statistics - OA Level. https://doi.org/10.5255/UKDA-SN-7713-1

ONS. (2014) 2011 Census: Origin-destination statistics on migration for local authorities in the United Kingdom and on workplace for Output Areas and workplace zones, England and Wales. Available from: https://www.nomisweb.co.uk/census/2011/all_tables?release=UK.1 (accessed date December 18, 19). Office for National Statistics.

ONS. (2019a). Mid-year population estimates quality and methodology information. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/annualmidyearpopulationestimatesqmi (accessed July 07, 2019). Office for National Statistics.

ONS. (2019b). Local area migration indicators, UK. Available from: https://www.ONS.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/datasets/localareamigrationindicatorsunitedkingdom (accessed date December 12, 2019). Office for National Statistics.

ONS. (2020a). Population estimates for the UK, mid-2019: methods guide. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/methodologyguideformid2015ukpopulationestimatesenglandandwalesjune2016#appendix-4-historic-geography-changes-2009-to-2019 (accessed date September 09, 2020). Office for National Statistics.

ONS. (2020b). Internal migration: detailed estimates by origin and destination local authorities, age and sex. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/datasets/internalmigrationbyoriginanddestinationlocalauthoritiessexandsingleyearofagedetaileddestimatesdataset (accessed date September 14, 2020). Office for National Statistics.

Park, R.E. & Burgess, E.W. (1925) *The city: Suggestions for investigation of human behaviour in the urban environment*. Chicago: The University of Chicago Press.

Rabe, B. & Taylor, M. (2010) Residential mobility, quality of neighbourhood and life course events. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3), 531–555. https://doi.org/10.1111/j.1467-985X.2009.00626.x

Ravenstein, E.G. (1885) The laws of migration. *Journal of the Statistical Society of London*, 48(2), 167–235.

Rees, P., Stillwell, J., Convey, A. & Kupiszewski, M. (1996) *Population migration in the European Union*. Chichester: John Wiley and Sons Ltd.

Royal Mail. (2020) PAF. Available from: https://www.poweredbypaf.com/product/paf/ (accessed date September 09, 2020). Royal Mail.

Smith, D.P. & Denholm, J. (2006) *Studentification: A guide to challenges, opportunities and practice*. London: Universities UK.

Smith, D.P., Finney, N., Halfacree, K. & Walford, N. (2015) *Internal migration: Geographical perspectives and processes*. Farnham: Ashgate Limited.

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

Stapleton, C.M. (1980) Reformulation of the family life-cycle concept: Implications for Residential Mobility. *Environment and Planning A*, 12(10), 1103–1118.

Stillwell, J. & Thomas, M. (2016) How far do internal migrants really move? Demonstrating a new method for the estimation of intra-zonal distance. *Regional Studies, Regional Science*, 3(1), 28–47. https://doi.org/10.1080/21681376.2015.1109473

Ordnance Survey. (2020). Address data. Available from: https://www.ordnancesurvey.co.uk/business-government/address-data?gclid=Cj0KCQjw-uH6BRDQARIsAI3I-UdeqEO4wLCL2CzNWbHTq-A5QWhSCTSLiEtoJ_CYvOgNiIGD9qKTuRUaArFPEALw_wcB (accessed date September 09, 2020). Ordnance Survey.

Timms, D. (1975). *The urban mosaic: Towards a theory of residential differentiation*. Cambridge Geographical Studies No. 2. London: Cambridge University Press.

Travers, T., Tunstall, R., Whitehead, C. & Pruvot, S. (2007) *Population mobility and service provision: A report for London Councils*. London: London School of Economics.

Tyrrell, N. & Kraftl, P. (2015) Lifecourse and internal migration. In: Smith, D.P., Finney, N., Halfacree, K. & Walford, N. (Eds.) *Internal migration: Geographical perspectives and processes*. Farnham: Ashgate Limited.