

Detecting Address Uncertainty in Loyalty Card Data

Alyson Lloyd¹  · James Cheshire¹

Received: 5 October 2017 / Accepted: 5 January 2018
© The Author(s) 2018. This article is an open access publication

Abstract There is a fundamental need to better appreciate the dynamics and uncertainty of large consumer datasets, particularly if they are to be utilised to model social and geographical phenomena. This research, the first to utilise a major UK retailer’s loyalty card dataset, presents a novel data-driven approach for quantifying uncertainty in consumer addresses. Uncertain cases were identified through the linkage of locational and behavioural attributes and consumer mobility patterns recorded at a small area level. Such methods are not only important for the reliable adoption of large commercially generated datasets in research, but also for retailers if utilising this information to inform location-based marketing strategies. Results are contextualized with dynamics in the general population, demonstrating comparable relationships with Census migration patterns.

Keywords Big data · Retail · Spatial analysis · Uncertainty · Mobility

Introduction

Loyalty card data offer a typical example of a contemporary “Big Data” source, allowing compilation of behaviours that inform brand choices, household inventories, promotional impacts and long term behavioural patterns. In addition, customer meta-data such as age, gender and postcode are collected via the application processing system for a loyalty card, adding a dimension of demographic and spatial data that can be attributed to transactional behaviours. These novel forms of data capture a broad range of socio-spatial processes particularly since the postcode field provides an important means of linkage to conventional statistical geographic units and the data associated with them, such as existing national statistics.

✉ Alyson Lloyd
Alyson.lloyd.14@ucl.ac.uk

¹ Department of Geography, University College London, Pearson Building, Gower Street, London WC1E 6BT, UK

These data fall under the classic definition of Big Data, with its five most commonly recognised dimensions of volume, velocity, variety, veracity and value (Demchenko et al. 2013). Here we demonstrate the increasing importance of veracity; referring to issues of uncertainty such as the accuracy, trustworthiness, quality and credibility associated with these new forms of data (Schroeck et al. 2012). Uncertainties arise due to the nature of these data being produced as a by-product of alternative commercial agendas, rather than conforming to the rigours of more traditional approaches to data collection. Since they are often hard to obtain for academic research, this investigation offers a unique opportunity to examine ambiguities surrounding the veracity of such data and develop heuristics to address them. It is the first to do so for a major, national, consumer dataset in Great Britain.

Loyalty Card Data

Since their widespread adoption in the 1990s, and the recognition of the cost benefits of retaining rather than obtaining customers (Kotler 2002; Reichheld 2001), loyalty schemes have become extremely popular for both retailers and consumers. In their most basic form, these schemes involve awarding points according to how much a customer spends, which can then be redeemed as discounts on future purchases. Data are collected firstly through the process of application (i.e. either in store, or online), which commonly asks for demographic (age, gender) and address information. Secondly, customers are typically provided with a membership card which records their purchasing habits at the point of sale.

For retailers, these schemes can be defined as an integrated system of marketing actions that aim to make customers more loyal by offering monetary based rewards for loyal shopping behaviour (Yi and Jeon 2003). However, the data produced as a result has a wide range of uses. Commercial organisations utilise the transactional data to inform customer needs (Webber et al. 2015) and demographic information to provide a deeper understanding of individuals. In addition, the postcode information can be utilised for marketing strategies such as mail-based rewards or location-based events targeting, yet also for linkage to area classifications that categorise people based on the demographics of the neighbourhood in which they live (see for example, Gale et al. 2016). These classifications provide useful context about social structures and common characteristics between people and places and have been widely applied by businesses to infer lifestyles, social attitudes or identifying the best locations from which to serve and retain their customers.

Yet, this information also has a number of applications for social and geographical research. The volume of large-scale digital datasets being generated by these retail activities is unprecedented and offers a number of advantages for mining a broad spectrum of human patterns and processes. For example, these data are collected on an on-going basis, capturing both planned and mundane spatiotemporal dynamics on a diversity of scales, represent large populations and convey information about actual behaviour rather than stated preferences. In addition, the postcode field offers a valuable spatial reference for linkage, allowing a consistent basis for comparison across multiple forms of data. For instance, data can be appended to existing national statistics to infer relationships between population characteristics, neighbourhood types or inferred heritage (Webber et al. 2015). This permits interpretation of geodemographic behaviours and also aids understanding of potential biases in the data in terms of

general population characteristics. Furthermore, it allows for linkage to any alternative big data or independently collected behavioural databases, on a relatively granular scale. These novel forms of data may allow us to build a finer-grained and more accurate description of societal phenomenon than any metric previously utilised, which has been implicated in providing a valuable source of population intelligence if aiming to move away from the ten yearly Census approach through the integration of novel forms of population and social statistics (Stillwell 2016).

However, these data also present a unique set of challenges that need to be overcome if their potential is to be realised (Mayer-Schönberger and Cukier 2013). In the case of loyalty card data, the accuracy of customer metadata attributes are entirely dependent on human input when signing up to a scheme. Address attributes are further dependent on the motivation to update this information in the event of a location change. Whilst data cleaning regimes are well established to identify spelling errors or syntax mistakes at the data input stage, the accuracy of addresses may be substantially harder to isolate, particularly because they are dynamic. In the twenty-first century, places of residence change more frequently (Van der Klis and Karsten 2009) and populations are moving in more dynamic, complex and traceable ways than ever before (Sheller 2011). Data pertaining to changes in residence are seldom able to be captured by traditional methods, however, the 2011 Census estimated that 7.5 million people changed address within the prior to the Census. Recent research attempting to identify annual population change through novel forms of data, such as consumer registers (Lansley et al. 2017), also estimate a similar magnitude of migration. It is therefore critically important that we begin by assessing the accuracy of the spatial information provided by novel forms of data, to avoid obscuring, rather than revealing social and spatial processes (Graham and Shelton 2013).

Research Potential

In order to make best use of large consumer datasets in the social and geographic sciences, there is a need to develop suitable heuristics to both assess and improve data veracity. Miller and Goodchild (2015) outline two ways of approaching this: either by restricting the assumptions and generalisations drawn from analyses, or by attempting to clean and verify the data. One solution regarding the latter is to adopt the *knowledge solution*, which postulates that we may draw on existing theory to ascertain whether or not purported fact is false, or likely to be false. Consequently, we may attempt to create informed assumptions based on what has been termed ‘logical consistency’ in the literature of geographic information science (i.e., Gupta and Morrison 2013), and whether a purported fact is consistent with what is already known about the geographic world. Yet, limited research has been able to apply this in practice, due to these data being substantially hard to access outside of the commercial settings in which they are often created.

Here we present insights from our work with a large retailer’s loyalty card dataset obtained by the ESRC Consumer Data Research Centre. It offers a unique opportunity to study the dynamics of and inaccuracies within a commercial dataset. Initial spatial analyses of customer postcodes and transactional behaviours suggested instances of consumer behaviour that may deviate from expectations. For example, irregular travel patterns such as routine transactions at a store a long distance from a stated home address. Therefore, our primary research objective was to develop a means of quantifying potentially inaccurate address information in the absence of reference data. To

achieve this, data-driven heuristics were constructed that utilised customer transactions to estimate the credibility of their address information, by drawing on current knowledge and theory of spatial behaviour. We then demonstrate how this information may be further utilised to estimate the new locations of these individuals. Our second research objective was to explore and attempt to contextualise these findings in relation to existing population statistics. This aimed to provide a pragmatic means of validation since the postcodes offered a basis by which each customer record could be linked to conventional statistical geographic units and then compared to and augmented with existing national statistics.

Data

Loyalty card data were provided by a major high street retailer (HSR) with a national network of stores, accounting for every transaction that had occurred within Great Britain (GB) between April 2012 and March 2014. The loyalty customer metadata consisted of demographic information provided at sign up, including gender, date of birth and postcode. Transactional data included store of purchase, product type, amount spent and a timestamp for over 400 million records. Store data were provided for all stores within GB including point location and metadata such as size and type. The details provided below, and the reporting of results, are necessarily constrained by strict controls to prevent the disclosure of either personal or commercial information.

Data Cleaning & Pre-Processing

A number of measures were taken to clean these data. Firstly, transactional volumes varied substantially between customers. Overall, 0.66% of customers had never transacted, 2.6% had transacted only once and 12.1% less than 10 times over the two financial years. For the purpose of this analysis, active customers were defined as those that had transacted more than five times within the last financial year (April 2013–March 2014). This threshold was selected with the intention of eliminating inactive customers whilst also retaining the maximum possible sample size. Secondly, 0.2% of the accounts had missing or invalid postcodes and were therefore excluded from analysis. These stages resulted in a sample of approximately 15.8 million customer accounts.

Cleaning measures were also applied to customer metadata as these were utilised for interpreting characteristics post-analysis. Converting the date of birth field to age revealed an abnormal range, with a minimum of 2 and a maximum of 359. This could have been due to customer input error. Therefore, customers were selected between the ages of 16–85, since this range captures the majority of the adult population (still, errors falling inside of the normal human age range would not be identifiable in these data). Furthermore, 0.4% of gender attributes were withheld or missing. These stages removed 20.9% of the active customer database leaving a sample of approximately 12.5 million accounts with both sufficient volumes of transactional data and complete metadata attributes.

Finally, due to the sensitive nature of the dataset, customer locations were aggregated from postcode level to small area Census derived geographies. For the analysis of GB, this included Middle Layer Super Output Areas (MSOA) for England and Wales

and Intermediate Zones (IZ) for Scotland. MSOAs are the most aggregate level of geography derived from 2011 Census variables and IZs the equivalent geographies for Scotland. Both are built up from lower level Output Areas (OA), which describe small area characteristics. Whilst we acknowledge that aggregation is best avoided, we deemed it necessary here for disclosure purposes, the implementation of our group level model and to facilitate linkage to national statistics from the Census.

There are 7194 MSOAs in England and Wales with a minimum of 5000 residents per area and an average population of 7500. In Scotland, there are 1235 IZ's with a minimum of 2500 and an average of 4000 residents per area. Figure 1 shows the distribution of cardholders per MSOA and IZ across GB, taking into account underlying population volumes as estimated by the 2011 Census. For simplicity, we refer to the collective use of these geographies as MSOA's for the remainder of this paper.

Cardholders demonstrated a national extent of coverage, with an average of 2248 customers per MSOA.

Detecting Irregular Travel Patterns

The first stage was to perform exploratory analyses to identify potentially irregular travel patterns in the data. Using abductive reasoning, interactions between customer addresses and store visiting behaviours were evaluated based on existing knowledge of spatial behaviour and human mobility. For example, despite the increasing complexity of mobility systems, assumptions can be made as to what observable patterns in the data were likely to be false. Notwithstanding increased ease of mobility due to available transportation (Sheller and Urry 2006), daily mobility is still characterised by a deep-rooted regularity (i.e. Gonzalez et al. 2008; Song et al. 2010). A home location can be considered as one of the moorings that define spatial movement (i.e. journeys are likely to begin from and end at home) and the location in which one lives therefore poses spatial and temporal constraints that affect the daily movement patterns and lifestyle of an individual (Ellegård and Vilhelmson 2004). Boundaries will be constrained by physical barriers and interactions between locations declines with the increasing distance, time, and cost between them (although it may be positively associated with the amount of activity at each location; Isard 1956). In addition, a home boundary can be seen to represent an area in which the majority of time is spent and movement can be interpreted as when changes in the 'spatial points of reference' of a home mooring occur (Behr and Gober 1982).

Figure 2a illustrates an example of the travel flows from active customers' origin MSOA to their most frequently visited store, for one store type that primarily serves the local surrounding communities. These patterns indicated potential instances of deviation from expectations based our knowledge of spatial behaviour. For example, it is unlikely that customers frequently travel substantially long distances (i.e. from Scotland to the South coast of England) to visit store destinations. The spatio-temporal patterns of individual customers identified further examples of ambiguity. Figure 2b shows an example of this for a customer registered to an MSOA in Northern England.

Whilst transactional behaviours appeared logically consistent with their postcode at the beginning of records (i.e. within a local store network), their network of stores

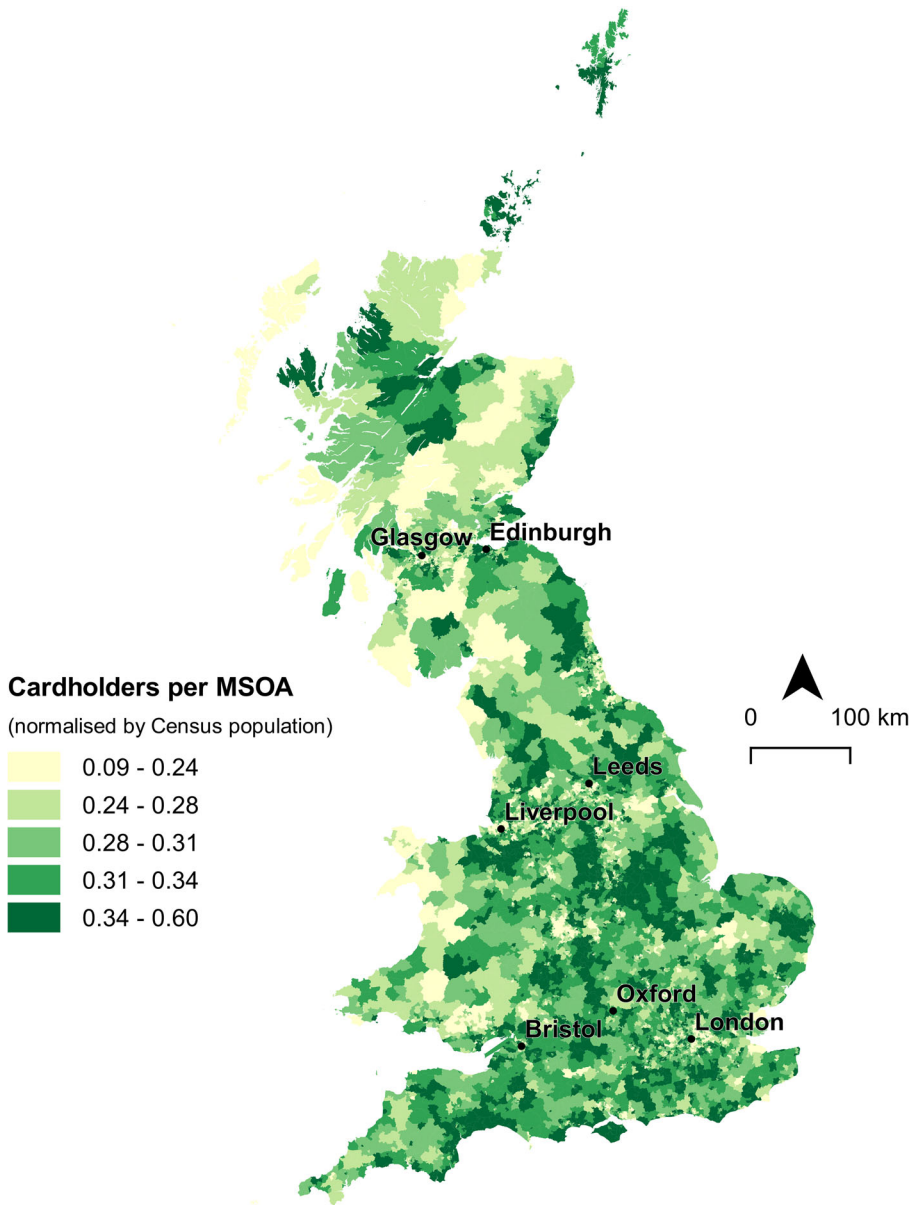


Fig. 1 Cardholders per MSOA in Great Britain normalised by census population estimates (classified by quantiles)

exhibited a permanent shift to inconsistent geographical areas after certain time periods (Oxfordshire in 2013, Bristol in 2014). This could typically be defined as an absence of further transactions within their initial network for the remainder of their recorded activity. These observations suggested a change in location that was not reflected by the postcode information provided in the data and therefore a need to investigate the extent to which these uncertain cases existed across the customer database.

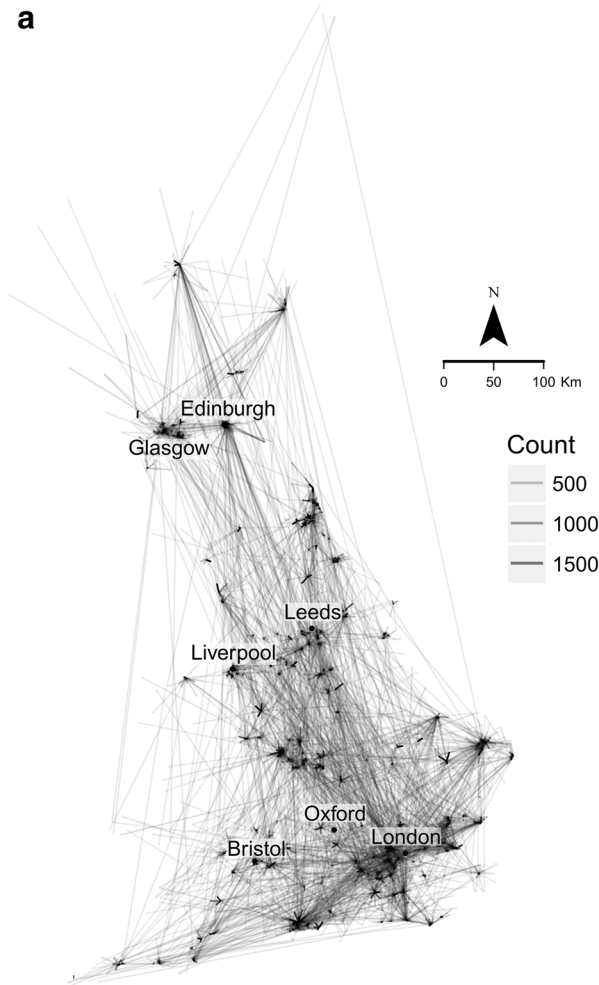


Fig. 2 **a** Flows from customer origin MSOA's to their most frequently visited store, for one store type. **b** An example of ambiguous spatio-temporal transactional behaviour

Method

To quantify irregular travel patterns, a data-driven methodology was constructed by drawing on knowledge and theory from multidisciplinary domains. The concept was applied that an individual will be anchored to their immediate geographical neighbourhood to some extent, and it is therefore possible to use store locations as spatial points of reference to infer the most likely patterns for different residential areas in the database. This could then be applied to interpret behaviours that were not consistent with residential areas, or when changes in these spatial points of reference occurred. The address information provided home anchor points, and the method intended to define the importance of each store location to different home anchor points across Great Britain. Figure 3 shows an overview of the process applied.



Fig. 2 (continued)

The method involved creating a trip distribution (or destination choice) matrix from the data, which described the number of trips going from each area of residence (origin) to store location (destination). This allowed analysis of interactions between all origins and destinations in the database. Whilst this methodology is traditionally applied for predictive purposes in travel behaviour research, the principles could be applied in this data-driven context. Secondly, theory from retail centre catchment research was utilised, which typically involves the selection of one or more threshold values that represent the proportion of customers likely to patronise a certain store or retail centre (Dolega et al. 2016). This concept could be applied here to define thresholds that categorised trip distributions per residential area across Great Britain. Finally, an algorithm was designed that utilised this catchment information to assess the frequency at which individual customers performed irregular travel patterns throughout their transactional histories.

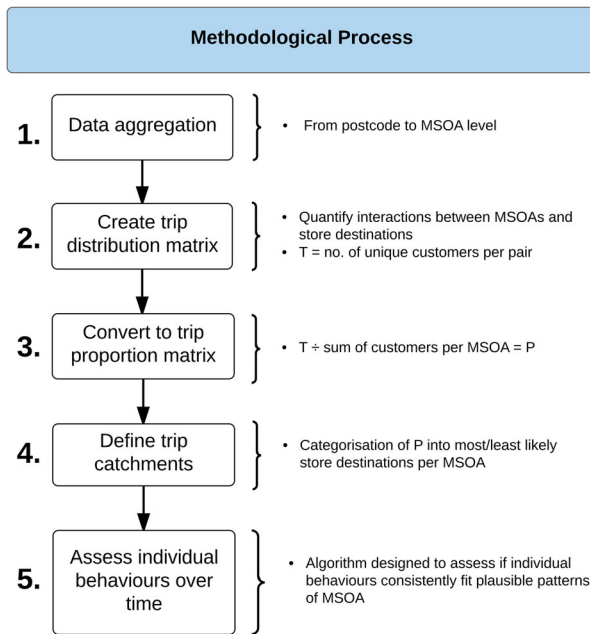


Fig. 3 Overview of methodological process

Deriving Origin-Destination Matrices

To create an O-D matrix, customer origins were aggregated to the MSOA level, as this was the lowest level of aggregation able to produce large enough population groups to distinguish interpretable distributions. More granular spatial units led to noisy and unbalanced trip representations, an issue that is alleviated when zone sizes become larger (Chung and Kuwahara 2007; Zandbergen 2009). For example, some areas at the more granular OA level contained few, if any, customers, which would not be sufficient to derive normative behaviours for an area. Conversely, more aggregate units, such as Local Authorities, reduced the sensitivity of the analyses to changes in store networks that were identifiable at the MSOA level.

The O-D matrix was created by obtaining all MSOA to store journeys that had occurred within these data. This resulted in 5,833,028 unique combinations. For each combination, the number of customers that had performed a journey (T) was obtained. This resulted in a matrix describing the frequency of customers that had performed each pair. Therefore, each unique trip a customer had performed was recorded. Table 1 shows an example of the matrix format. Subsequently, trip distributions were converted into trip proportions, by dividing T values by their O sum (total number of customers per MSOA). This was to interpret trips in relation to the differing volumes of customers per area.

The O-D proportion matrix allowed interpretation of the relative frequency with which the pairs were performed per MSOA. On average, 688 unique stores were visited per MSOA, with a minimum of 111 and a maximum of 1248. Individual customers visited an average of 11 different stores over the 2-year period.

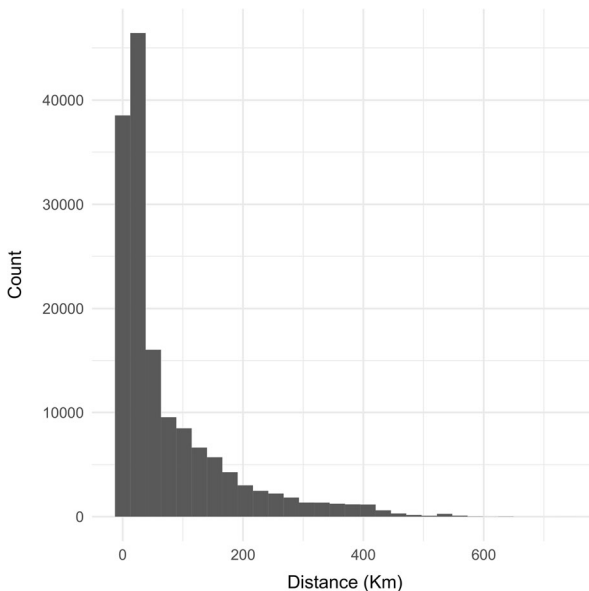
Table 1 Example trip distribution matrix

MSOA	Store ID					Sum
	1	2	3	4	5	
E02000001	T ₁	T ₂	T ₃	T ₄	T ₅	O ₁
E02000002	T ₆	T ₇	T ₈	T ₉	T ₁₀	O ₂
E02000003	T ₁₁	T ₁₂	T ₁₃	T ₁₄	T ₁₅	O ₃
E02000004	T ₁₆	T ₁₇	T ₁₈	T ₁₉	T ₂₀	O ₄
E02000005	T ₂₁	T ₂₂	T ₂₃	T ₂₄	T ₂₅	O ₅
Sum	D ₁	D ₂	D ₃	D ₄	D ₅	

Threshold Selection and Distance Constraints

Thresholds were defined to categorise trip distributions and select destinations that fell above or below these. This aimed to identify the point at which stores no longer constituted regularly patronised destinations for an MSOA. Stores above these thresholds were defined as ‘primary’ destinations (i.e. the highest 40% of visited locations) and those below as ‘non-primary’ destinations. Thresholds were calculated individually for each MSOA to reflect the unique dynamics of each.

Trip distribution tails for all MSOAs were positively skewed, a feature that can be explained by the behavioural dynamics of the data in context. For example, relatively few destinations are highly patronised by customers of a given area, largely due to effects of proximity. Figure 4 illustrates an example of trip distributions and (Euclidean) distances travelled, calculated from MSOA centroids to store locations.

**Fig. 4** Example trip distribution tail and distance travelled

Due to these distance decay dynamics, a plateau in patronage could be observed following the most highly patronised group of stores, which was used to select suitable thresholds. Across MSOAs, these ranged between 31% and 55%. Trips falling outside of these thresholds primarily described less patronised destinations, such as those far in proximity. A higher percentage threshold for an MSOA typically reflected an area with a denser store network, as more nearby store locations produced wider patronage patterns. Lower thresholds typically reflected more rural MSOAs where fewer locations were highly patronised due to low store competition. On average, there were 37 primary stores per MSOA, a minimum of 2 and a maximum of 64.

A final methodological step was to introduce a distance constraint to these store categorisations in order to alleviate issues identified in the data. For example, distances were a key indicator of whether or not behaviours could be deemed logically consistent with an area. Yet, in some instances, destinations within relatively close proximity to a residential area were categorised as non-primary and those very far as primary. Exploratory analysis suggested that this was likely due to two factors; firstly, the limited time span of the data producing small counts for relatively local stores (therefore a reflection of lack of data rather than true customer behaviour) and secondly, destinations that generally exhibited high counts across many areas, but were unlikely to be frequently patronised store (such as Oxford Street in central London).

To create a suitable constraint, the average Euclidean distance travelled from MSOA centroids to primary stores was calculated for each area. It is acknowledged that alternative (i.e. network) measures would provide more accurate portrayals of travel distances, however, the aim of including this constraint was to isolate cases that were significantly above or below normative behaviour in relation to overall Euclidean distances per area (i.e. given that this may vary considerably between rural and urban areas), rather than to quantify precise travel behaviours. These irregular instances were minimal, however customers who were identified in the succeeding analysis based on these re-categorised stores were flagged in the output for further investigation.

Implementation

The resulting output from these steps was a list of primary stores per MSOA that could be utilised to interpret customers' spatial behaviour. Two fundamental patterns of irregular behaviour needed to be considered when implementing this information. Firstly, for the purpose of this analysis, *address errors* were defined as customers who had never transacted at a primary store location. Secondly, *address changes* were defined as customers who demonstrated a change in patronage behaviour within the time span of the data. These could typically be identified as a permanent shift to a unique network of stores that was outside of their registered area's primary destinations. The algorithm identified 'permanent shifts' by returning customers for which no further occurrences of primary store transactions were recorded after certain timestamps in their transactional histories. A constraint was necessary to avoid returning customers who may have exhibited deviating behaviours in only their most recent transactions (i.e. due to the data only capturing a two-year period, this would not provide enough information to classify a location change). All identified cases were therefore subject to a time constraint of 1 month to ascertain if the person had spent sufficient time transacting in the new area to be defined as a permanent change. This time period

was chosen based on a trade-off between accounting for the relatively high average transaction interval of the active customer sample (12 days, yet this was positively skewed) and avoiding omitting identification of those with more frequent transactions. Figure 5 gives an overview of the algorithm designed to identify these cases.

Account numbers were individually selected alongside their time-ordered transactional histories. Catchment thresholds were then obtained for a customer's MSOA and their transactions categorised. If an *address error* was not identified at this stage, accounts were assessed for an *address change* by analysis of their time-ordered transactions for identification of a change in store networks (i.e. no further occurrences of primary transactions). If a change point was detected and patterns exhibited a new network of stores, account numbers were appended as an *address change* and a

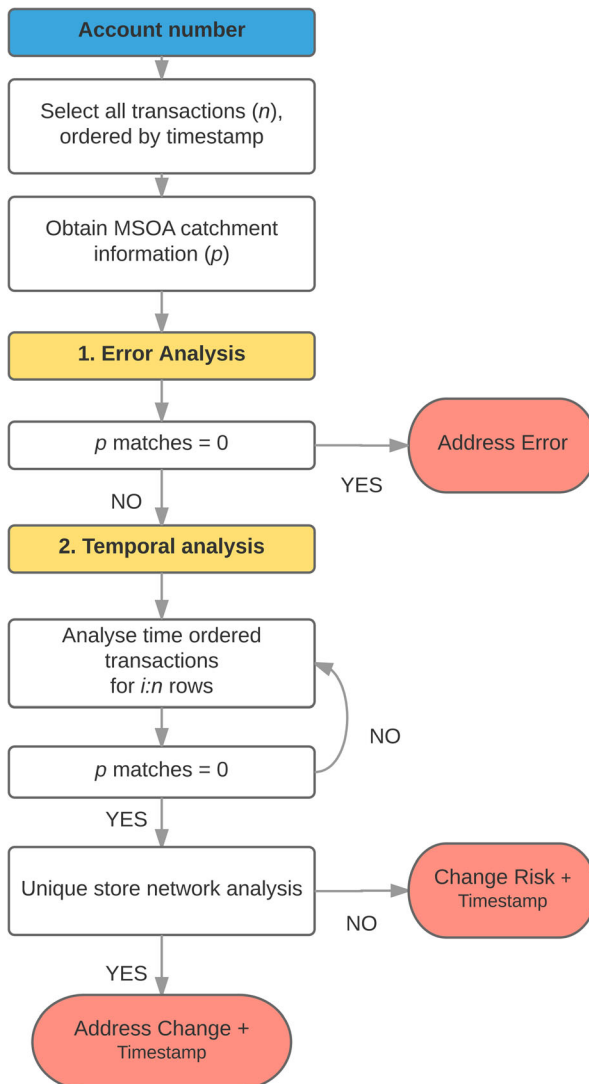


Fig. 5 Overview of the algorithm function for detecting uncertain addresses

timestamp of this change recorded. If at least one store had been previously visited, the account was appended as *change risk*.

Comparison to Census Statistics

To contextualise the characteristics of flagged customers, comparative analyses were conducted with a number of Census outputs. Due to available statistics, only England and Wales data were utilised. Firstly, accounts were compared across 2011 Output Area Classification (OAC) groups. This classification, derived from Census variables, describes geo-demographic population characteristics across 8 supergroups, 26 groups and 76 subgroups at the OA level. Customer postcodes were thus aggregated to the OA level and frequency of estimated moves compared across the groups. Counts were normalised by total customers per group in the database, to account for underlying variation in volumes.

To contextualize the estimated moves in terms of migration patterns, results were also compared to Census migration statistics, which describe moves that occurred between MSOAs in England and Wales between 2010 and 2011. Events captured using the loyalty card data were selected between 2013 and 2014, as this were the only available full year of data comparable to Census dates. At this stage we are seeking to establish whether the moves identified within the loyalty card data follow broadly expected flows, rather than use them to make broader inferences across the population. Card estimates were adjusted to reflect Census population volumes per MSOA by creating a coefficient (total Census population per MSOA divided by total card population per MSOA). Card migration counts were then multiplied by this value. Relationships between migration estimates were measured using Spearman's rank correlation.

Results

Implementation of this algorithm returned a total of 447,141 accounts – approximately 3.6% of the analysis sample. This comprised of 213,395 estimated address errors and 233,748 address changes. Whilst it is unlikely that the largest proportion of customers provided incorrect address information at sign up, it is possible that an address change occurred before the time period of the available data. In addition, a large proportion of the address change customers could only be categorised as change risk due to lack of available transactional data (45% of risk customers exhibited less than 10 transactions in comparison to 6% of those conclusively categorised), resulting in final analysis samples of 213,395 address errors and 169,943 address changes. Analysis of spend characteristics suggested that these customers had not deserted the card scheme, with an average spend of £344, 37 transactions and 64 products over the 2 financial years.

Figure 6 shows a comparison of the travel flows from customers' origin MSOA to their most frequently visited store (showing one store type), using a) the un-cleaned data and b) the cleaned data.

Applying this cleaning method produced flows that were consistent with expectations for this store type, which primarily serves local surrounding communities. In comparison to the raw data, we speculate that the majority of patterns that were inconsistent with our existing knowledge of spatial behaviour could be identified (sample reduction between these figures was 1798 customers, or 2.7%).

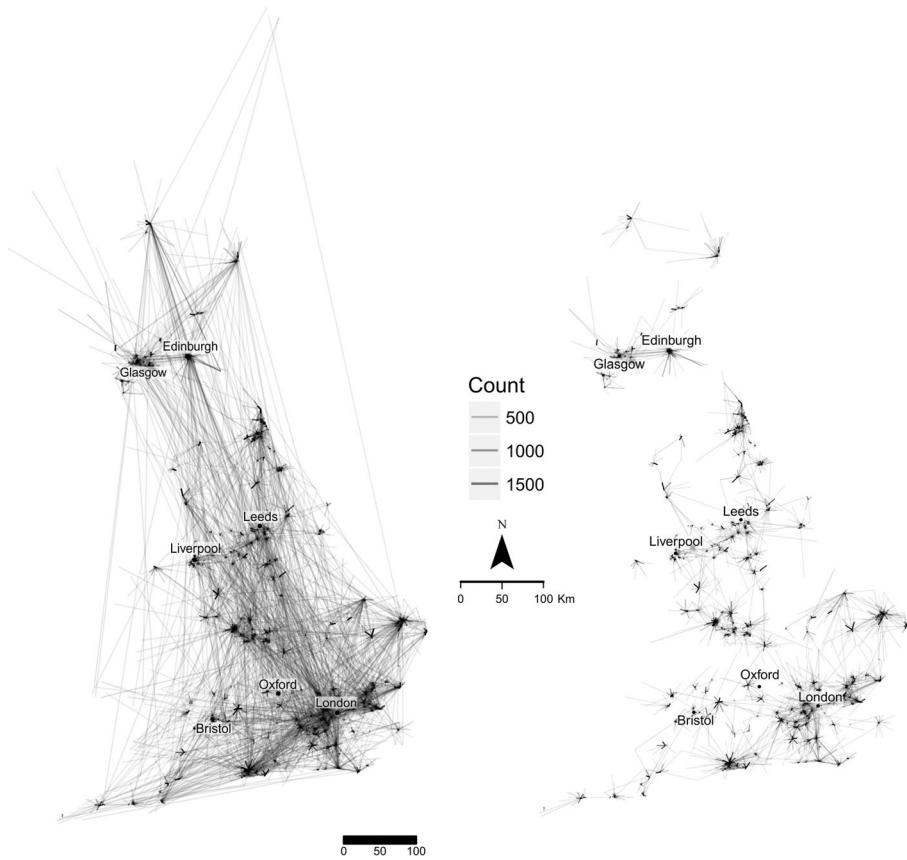


Fig. 6 Cleaned data flows comparison for one store type, across great Britain

Analysis of the demographic attributes of these customers was possible using the metadata provided at sign up. Figure 7 shows the age distributions of customers recorded at the time of their change point, normalised by total customers per age group.

This suggested that customers flagged as exhibiting a location change were considerably skewed towards younger cohorts, particularly between the ages of 18–20. This could be indicative of a more transient group with a greater risk of failing to update their address information with more frequent moves. Figure 8 shows the comparison of these customers to the OAC at the super group and group level. This suggested that the largest proportion of these customers were registered to cosmopolitan areas and in particular, student populated neighbourhoods.

Analysis at the subgroup level indicated that the highest proportions were registered to: student digs, student communal living, students and commuters and multi-cultural student neighbourhood groups. Higher proportions amongst other subgroups indicated those also less likely to have a long-term stable location, such as young families and students (Ethnicity Central) and private renting new arrivals (Multi-cultural Metropolitans). Supergroups with the lowest proportions of flagged customers included suburbanites and hard pressed living. These trends suggest that the method was able to highlight areas that we expect to have more

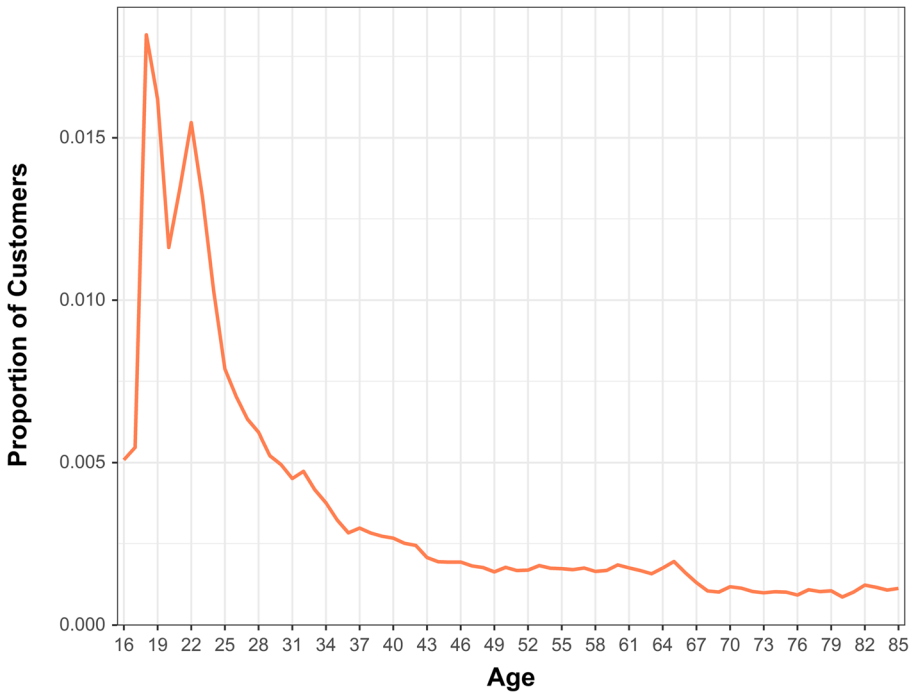


Fig. 7 Ages recorded at time of estimated change point, normalized by total customers per year of age

transient residents, whilst also providing insight into the specific characteristics of customers where inaccurate address attributes may be more prevalent in these data.

Finally, correlation of migration events between the datasets showed a moderately strong positive correlation ($\rho = 0.53, p < 2.2e-16$). However, in light of the previous observations, it was likely that the card migration was skewed by the amount of student migration captured by the method. Correlation with Census student migration estimates at the local authority level indicated a strong positive relationship of 0.87, $p < 2.2e-6$ (see Fig. 9).

We speculate that the ability of the method to flag high levels of student migration could be due to the highly transient residential nature of this demographic group. However, it is also possible that the method is better equipped at flagging long distance moves, as these cases will fundamentally exhibit a more discernible change in store network. It follows that an unavoidable limitation is that this method is not able to detect location changes that do not cause a modification in store visiting behaviours. This may limit the extent to which we are able to detect close proximity moves, and will also be constrained by the geography of the store locations, as these are the only spatial point of reference for observing irregular behaviours. This may affect the method where store networks are less dense, for example, more rural areas.

Estimating Relocations

Having quantified store-visiting behaviour at a small area level, it was possible to estimate potential areas of relocation for flagged customers by identifying the

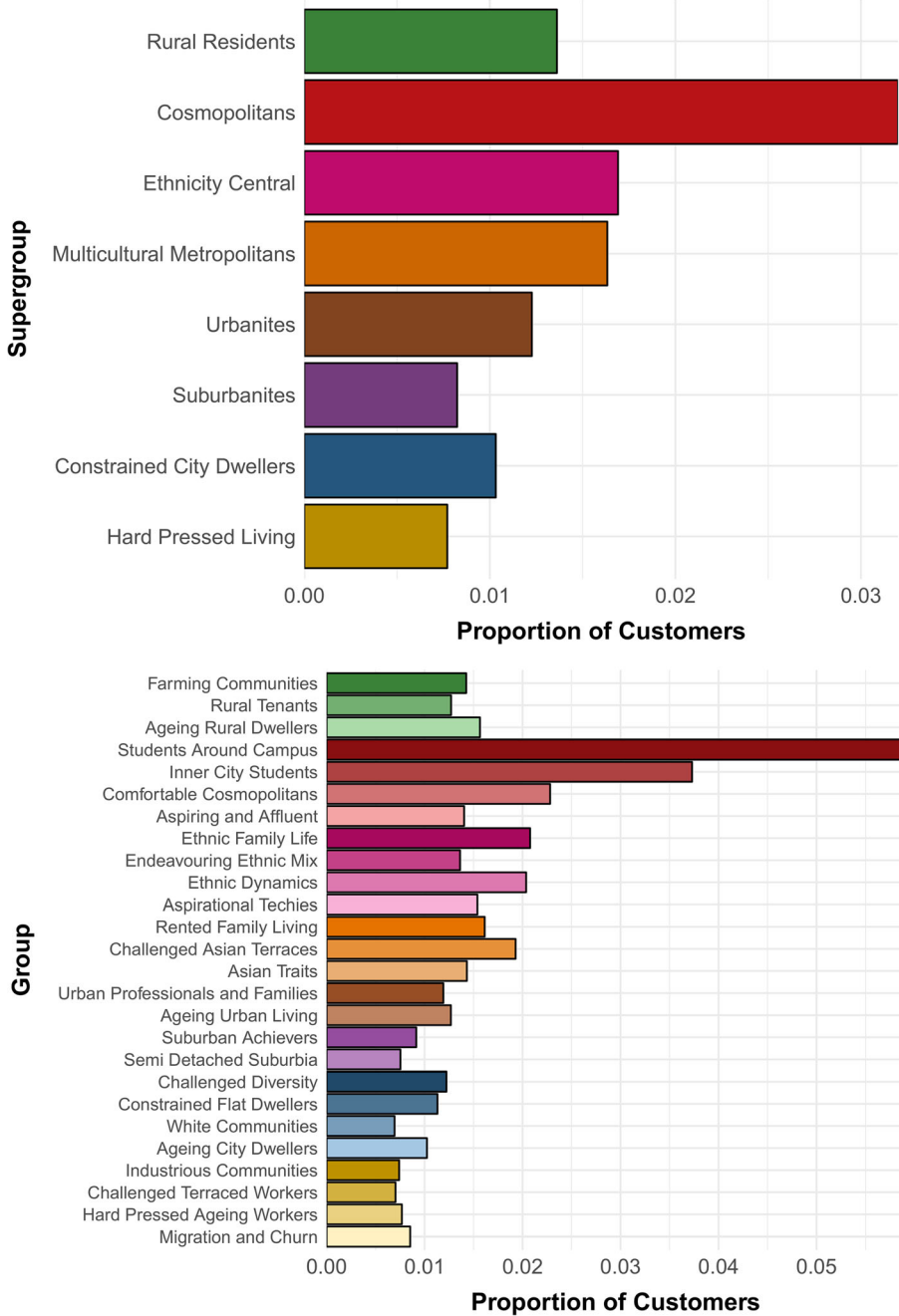


Fig. 8 Frequency distribution of migration counts across OAC a) supergroups and, b) groups

localities that their new store visiting behaviours may be consistent with. For example, providing enough transactional data were available, behaviours could be matched with areas for which these stores were a primary destination. Conducting

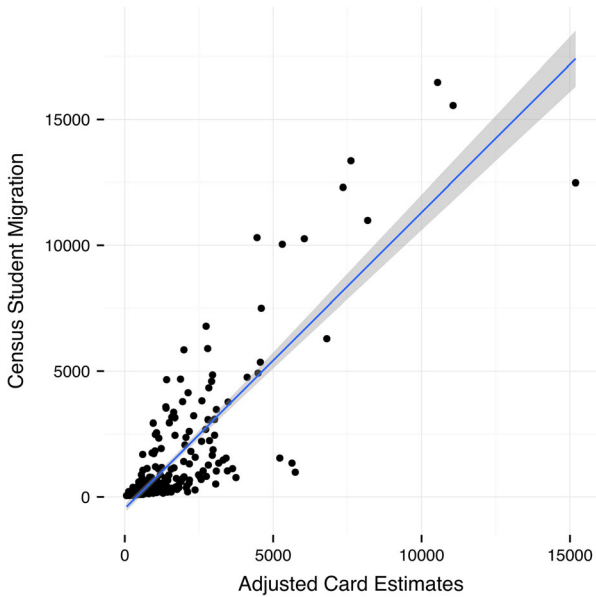


Fig. 9 Scatterplot of correlation between card migration and census student migration estimates

these analyses aimed to understand the extent to which we could extract further insights from this data-driven methodology. Results were then contextualised using Census migration statistics.

Method

In order to estimate potential relocations, an algorithm was designed to match the new store networks to the primary store networks of different small areas. This was conducted at both the MSOA and also Local Authority (LA) level to assess the granularity at which such analyses could be accurately implemented. The algorithm selected transactional histories for flagged customers, using all transactions for address error customers and transactions succeeding a change point for address change customers. Stores visited by these individuals were then matched with primary stores identified per MSOA or LA using simple pattern matching. Primary stores per LA were computed by aggregating this information from the MSOA level (created from the computations in Section 3). Per LA, there were an average of 83 primary stores and a maximum of 262. Outputs included the number of total area matches and the number of primary stores that were matched from each area.

Results were contextualised with existing migration statistics captured by the 2011 Census Origin-Destination data (describing the origins and destinations of moves that occurred in England and Wales between 2010 and 2011). These analyses similarly aimed to investigate whether the moves identified within the loyalty card data followed broadly expected patterns of migration, rather than use them to make inferences across the population.

Table 2 Average data required for relocation estimation accuracy

LA Matches	Average Transactions	Average Stores
1	180	23
2	56	13
3	23	2
4	6	1

Results

Due to the segment of customers being an actively transacting group, 97.45% could be accurately matched to one LA. Matches at the MSOA level were able to match 13,000 address errors and 4290 address change customers to one small area. These customers had an average of 121 transactions and a mean of 24 stores, suggesting that, unsurprisingly, customers with a larger overall store network could be matched with finer granularity. However, due to these relatively low match rates, only LA level data were considered for further analysis. The maximum number of matches obtained for any individual was 4, but these results depended largely on the amount of data available for a customer. Table 2 shows the average transactions and available store data per customer required to suitably match individuals to their relocated areas.

Results indicated that whilst there may be a relationship with inter-regional flows, the card data substantially underestimated intra-regional migration from what we may expect. This is likely due to the limitations acknowledged in Section 3.2, such as students dominating the sample or the inability of the method to highlight local migration patterns. Figure 10 demonstrates a comparison of inter-regional flows using the card and Census data.

Spearman’s correlation indicated a moderately strong positive relationship between inter regional flows estimated using the card and Census data ($\rho = 0.691207, p < 2.2e-16$).

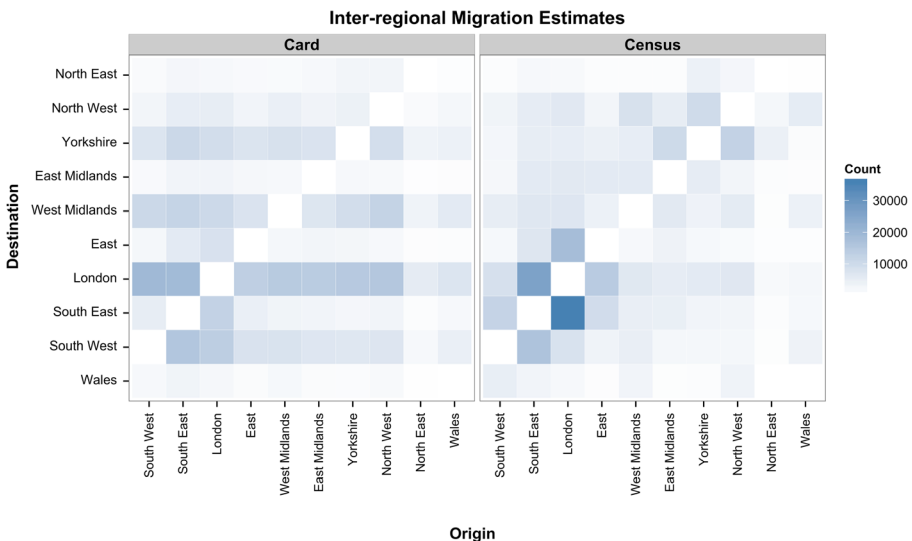


Fig. 10 Inter-regional migration using loyalty card data and census estimates

Comparable proportions of movement between regions can be observed, although the card data may over estimate moves between areas in some cases (i.e. migration to London). This could be due to the card data covering a different and longer temporal period.

Discussion

These analyses utilised knowledge and theory from multi-disciplinary domains to identify inaccurate addresses based on customers' stated home location and their transactional histories in a large consumer dataset. Results suggested that a segment of the population within postcode referenced consumer data might be unrepresentative of a current place of residence. Despite being unable to unequivocally verify these findings due to the absence of reference data, comparisons with existing national statistics suggested that the method was able highlight customers that we may expect to have more transient residential locations and demonstrated comparable trends to inter-regional migration. These insights provide promising prospects for the use of large consumer datasets in social science research. For example, they highlight the ability to identify potential uncertainties using the data alone, where linkage to reference data is not possible. In doing this, we are able to expose data veracity issues inherent in consumer data and propose heuristics by which we can attempt to address them. Such methods may be adopted in other relevant systems or settings in an attempt to clean spurious patterns, for example, in any dataset where home locations and behavioural points of reference are accessible.

Whilst further investigations are needed, the observed relationships with Census data present a promising example of the potential to use alternative datasets as a means of creating more frequent indicators of migration. This could contribute to filling the data void that occurs during inter-censal periods. For example, objectives for the Census 2021 and beyond suggest the integration of more address-level intelligence from administrative, commercial and open-data sources to help estimate non-response rates and to ultimately move away from the 10 yearly Census approach (Stillwell 2016). Key to this will be the ability to link data efficiently and accurately, yet we highlight here how preliminary data treatment is necessary to ensure the veracity of commercial data being integrated. Nevertheless, the methods presented here show an example of how we can attempt to mediate such effects, whilst also demonstrating a means of highlighting addresses, areas and specific population characteristics that may be more transient in nature. This information could also facilitate targeting of non-responding households.

This approach may also be of interest to other consumer data collectors and users who are operating reliant on consumers keeping up to date address records. For example, results provide insight into the extent of customers who may no longer live at their stated address and will therefore no longer receive their mail-based rewards, or be correctly identified for location-based targeting efforts. This could have negative impacts on proceeding loyalty behaviour and if aiming to distribute a limited number of offers, these errors could hinder the impact of such campaigns. Understanding the demographic and geo-demographic characteristics of the customers most at risk of these uncertainties could help to mitigate these negative effects.

We are, however, keen to recognise a number of limitations to our approach. Firstly, as outlined in Section 3.2, this method is unable to detect changes in location that do not cause a modification in store visiting behaviour, which likely limits the identification of

local migration, particularly in areas with less dense store networks. This study therefore highlights important considerations for the adoption of such data as indicators of population and social statistics, primarily that analyses are heavily dependent on the data available, which limits the scope of insights than can be derived. In this case, the dynamics and time-span of these data predispose our heuristics to be best suited to extracting moves over longer distances. Secondly, due to these data being static and historical in nature, there were a number of accounts flagged as showing deviation from normal behaviour, yet were unable to be conclusively classified due to a lack of transactional data. Since the data cover a two-year window, and many customers transact infrequently, more longitudinal records would be required to understand these patterns further.

Despite these limitations, access to a unique dataset has facilitated the construction of a method to detect these uncertainties, and there is prospect for this technique to be adapted for implementation on real-time Big Data. For example, despite being large in volume, many limitations arose from the limited time period of this sample, and the lack of complete data pertaining to both individuals and stores as a result. This inherently restricted the extent to which we could infer dynamics such as changes in transactional behaviour over time. Implementation on more longitudinal and frequently updated data would undoubtedly improve estimations of the small area trip distributions derived here. This may allow analysis at a finer spatial granularity, which would be more sensitive to network changes. Nonetheless, it is hoped that applying this data-driven method offers insight into the veracity of the address information in consumer datasets, which would not be practically obtainable using traditional methods.

Funding This study was funded by the Economic and Social Research Council (grant number 1477365).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Behr, M., & Gober, P. (1982). When a residence is not a house: Examining residence-based migration definitions. *The Professional Geographer*, 34, 178–184.
- Chung, E., & Kuwahara, M. (2007). Mapping personal trip OD from probe data. *International Journal of Intelligent Transportation Systems Research*, 5, 1–6.
- Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). May. Addressing big data issues in scientific data infrastructure. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 48–55). IEEE.
- Dolega, L., Pavlis, M., & Singleton, A. (2016). Estimating attractiveness, hierarchy and catchment area extents for a national set of retail centre agglomerations. *Journal of Retailing and Consumer Services*, 28, 78–90.
- Ellegård, K., & Vilhelmsen, B. (2004). Home as a pocket of local order: Everyday activities and the friction of distance. *Geografiska Annaler: Series B, Human Geography*, 86, 281–296.

- Gale, C. G., Singleton, A., 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.
- Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*, 779–782.
- Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*, 255–261.
- Guptill, S. C., & Morrison, J. L. (Eds.). (2013). *Elements of spatial data quality*. Amsterdam: Elsevier.
- Isard, W. (1956). Regional science, the concept of region, and regional structure. *Papers in Regional Science*, *2*, 13–26.
- Kotler, P. (2002). *Marketing places*. New York: Simon and Schuster.
- Lansley, G., Li, W., & Longley, P. (2017). (Online) Representing Population Dynamics from Administrative and Consumer Registers. In *GISRUK 2017 Conference Proceedings*, Manchester University, Manchester. Available from: http://huckg.is/gisruk2017/GISRUK_2017_paper_80.pdf. Accessed 6 Dec 2017.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.
- Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*, 449–461.
- Reichheld, F. F. (2001). *Loyalty rules!: How today's leaders build lasting relationships*. Boston: Harvard Business Press.
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data. *IBM Global Business Services*, pp., 1–20.
- Sheller, M. (2011). Mobility. *Sociopedia.isa* (online), 1–12. Available from: <http://www.sagepub.net/isa/resources/pdf/mobility.pdf>. Accessed: 5 Dec 2017.
- Sheller, M., & Urry, J. (2006). The new mobilities paradigm. *Environment and Planning A*, *38*, 207–226.
- Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, *327*, 1018–1021.
- Stillwell, J. (Ed.). (2016). *The Routledge handbook of census resources, methods and applications: Unlocking the UK 2011 census*. London: Routledge.
- Van der Klis, M., & Karsten, L. (2009). Commuting partners, dual residences and the meaning of home. *Journal of Environmental Psychology*, *29*, 235–245.
- Webber, R. J., Butler, T., & Phillips, T. (2015). Adoption of geodemographic and ethno-cultural taxonomies for analysing big data. *Big Data & Society*, *2*, 2053951715583914.
- Yi, Y., & Jeon, H. (2003). Effects of loyalty programs on value perception, program loyalty, and brand loyalty. *Journal of the Academy of Marketing Science*, *31*, 229–240.
- Zandbergen, P. A. (2009). Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS*, *13*, 5–25.