The Provenance of Loyalty Card Data for Urban and Retail Analytics

Abstract

The deployment of loyalty card and other consumer data in geographic research brings opportunities to explore and understand patterns of purchasing behaviour in unprecedented detail. However, valid generalisation requires thorough evaluation of their potential bias. We argue that, in competitive markets where consumers can choose to shop across competing retailers, loyalty card data from just one of these may not represent a 'complete' view of all purchases, and that this 'completeness' must be controlled for when assessing bias. To this end, we undertake a UK wide analysis of loyalty card data assembled by a major UK grocery retailer and provide guidelines for their effective deployment in the domains of urban and retail analysis. We assess, for the first time, the 'completeness' of circa 500 million customer transactions recorded by a major customer loyalty programme in representing the overall purchasing patterns of circa 16 million consumers across the entire UK, and develop a method by which to do this. Moreover, no operator has complete national store coverage, and so we suggest ways of accommodating this when conducting analysis using loyalty card data. We illustrate the importance of these issues before providing recommendations for the wider use of consumer loyalty card data.

Keywords

Retail Geography; Loyalty Card Data; Consumer Data; Urban Analytics; Retail Analytics

1 Introduction

Recent years have seen the deployment and use of new forms of Big Data to understand urban and regional activity patterns at higher spatial and temporal granularities than hitherto possible using conventional statistical sources (Lansley & Cheshire, 2018; Singleton & Arribas-Bel, 2019). This has entailed the re-use of consumer data in what are essentially data-driven approaches to provide thicker quantitative descriptions of spatial processes and outcomes (Miller & Goodchild, 2015). These new forms of 'found' data (Connelly et al., 2016) are very different from traditional sources in that they lack most of the quality controls and reference points associated with conventional surveys or other statistical sources and many are by nature 'incomplete'. Yet at the same time, they offer considerable opportunities to uncover and understand processes, contexts and behaviours that may not otherwise be detected, at spatially extensive yet granular scales and with timely refresh intervals.

Viewed in this general context, consumer data offer great potential to extend analysis of social concerns such as deprivation, not only in the range and frequency of measures of well-being but also by supplementing data describing night-time geographies of poor physical and social conditions (Norman, 2017) with activity-based measures of daily behaviours (Martin et al., 2015). Suitably balanced, a melange of traditional and consumer data sources can greatly extend our understanding of changing activities and behaviours, both on- and off-line, facilitating, for example, intersectional analysis of deprivation and emergent 'convenience cultures' (Wrigley et al., 2019), 'food deserts' (Wrigley, 2002) and e-retailing (Treadgold & Reynolds, 2016).

Whilst customer loyalty card data are clearly integral to this agenda, they are only rarely made available for academic research, often being restricted so as to preserve the anonymity of cardholders and to protect commercial interests. In most cases, only specific regional study areas or a subset of stores are made available. Where they have been, they have made useful contributions towards understanding seasonal (Newing et al., 2015) and temporal (Hood et al., 2016) demand, e-retailing (Kirby-Hawkins et al., 2019), and analysis of diets (Jenneson et al., 2020; Nevalainen et al., 2018; Green et al., 2020). Within the retail, marketing and information systems (IS) domains, they have proved equally valuable for developing methods, models and segmentations to understand various aspects of consumer behaviour (Sarantopoulos et al., 2016; Griva et al., 2018).

In contrast to more traditional and representative social surveys, assemblages of loyalty card data accrue with no structured design. Identification of a corpus of relevant cardholders, behaviours and locations is thus a necessary precursor to using such data when addressing both society-wide and retailer-specific concerns. However, to date, there is an absence of substantial evaluations of loyalty card data and their readiness for wider re-use. Inferences about behaviour are only robust and defensible if users of research are provided with clearly documented metadata detailing the extent to which data are representative of consumer behaviour across all of society, or within clearly specified population segments.

In this paper, we offer what we believe is the first UK-wide evaluation of the representativeness of data collected using a major grocery retailer's loyalty card. This is achieved using 52-weeks of data from 2015 for circa 15.9 million cardholders. As we demonstrate, effective use of loyalty card data in any local case study is contingent upon the saturation of store networks and market share. Moreover, we illustrate that even in cases where store networks have large coverage, consumers remain free to avail themselves of competitor locations and so loyalty card data do not present the entire picture. To this end, we view representativeness through the lens of 'completeness' by developing a new method to segment cardholders. The concept of 'completeness' might be thought of as in some ways analogous to loyalty segmentation based on 'share-of-wallet' or measures of frequency, recency and spend. However, here we extend these aggregate measures by examining loyalty card data at the category level to estimate the degree to which a customer's transactions reflect all purchases they may make, regardless of which retailer they conduct those in. We facilitate this analysis of purchasing norms by comparing each household's expenditure to the consumer expenditure survey the Living Costs and Food Survey (LCFS) (DEFRA & ONS, 2019) before segmenting all cardholders into four groups ranging from 'complete' to 'incomplete' and comparing these groups to the wider population. Having accounted for the 'completeness' of cardholder purchasing, we demonstrate why doing so is imperative for investigations into shopping behaviour which re-use loyalty card data.

The paper is set out as follows. First, we describe the organisational setting to the collection of loyalty card data, including how they are used to understand the loyalty and behaviours of consumers, before highlighting the poor understanding of the provenance of these data in comparison to the wider population. In the main section, we first evaluate how the coverage of store networks geographically constrains the population represented within loyalty card data for different retailers. Next, we address the issue of cardholder completeness. These two aspects are brought together in the section that follows, which evaluates the representativeness of the loyalty card data in comparison to the overall population of the UK. Finally, we illustrate why this work is crucial to effective exploitation of loyalty card data using analysis of interpurchase periods between different consumer groups.

2 Background and motivation

Loyalty cards provide an important source of consumer data, produced at scale to understand purchasing behaviour and to inform marketing practices (Longley et al., 2018). Here, we use the term' loyalty card data' to describe transactional records of the location and timing of account holder purchases either online or instore. UK retailers such as Tesco, Sainsbury's and Boots (amongst others) have long used loyalty card schemes to attract and retain customers, and to understand consumption patterns. Insights into the purchasing behaviours of specific consumer types and the specific products that they purchase can be gained (Bradlow et al., 2017; Griva et al., 2021), yet academic access to individual-level purchasing records has been restricted for a variety of commercial and data protection reasons (Lansley & Cheshire, 2018). When loyalty card data have been made available, they are often for a subset of records, such as limited to a single geographic region (Green et al., 2020; Newing et al., 2015). For example, Felgate et al. (2012) used Tesco loyalty card data to assess the impact of promotions, using only 1.4 million records; Griva et al. (2018) use transaction data from only a sample of stores; and Sarantopoulos et al. (2016) use transactional records from just six stores. As such, the arrangements negotiated hitherto for academic access remain somewhat partial and piecemeal. This, and a variety of other challenges, must be addressed and overcome if potential benefits to wider society are to accrue (Birkin, 2019; Royal Statistical Society, 2019).

2.1.1 Representation and data sparsity in consumer data

In practice, the data owner has little control over the magnitude or composition of customer response to loyalty schemes, and no single player is likely to relate to any clearly defined population of customers, since no players are monopoly providers in competitive markets. This stands in stark contrast to conventional social survey procedures, in which populations are generally clearly defined and individuals within them have known and pre-specified chances of selection (see, e.g., Dixon & Leach, 1970). Traditional survey methods, however, come with large costs to reach what is a relatively tiny fraction of the population and are experiencing ongoing declines in response rates over time (Stedman et al., 2019).

Consumer data sources are thus normally far from representative of entire populations, and data exploration and selection is required to understand and accommodate the sources and operation of resultant bias. Heuvelink (1999) defines error as the difference between reality and our representation of it. No loyalty card profile can provide a complete record of all individual or household purchasing (Lansley & Cheshire, 2018), and loyalty card data are characterised by messiness, multi-dimensionality and incompleteness in the absence of rigorous quality controls (Goodchild, 2013). In-store card swipes may over-represent characteristics of habitual visitors (Wright & Sparks, 1999), and customers may not swipe their cards frequently after signing up (Cortiñas et al., 2008). Moreover, the incentives offered by competing loyalty schemes in the market can lead to complex switching and multi-shopping behaviours (Gijsbrechts et al., 2008). These issues conspire to further degrade representations of typical purchasing behaviours and generalisations of them. In practical terms, consideration must be given customer visit frequencies and types alongside the products, promotions and places of purchase, so as to not adversely affect results (Griva et al., 2021). Given this, we argue that evaluating the representativeness of loyalty card data makes sense only if done so with due consideration of how much of a consumer's shopping is captured.

2.1.2 Methods to assess and segment customer loyalty

In early research into loyalty card schemes, efforts dedicated to examining whether loyalty cards created increased customer loyalty (e.g. Smith et al., 2003), with often conflicting results (Meyer-Waarden, 2007). Part of this stream of research involved segmenting customers according to the loyalty they displayed in a retailer through a range of behavioural measures such as frequency, retention over time, and share of wallet (Allaway et al., 2006). Such measures are easily extracted from transactional records associated with a consumer and involve the analysis of either a single category or at the overall firm-level (Park et al., 2014). One such example is Recency, Frequency and Monetary value (RFM), which combines a customer's transactional patterns at the retailer, including how much they spend, over a selected period and which is used to segment groups within the overall customer base (Frasquet et al., 2021). Another approach that has been suggested is to measure and model the time between transactions to identify regular and routine shoppers (Kim & Park, 1998). Others (Allaway et al., 2006) have combined a selection of similar transactional variables (e.g. the number of items purchased, total period shopping, the amount spent) into k-means clustering to classify consumers into six loyalty types.

Some of these aggregate segmentation models remain in widespread use across industry in part due to their relative simplicity to apply to large corpora of customer records. However, such models pay no attention to the universe of transactions that may take place beyond the retailer's walls. Where resources permit, issues of cross-shopping have been evaluated using panel-data (e.g. Melis et al., 2016), or by sequence analysis made possible by the availability of records conducted in competing banks (Ho et al., 2021). In essence, our understanding of customer loyalty is largely based on the records of just a single retailer, or is reliant upon expensive investigations that collect bespoke panel data (Meyer-Waarden, 2007).

2.1.3 Research Gap

Whilst more generally, there have been attempts to understand issues of population coverage using some new forms of data (e.g. Longley et al., 2015; Leak & Lansley, 2018, Lovelace et al., 2016), the issue of completeness of customer purchasing records and self-selection of programme participants has yet to receive sustained focus. Methodologies to enable large-scale evaluations of the provenance of loyalty card data are in their nascent stages. In one recent example of the use of loyalty card data for analysis of consumption of different food types at the neighbourhood level (Aiello et al., 2020), consideration was given to the uneven distribution of the store network and the profiles of cardholders compared to a relevant population. However, no attempt was made to analyse these factors alongside purchasing frequency, which may have led to a more biased representation of diets. Another study contacted a sample of loyalty cardholders from Finland's largest grocery retailer. An important step taken was to understand the sample in comparison to the general Finnish adult population, done so by constructing post-stratification weighting to control for non-response. The findings of this research were that, despite divergence from the population at large, the volume of data-enabled analysis of sub-groups cost-effectively (Vuorinen et al., 2020). It is in this context that we address the gap in our understanding of how representative loyalty card data are by analysing nationwide loyalty card data for the first time.

3 Evaluating the completeness of loyalty card data

As illustrated above, all loyalty card data can offer insights into aspects of consumer behaviours, however the most panoptic view of consumption emerges from the schemes of major grocery retailers, even when taken alone rather than in combination. The challenge coupled with this is that the UK grocery market is catered for by retailers targeting a range of market segments, none of whom has universal geographic coverage. In practice, investigations within this setting require the restriction of analysis to locations in which our retailer's store network potentially enables consumers to satisfy their needs, whilst simultaneously acknowledging that even in such locations, competing retailers offer the same opportunity for needs to be fulfilled. The evaluations presented here grapple with these dual issues. We begin with an outline of the data used before presenting a two-stage selection process to control for geographical coverage and, following this, the 'completeness' of purchasing histories.

3.1 Loyalty card data

We base our extensive UK-wide loyalty card analysis upon that provided by a major UK retailer for 2015. The data records circa 540 million transactions made by circa 15.9 million cardholders belonging to circa 13.9 million households, comprising transaction histories ranging from a single to more than 400 transactions during the 52-week study period. Each transaction details the value and number of items purchased in each of 70 product categories such as 'apples, bananas and pears', 'coffee', 'UHT fruit juices', 'fats' and 'pasta', as defined and aggregated by the retailer. The transactions are attributed using anonymised and hashed customer and household identifiers, the geographic component of which allows place of residence to be assigned to a 2011 Census Output Area. We have used this to append the 2011 Output Area Classification (OAC: Gale et al., 2016) Supergroups and Groups. The customer and household identifiers are subject to a standard sequence of data cleaning and matching processes conducted by the retailer. For reasons of information governance, no account-level demographic or household size information is provided. Commercial sensitivities prevent us from reporting precise numbers of customers or their transactions, and we report rounded figures only.

3.2 Store network coverage

The fundamental premise of spatial interaction modelling is that distance has an attenuating effect upon the likelihood of a customer visiting a store (e.g. Birkin et al., 2002), although patronage may also of course be influenced by workplace location or leisure activities (Hood et al., 2016). Regular purchases are thus usually made close to place of residence or one frequented for another purpose, and thus store location is crucial to retail patronage patterns.

It follows that the 'completeness' of a consumer's loyalty transaction record in large part reflects the spatial configuration of supply through the retailer's store network, since this shapes the likelihood of providing some or all of consumer requirements. In the case of the UK's major grocery retail corporations, the contemporary national store network is the outcome of historic patterns of growth, merger and acquisition (Reynolds & Wood, 2010). These outcomes can be viewed using Geolytix's (2019) Retail Points, an open dataset that records the locations and approximate sizes of every grocery store operated by the main retailers in the UK. Table *1* presents the percentages of the 2011 Census Output Area Populations living within 5 miles of a store for nine large national retailers, as of September 2019. It is clear that even large national retailers such as Co-Op, Tesco, Asda and Sainsbury's are not within 5 miles of every Census Output Area.

Retailer	Percentage of Census 2011 Population living within 5 miles of a store
Aldi	83.4%
Asda	79.1%
Co-Op	95.4%
Lidl	83.3%
Marks and Spencer	84.1%
Morrisons	76.1%
Sainsbury's	84.3%
Tesco	92.8%
Waitrose	52.7%

Table 1: *Percentage of UK population resident within 5 miles of major grocery retailers (source: Author calculations based on Geolytix, 2019)*

Figure *1* breaks down this national picture of population coverage into geodemographic groups using OAC Supergroups. As can be seen, only Co-Op has over three-quarters of the UK's Rural Residents living within 5 miles of one of its stores, and most retailers have less than half. However, most retailers have a store within 5 miles for over three-quarters of the population of all other OAC Supergroups. Only Waitrose, as the national picture above suggests, has relatively low proportions of some Supergroups resident within 5 miles of its store network: this is particularly pronounced for the less-affluent OAC Supergroups of Hard-Pressed Living and Constrained City Dwellers, with accessibility restricted to 32.3% and 42.9% of these Supergroups respectively. It is seemingly the case that some retailers, such as Waitrose, Sainsbury's and Marks and Spencer, target more affluent groups and is of concern if research purporting to be society-wide re-uses their loyalty card or transaction data. This said, both Sainsbury's and Marks and Spencer have stores within 5 miles of over 80% of the population within these two OAC Supergroups, despite historic patterns of organic growth and targeted openings.

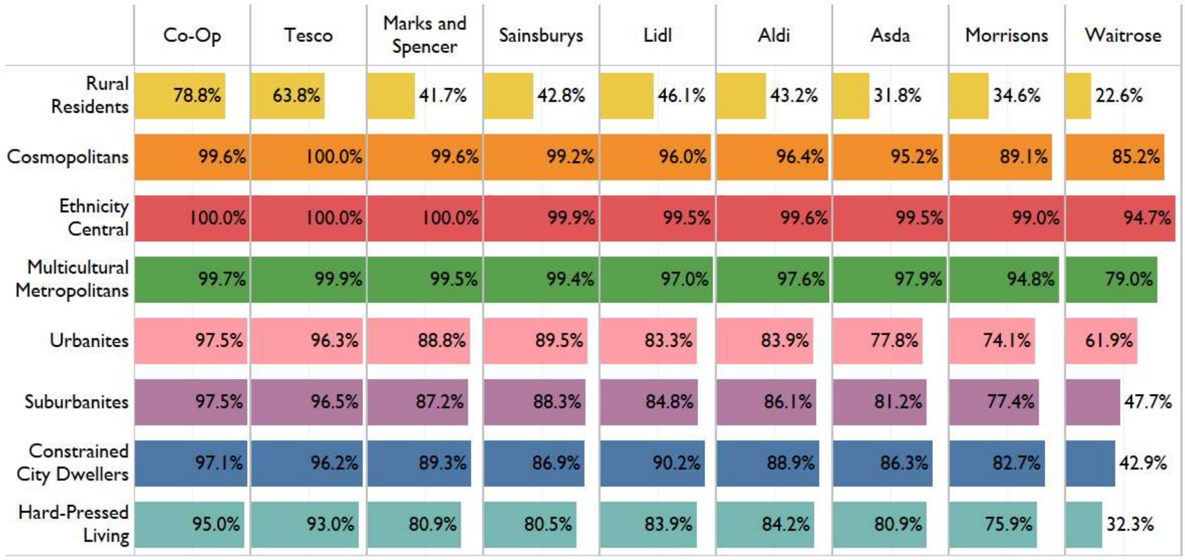

Figure 1: *Percentage of UK population living within 5 miles of major UK grocery retailers' stores, by OAC Supergroup (source: Author calculations based on Geolytix, 2019).*

3.2.1 Implications and initial treatments to loyalty card data

The implications of this are twofold. First, that store networks fundamentally shape regional and local consumer behaviour and the resultant observations found within loyalty card data. Second, most nationwide retailers offer sufficient opportunities for much of the population to visit their stores, potentially allowing for most consumer and geodemographic types to be represented within the loyalty card data that ensues despite a degree of unevenness in coverage. The practical impact of the first consideration is that we must control for this in our analysis. Signing up to a loyalty card, and indeed using one, is possible regardless of a customer's residential address or proximity to a store, resulting in transactional records being found throughout the UK, as shown by Lloyd & Cheshire (2018). These records may prove to distort geographical measures and indicators we may choose to use in our analysis, such as the distance travelled to a store or purchase frequencies.

Accordingly, prior to further analysis, we implemented a 'cookie-cutter' operation around each store to remove any neighbourhoods which are deemed to be too far away and subsequently, to identify the cardholders resident within them. To do this, we applied drivetimes varying in size, and determined by the rurality/urbanity of the store alongside its size. For example, stores under 300sqm in urban areas were deemed accessible only if within five minutes drive, or ten minutes in rural areas. Following this, checks were used to identify output areas in which more than the national median of transaction counts for the 52-week study period (1,500) were undertaken, and these are included in the subsequent analysis irrespective of drivetime accessibility. This led us to exclude any cardholders resident in some 30,320 output areas from further analysis. Although this substantially reduces the geographic coverage of our subsequent analysis to 86.9% of UK 2011 census output areas, the wastage is only 5.1% of cardholders, 4.8% of households and 2.5% of transactions.

In addition to understanding the geography of retail supply, we also consider consumer reliance upon a single retail chain, as manifest in purchasing behaviour. As such, we next focus upon the 'completeness' of purchasing indicated by the breadth and volume of items purchased across time, as recorded within each loyalty card account.

3.3 Defining 'complete' cardholder and household records

For obvious reasons, external validation of the level of completeness of a loyalty card account cannot be undertaken using rival loyalty card scheme data, and transactions for which no card is swiped go unrecorded. As discussed in section 2, previous work segmented customer loyalty using panel data or RFM techniques focused on the transactions conducted at a single retailer. Here, we develop a new perspective upon completeness (and upon loyalty segmentation, though this is not our primary aim) by considering the breadth of purchases made in different product categories compared to the consumption domains of the Office for National Statistics (ONS) Living Costs and Food Survey (LCFS). From this perspective, 'completeness' measures the extent to which the retailer accounts for the estimated total purchases of the household for each week and in aggregate, for each period. Comparison to the LCFS overcomes the constraint of having no information available on purchases made outside the retailer whilst simultaneously enabling us to understand purchasing across categories, in contrast to aggregate measures such as RFM.

The LCFS surveys circa 5000 households' purchases each year using a detailed diary kept for more regular purchasing and an interview to capture more infrequent purchases such as cars or white goods. In particular, the diary records all grocery purchases from all retailers by category. The 2015-16 release achieved a 46% response rate and a sample size of 4,912 households (Bulman et al., 2017). Here we use the detailed derived household expenditure data available through the UK Data Service (DEFRA & ONS, 2019). This contains the average weekly expenditure for each household, in supermarkets, for 96 categories of grocery items, split into groups according to the Classification of Individual Consumption by Purpose (COICOP) scheme. These data also contain a variety of household-level variables such as household size, income, urbanity and 2011 OAC Supergroups.

3.3.1 Method: Comparing cardholder purchases to the LCFS

We begin to assess the 'completeness' of loyalty card records by aggregating circa 15.9 million cardholders into circa 14 million households, while retaining flags to identify multiple cardholders within households. We accommodate non-stationarity in purchasing behaviour and customer churn by splitting the annual series into four 13-week periods, each containing between 12 and 13 million unique cardholders drawn from circa 11.5 million households in each cohort. Following this initial pre-processing, the remaining steps of analysis are summarised in 2, with more detail for each step following.

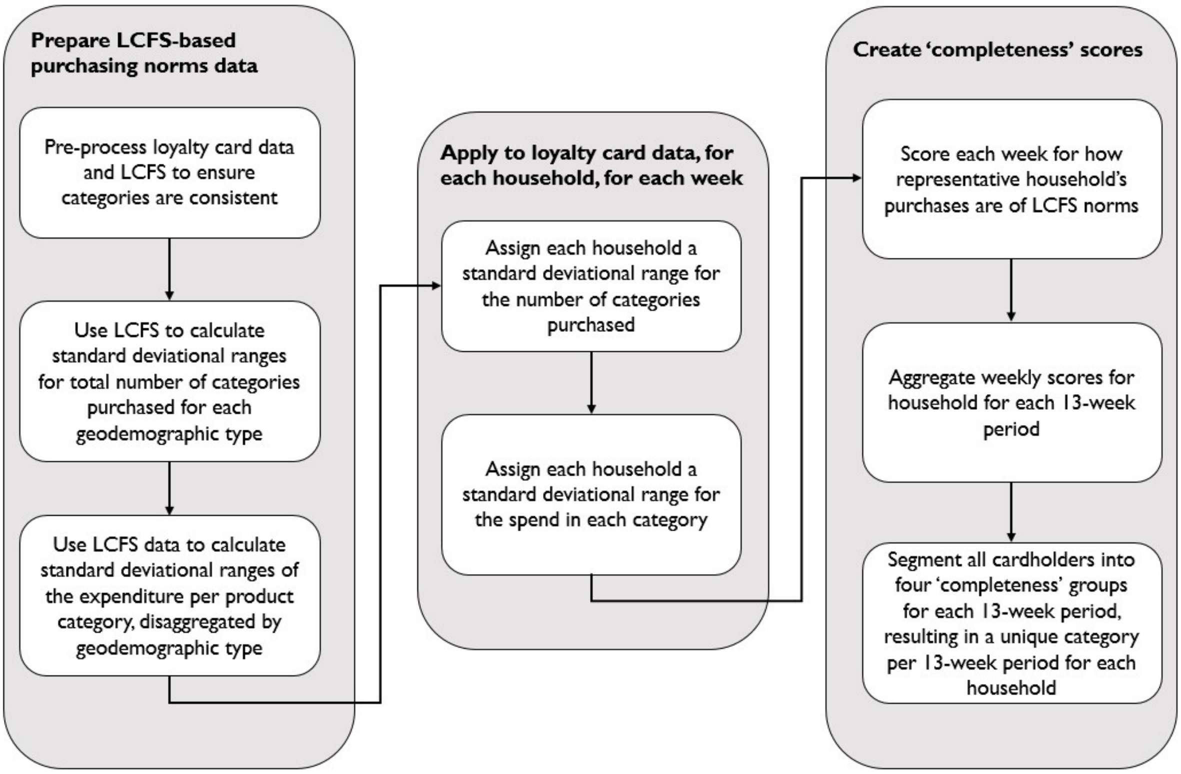

Figure 2: *Summary of analysis steps*

Initial data processing was performed to match the transaction data (which are pre-aggregated to 70 organisational categories) to the LCFS COICOP scheme. The LCFS was then used to calculate the mean number of categories purchased each week along with standard deviational ranges, disaggregated by OAC Supergroups to account for variation in purchasing power and sustenance needs. These ranges were used in the second stage to categorise the equivalent weekly purchasing for each household in the loyalty card data and are shown for each OAC Supergroup in Table 2. It can be seen, for example, that according to the LCFS, households in the Rural Residents OAC Supergroup who are within the 'around average' range (-0.5 to 0.5 s.d.) purchase between 21 and 35 categories per week.

OAC Supergroup	Mean expenditure (£) per week	Mean number of categories per week	< -1.5 s.d. (low)	-1.5 to -0.5 s.d. (below average)	-0.5 to 0.5 s.d. (around average)	>0.5 s.d. (above average)
Rural Residents	£80.32	29.5	1-11	12-20	21-34	35+
Cosmopolitans	£55.49	24.3	1-9	10-16	17-27	28+
Ethnicity Central	£51.34	23.4	1-7	8-14	15-26	27+
Multicultural Metropolitans	£58.58	25.0	1-8	9-15	16-28	29+
Urbanites	£69.03	27.9	1-11	12-18	19-31	32+
Suburbanites	£79.09	29.6	1-13	14-21	22-33	34+
Constrained City Dwellers	£50.35	23.9	1-7	8-14	15-27	28+
Hard-Pressed Living	£62.46	26.6	1-9	10-17	18-30	31+

Table 2: *Number of categories purchased per week by OAC Supergroup by standard deviational range (source: Author calculations from LCFS adjusted categories)*

We repeated this process at the category level, establishing standard deviational ranges for the amount of money spent on each category per week. For example, within the LCFS, the Rural Residents OAC Supergroup spends between £1.17 and £2.25 per week on average (the -0.5 to 0.5 standard deviation range) on Bacon and Ham.

To assign the LCFS norms to the loyalty card data for each household, we began by aggregating multiple transactions into a total weekly expenditure for each category to facilitate comparison to equivalent temporal intervals in the LCFS whilst simultaneously accommodating those who conduct mainly 'top-up' shopping across a week instead of a large weekly shop. Using the OAC Supergroup of the household's place of residence, we then appended the relevant standard deviational range for a) the number of categories purchased in each week and b) the amount spent in each of those categories. For example, if a household in the Rural Residents Supergroup purchases 25 categories, they are labelled with 'around average' for a), and if they spent £1.99 on Bacon and Ham, they are labelled 'around average' for one of the 25 categories in b).

These two measures, both the breadth of categories and the spend within each, mean that our completeness scoring is based on 'normal' purchasing within the week. This contrasts with RFM approaches which do not account for the degree to which 'leakage' to competing retailers and non-swiping behaviours occurs and tend to give higher scores to higher spenders, even if this is only within a single category (and therefore is unrepresentative of 'normal' purchasing).

The completeness scoring was conducted first by creating a score for each week that a household makes a purchase, as illustrated in Figure *3*. This shows indicative products found in the weekly shopping in each for a randomly selected set of households. Light grey colouring of an icon indicates below-average expenditure for that category relative to LCFS norms. As can be seen, within the lower segments, a smaller number of categories are evident, however, the scoring has been designed so as to not heavily penalise a household that purchases a breadth of categories but spends below average on these in any particular week. Inevitably, a limitation of our approach is that it may under-report completeness for some smaller (especially single person) households or those who do not purchase a broad variety of categories over time.

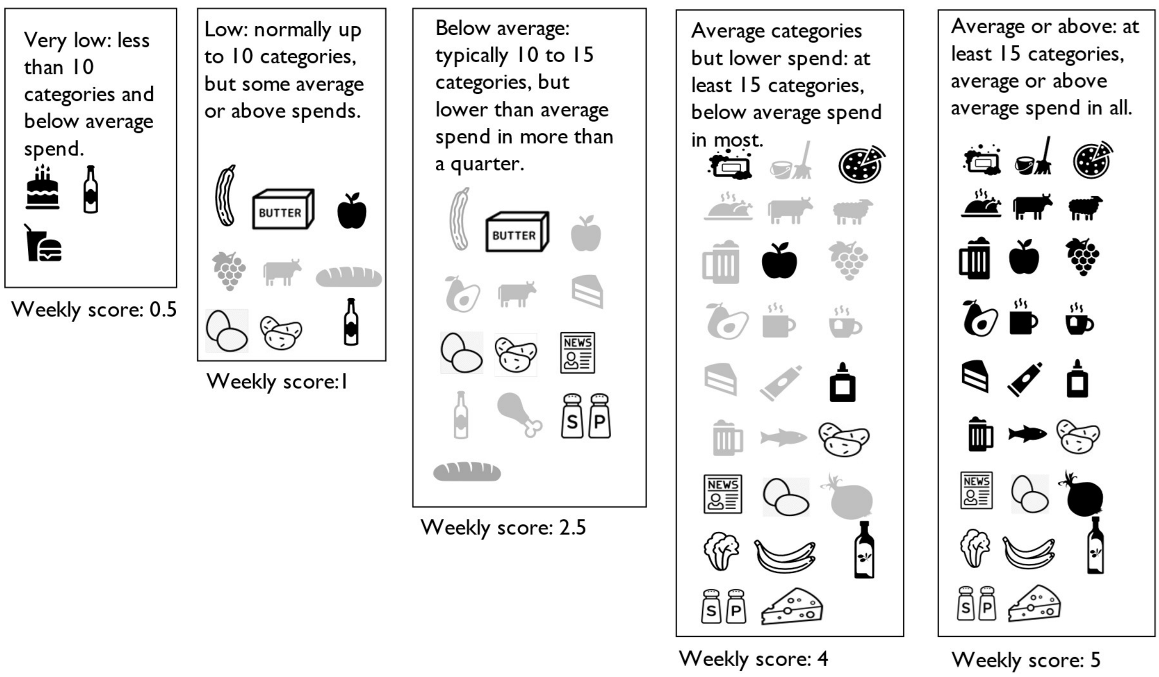

Figure 3*: Example weekly shopping baskets in each completeness category*

To create our periodic scores, we aggregated the weekly scores into a single total score for each 13-week period. The scores for each household were then assigned to one of four segments, defined by manually choosing breakpoints according to the patronage levels each represents, as indicated by the characteristics in Table *3*. The first represents the complete households, that achieve a score of 35 or more over the 13-week period, to attain which, for example, a household needs to purchase the 'average or above' shopping basket at least seven times over the 13-week period. The remaining segments are assigned as scores decrease, as illustrated in Table *3,* to the incomplete segment for scores of less than 10, where the household is only rarely observed. As such, attainment of complete status sets quite a high bar, but is also sufficiently flexible to allow for occasional missing weeks or lower expenditures from regular visits across most weeks within the study period.

3.3.2 Results

Using the definitions above, and after filtering out areas where the retailer has no store presence, as per section 3.2, an average across the four time periods of 60.8% of households' records are deemed incomplete, and just 11.5% are categorised as complete. Table *3* provides a fuller breakdown, detailing the percentage of households falling into four classes in each of the four 13-week periods (households not observed at all within the period are removed from the analysis). Just 5.1% of households are categorised as complete in all of the four periods across the year. In practical terms, this means not only that a low percentage of households records all grocery shopping using the loyalty card, but also that many do so for only part of a year (in each period, between 18.0% and 20.6% of households active within the year are not observed).

Overall Completeness Group	Typical Characteristics	Q1 % of households	Q2 % of households	Q3 % of households	Q4 % of households
Incomplete (score of less than 10)	Mostly small transactions or not observed in each period	60.0%	60.1%	60.8%	61.1%
Less Complete (score of 10-20)	Complete for one week and/or occasional lower scoring weeks	16.3%	16.3%	16.4%	16.3%
Partially Complete (score of 20-35)	Complete for some weeks but interspersed with absent weeks and lower scoring weeks	11.8%	11.9%	11.8%	11.5%
Complete (score of over 35)	All or almost all weeks categorised as complete or partial.	11.9%	11.8%	11.0%	11.1%

Table 3: *Percentage of households within each completeness category in each quarter (Q1-Q4) of the study period (source: Author calculations from scored loyalty card data)*

3.4 Representing population consumption

Despite much hubris about the potential of new Big Data sources, rather little work has been carried out to research their correspondence at the neighbourhood level with established data sources such as the census or other population estimates (Lansley & Cheshire, 2018). The results reported above demonstrate the importance of such checking, since even one of the UK's biggest loyalty programmes identifies only a low percentage of households as documenting all or even most of their shopping through it. Here we facilitate these comparisons by applying the household level completeness segments to the individual cardholders within each household (this is almost a one-to-one relationship, on average there are 1.1 active customers to each household). We then compare these counts, at UK Lower Super Output Area (LSOA) geography, to the corresponding 2015 Mid-Year Population Estimates, so as to identify the range of patronage levels within each neighbourhood and better understand the vastly uneven geography of coverage within our complete segment.

Using this benchmark, some 87.5% of UK LSOAs have cardholder population coverage of more than 5%, and 61% of LSOAs have coverage of more than 20% of the adult population. However, when considering that almost two-thirds of cardholders are within the incomplete segments, the usefulness of this decreases. If looking solely at the population coverage for the complete segment, the number of LSOAs registering greater than 5% coverage drops to just 19.6% of LSOAs (8341/42619). The range of population penetration for cardholders labelled as complete across the UK is illustrated by the LSOA geography in Figure 4. This shows the uneven geographic distribution of complete records, with many parts of the UK, particularly more rural locations, having either no, or less than 2.5% of households in this category.

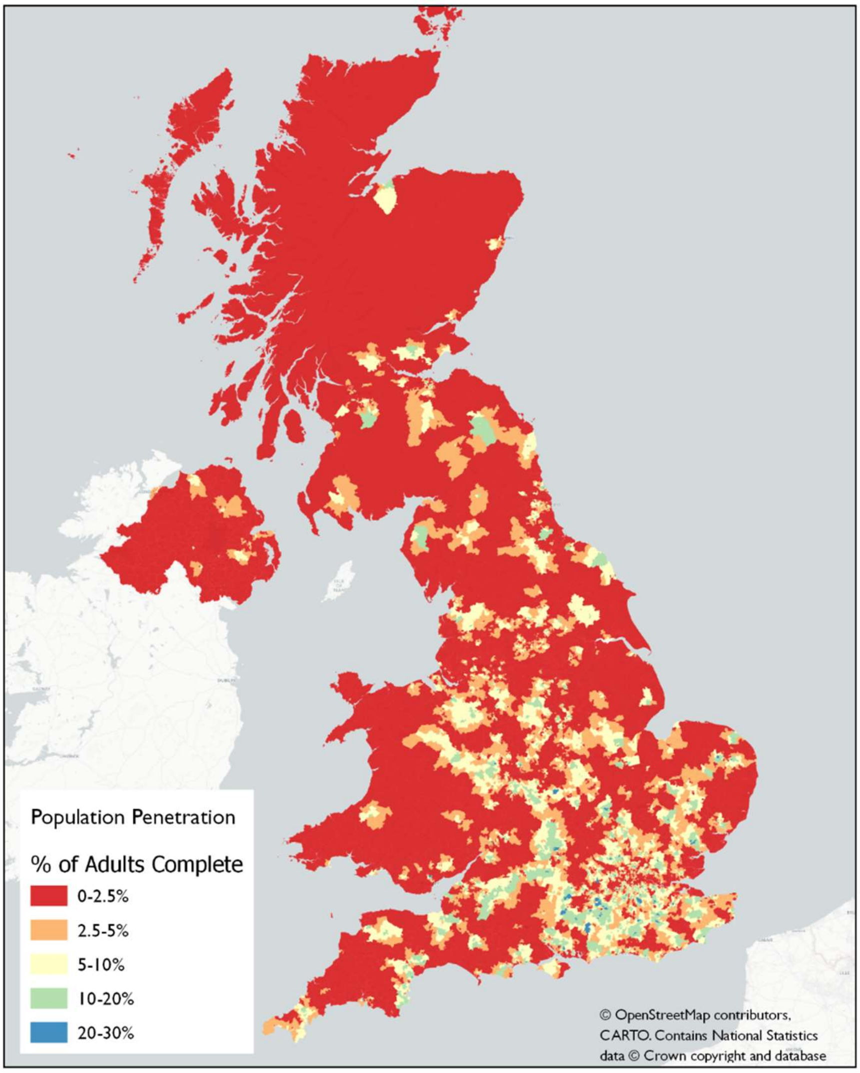

Figure 4: LSOA Map of the proportion of cardholders labelled as complete (average for all periods of analysis) (source: Author calculations)

We also compare the numbers of cardholders in each OAC Supergroup relative to the baseline UK 2011 Census population in Figure 5. The broad picture shows under-representation of Rural Residents, Constrained City Dwellers and Hard-Pressed Living Supergroups, with the latter two being substantially under-represented. There is substantial over-representation of Urbanites and Suburbanites, and on the whole, of Cosmopolitans and Ethnicity Central. Multicultural metropolitans are generally under-represented but are more likely to be found amongst complete households. However, although relatively over-represented, Cosmopolitans are under-represented within the most complete segment. This may be because of the dense and diverse food outlets and convenience culture on offer characteristic of these neighbourhoods, which may lead to smaller basket sizes and subsequently lower incentives to swipe loyalty cards. Over-representation in the remaining segments for this group indicates that the store network is present in such locations, and as such the under-representation in the complete segment suggests substantial differences in consumption and card swiping behaviours.

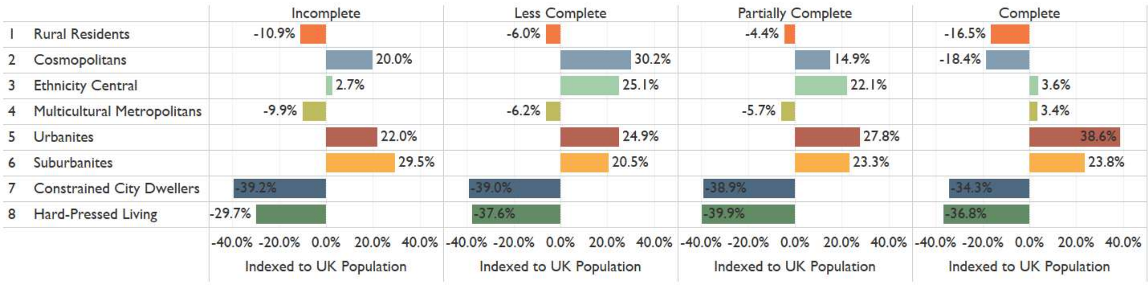

Figure 5: Percentage of cardholders by Completeness Segments and OAC Supergroup indexed to UK Base Population (Source: Autor calculations with loyalty card data and 2011 Census)

3.4.1 Discussion

The low and uneven incidence of complete purchasing profiles limits the inferences that may be drawn from loyalty card data and are set out nationally here for the first time. Yet the detail in purchasing behaviour is exposed in much greater depth than in conventional panel surveys. The data are also less vulnerable to attrition over time, and the geodemographic profiling makes the source of selection bias more apparent. Variability in population coverage can be understood in terms of the nationwide store network and local market share. Together, this makes it possible to understand the basis for generalisation from loyalty card accounts in different geographic locales, taking geodemographic factors into account as appropriate. Although the national corpus of loyalty card data cannot be thought of as a 'sample' in any scientific sense, the metadata created here offer a basis for generalisation through triangulation of rich loyalty card data with sample surveys and background geodemographic characteristics summarised from the census. In this way, value may be leveraged from these different sources: the LCFS recruits approximately 5,000 households over all OAC Supergroups, making robust generalisation impossible at regional and local scales, yet the loyalty card data provide detailed consumption data for a minimum of 65,000 complete cardholders within each Supergroup in each quarterly study period.

Figure 4 illustrates the challenge of using this rich data source to draw valid generalisations in localities where the incidence of full customer profiles is low. This nevertheless provides fertile ground for future research to apply small-area estimation techniques that weight regional and geodemographic group means to improve neighbourhood estimates where required in order to obtain universality of coverage (e.g. Ghosh & Rao, 1994). If a high bar is set for including only the most complete cardholder records, this would entail removal of some 89.5% of cardholder records, but this would still leave data pertaining to 1.2-1.3 million cardholders in each quarterly period. These retained cardholders record slightly under 125 million transactions over the entire 52-week period. In short, there is considerable scope to conduct highly detailed geo-temporal analysis using loyalty card data, using controls and techniques to accommodate the sources and operations of bias inherent in their collection.

3.5 Evaluation: the number of days between transactions

To illustrate these ideas, we conduct an initial examination of customer interpurchase periods between shopping trips, using loyalty card data for households in each completeness segment taken from the first quarterly period in our 2015 study period. The interpurchase period is useful for understanding how different consumer types conduct their shopping (Kim & Park, 1998): frequent purchases may be indicative of convenience style shopping habits as opposed to more routinised, less frequent shopping patterns. However, the concept of completeness developed above is key to validating the number of days between purchases. The mean inter-purchase period for all households is 11.2 days yet increases to 14.8 days for those labelled as incomplete (who have more than one transaction recorded), with periods of 9.0, 6.5 and 4.6 days for cardholder records that are less complete, partially complete and complete, respectively. These aggregate figures, including the apparently anomalous spacing of less complete and partially complete respectively, can be broken down by OAC Supergroup as shown in Table *4*.

OAC Supergroup	Complete households	Partially complete households	Less complete households	Incomplete households	All households
Rural Residents	5.1	7.2	10.1	16.1	12.1
Cosmopolitans	4.1	5.5	7.8	13.1	10.0
Ethnicity Central	3.8	5.4	8.0	13.5	9.7
Multicultural Metropolitans	4.5	6.4	8.9	14.8	10.9
Urbanites	4.7	6.6	9.1	14.8	11.1
Suburbanites	5.0	6.7	8.9	14.9	11.5
Constrained City Dwellers	4.3	6.4	9.3	15.4	11.2
Hard-Pressed Living	4.5	6.4	9.0	15.2	11.6

Table 4: *Mean interpurchase period (days) in each completeness segment by OAC Supergroup for the first quarter of analysis (source: author calculations)*

The complete segment alone therefore makes it possible to confirm differences in shopping habits between OAC Supergroups: Rural Residents and Suburbanites both space visits by five or more days, and Urbanites space theirs by 4.7 days, while Cosmopolitans and Ethnicity Central Supergroups have mean spacings of 4.1 and 3.8 days, respectively. This is consistent with analysis of convenience cultures, in which consumers shop ‘little and often’ (Hallsworth et al., 2010). Households living in neighbourhoods categorised as Constrained City Dwellers and Hard-Pressed Living have 4.3 and 4.5 days between transactions respectively, perhaps reflecting supply factors (smaller stores and food deserts: Wrigley, 2002), or economic access issues which give rise to variable purchasing power on a week-to-week basis (e.g. O’Connell et al., 2019).

We further illustrate how, using those records categorised as complete, and stratifying by geodemographics, region and the density of the surrounding supply of stores, it is possible to create a UK map of the estimated neighbourhood interpurchase period. For locations with no, or very low coverage, we ‘borrow strength’ from those with higher coverage to create synthetic estimates for each output area. For areas with higher coverage, we use composite estimates consisting of domain level synthetic estimates weighted with the direct estimates based on the proportion of cardholders to the population, as suggested in Ghosh & Rao (1994). The result is a map, for the whole UK, of the neighbourhood averages of the number of days between transactions, as seen in Figure 6. This map clearly shows the majority of inner

London cardholders having interpurchase periods of less than 4.1 days, but outer London cardholders and its more rural surroundings having an interpurchase period of slightly longer, including over five days. Outside London, the pattern is of lower interpurchase periods within city centres and inner cities, as exemplified by the case of Manchester. However, the general pattern within other cities is of slightly longer inter-purchase periods, of over 4.3 days. A similar story is found in most provincial and commuter towns, with town centres having lower interpurchase periods and rural fringes having slightly longer periods.

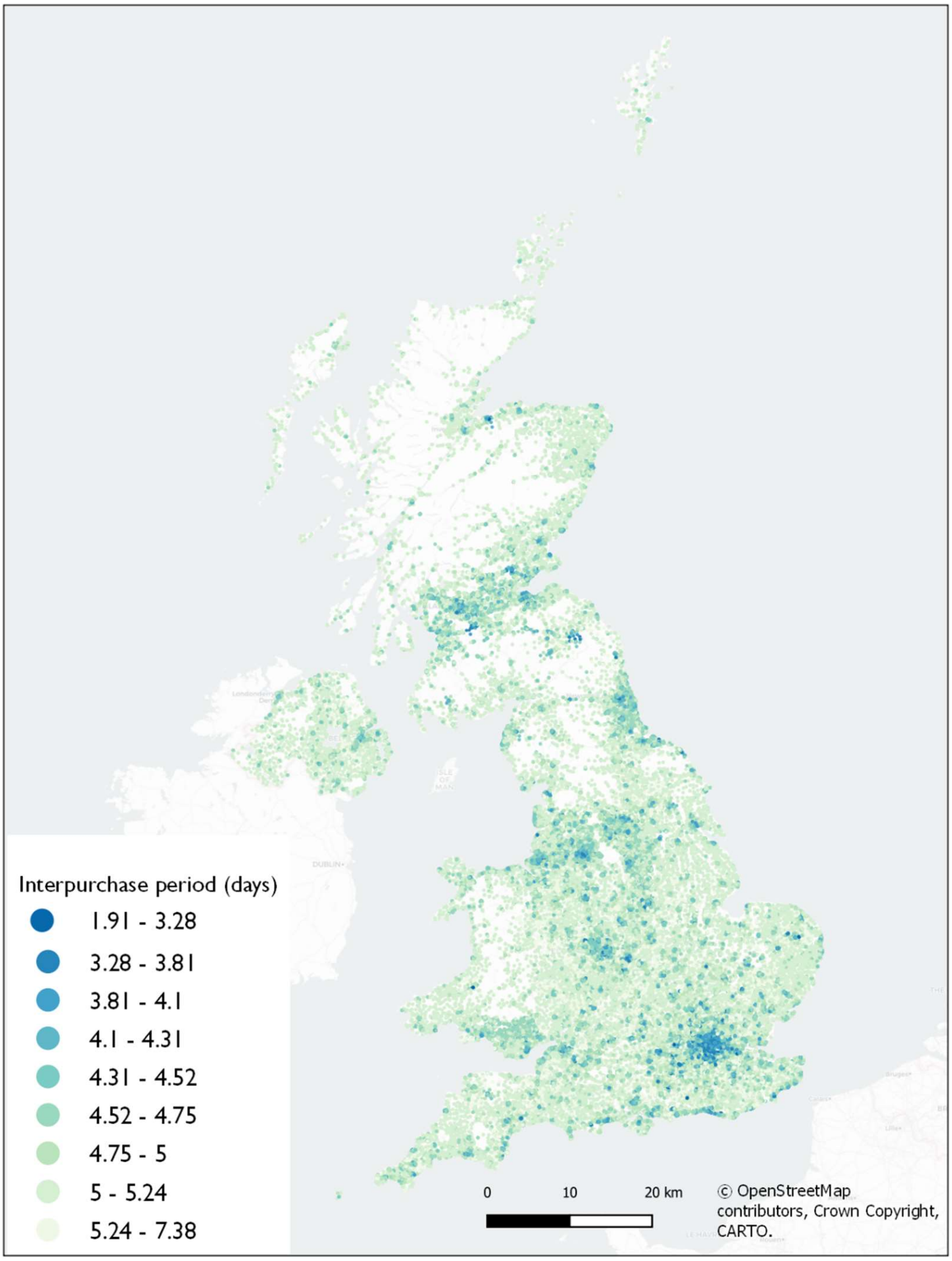

Figure 6: LSOA Map of estimated interpurchase periods across the UK (source: author calculations)

4 Conclusions

This paper has illustrated how loyalty card data might be re-purposed to better understand how communities go about their shopping in different circumstances and settings. To do this, however, considerable internal and external validation must be undertaken to ensure that consumer Big Data are truly analysis-ready. In this paper, we have subsumed these procedures within the concept of 'completeness' viewed from perspectives of both supply and demand. The value of consumer data is affected by the vagaries of market share across all parts of the UK and consequent accessibility of the store network. Loyalty card data arise from self-selecting samples of individuals that do not have any pre-specified probabilities of providing data yet provide vastly larger absolute numbers. Thus, in terms of volume, timeliness and detail, consumer data are superior to conventional surveys and censuses, but their value can only be reliably leveraged through triangulation and validation with respect to these sources.

The practical implication of our work is that, given suitable treatment, loyalty card data can prove valuable for better understanding of differences in consumption between segments of the population at large. The paper has illustrated how the strengths of loyalty card data can be selectively built on through linkage, comparison and profiling to traditional sources. Most importantly, we emphasise that researchers must recognise that uncritical use of loyalty card data can only be condoned in studies that recognise that they are exploratory in nature and that they are descriptive rather than inferential in scope.

Our theoretical contribution is to emphasise the limitations inherent in data-driven consumer data analytics, since the theoretical apparatus of inference is inapplicable. This is because consumer Big Data are assembled from self-selecting samples of loyalty card holders that may or may not be representative of a population, the characteristics of which are essentially unknown. Sustained focus on this area has been somewhat limited until now (Longley et al., 2015; Leak & Lansley, 2018). We further emphasise that, given that consumers exist within competitive markets, it is essential to consider representativeness through the lens of the 'completeness' of a consumer's purchasing that is recorded within their loyalty card histories. That is, to what extent the transactional records reflect purchasing norms. To do so, within the constraints of loyalty card data, in which purchase histories are only available for a single retailer, requires methodological innovation to draw on established sources so as to ground purchasing patterns in reality.

A limitation of our approach relates to its universal applicability, in that firstly, it is contingent upon the availability of detailed diary records of expenditure which may not be widely available across all regions or product types. Secondly, applications within other domains of expenditure will require further consideration of how to triangulate purchasing norms relevant to the retailer's industry, which may take place across a different range of products with different purchasing intervals.

By accommodating 'completeness,' the paper sets the foundations for future work using loyalty card data which explore retail applications such as the impacts of convenience culture, the revitalisation of high streets, and the understanding of multi-channel behaviour. Loyalty card data may enable focus on segments of the population that may be harder to reach, with, as illustrated here, coverage being of a sufficiently large scale to partially dispel preconceptions over the brand preferences of different consumer types. These data have merits over many traditional sources in their detailed recording of consumer behaviour for larger shares of the population (Lansley & Cheshire, 2018). They may also be suitable for a range of studies for the social good, such as food insecurity, sustainable economic growth, the spatial patterns of obesity (e.g. Morris et al., 2018), and the analysis of subsets of loyalty card users who have given explicit consent for the re-use and linkage of their purchasing habits to NHS records for exploring early interventions for specific health problems (Davies et al., 2018). In these ways, loyalty card data, pertaining to grocery purchasing in particular offer data-rich insights into people and place-based dynamics simultaneously, making rich contributions to urban and regional research.

Thus, loyalty card data have evident potential in many research domains that have detailed spatial and/or temporal focus. Not least, they remain valuable for better understanding patterns of consumption and the spatial interaction effects that surround this at both the individual and neighbourhood levels. These sources provide a valuable supplement to conventional census or survey information, particularly with respect to the creation of small area indicators for use within academia, policy and commercial strategy. Such indicators, especially if sensitive to the small-area heterogeneity that can characterise prevailing social conditions, may also contribute to analytical issues of bridging micro and aggregate representations of retail behaviour.

5 References

Aiello, L. M., Quercia, D., Schifanella, R., & Del Prete, L. (2020). Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. *Scientific Data*, *7*(1), 1-11.

Allaway, A. W., Gooner, R. M., Berkowitz, D., & Davis, L. (2006). Deriving and exploring behavior segments within a retail loyalty card program. *European Journal of Marketing*.

Birkin, M. (2019). Spatial data analytics of mobility with consumer data. *Journal of Transport Geography*. In Press.

Birkin, M., Clarke, G., & Clarke, M. P. (2002). *Retail geography and intelligent network planning*. John Wiley & Sons.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, *93*(1), 79-95.

Bulman, J., Davies, R., & Carrel, O. (2017). *Living costs and food survey: Technical report for the survey year 2015/16*. Retrieved from https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/livingcostsandfoodsurvey/livingcostsfoodtechnicalreport2015.pdf.

Connelly, R., Playford, C. CIRCA J., Gayle, V., & Dibben, C. CIRCA (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1-12.

Cortiñas, M., Elorz, M., & Múgica, J. M. (2008). The use of loyalty-cards databases: Differences in regular price and discount sensitivity in the brand choice decision between card and non-card holders. *Journal of Retailing and Consumer Services*, *15*(1), 52-62.

Davies, A., Green, M. A., & Singleton, A. D. (2018). Using machine learning to investigate self-medication purchasing in England via high street retailer loyalty card data. *PLoS one*, *13*(11), e0207523.

Department for Environment, Food and Rural Affairs, Office for National Statistics. (2019). *Living Costs and Food Survey, 2015-2016*. [data collection]. *3rd Edition*. UK Data Service. SN: 8210, http://doi.org/10.5255/UKDA-SN-8210-5

Dixon, C. CIRCA J., & Leach, B. (1977). *Sampling methods for geographical research*, CATMOG 17. Norwich: Geo Abstracts.

Felgate, M., Fearne, A., DiFalco, S., Martinez, M.G., 2012. Using supermarket loyalty card data to analyse the impact of promotions. *International Journal of Market Research.* 54 (2), 221–240. https://doi.org/10.2501/IJMR-54-2-221-240.

Frasquet, M., Ieva, M., & Ziliani, C. (2021). Online channel adoption in supermarket retailing. *Journal of Retailing and Consumer Services*, *59*, 102374.

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*, *2016*(12), 1-27.

Geolytix. (2019). *Retail Points Dataset.* Retrieved from: https://www.geolytix.co.uk/#!geodata

Ghosh, M., Rao, J.N.K., 1994. Small-area estimation: an appraisal. Statistical science 9 (1), 55–76. https://doi.org/10.1214/ss/1177010647.

Gijsbrechts, E., Campo, K. & Nisol, P. (2008). Beyond promotion-based store switching: Antecedents and patterns of systematic multiple-store shopping. *International Journal of Research in Marketing*, *25*(1), 5-21.

Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in Human Geography*, *3*(3), 280-284.

Green, M. A., Watson, A. W., Brunstrom, J. M., Corfe, B. M., Johnstone, A. M., Williams, E. A., & Stevenson, E. (2020). Comparing supermarket loyalty card data with traditional diet survey data for understanding how protein is purchased and consumed in older adults for the UK, 2014–16. *Nutrition journal*, *19*(1), 1-10.

Griva, A., Bardaki, C., Pramatari, K., & Doukidis, G. (2021). Factors Affecting Customer Analytics: Evidence from Three Retail Cases. *Information Systems Frontiers*, 1-24.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, *100*, 1-16.

Hallsworth, A., de Kervenoael, R., Elms, J., & Canning, C. CIRCA (2010). The food superstore revolution: changing times, changing research agendas in the UK. *The International Review of Retail, Distribution and Consumer Research*, *20*(1), 135-146.

Heuvelink, G. B. M. (1999). Propagation of error in spatial modelling with GIS. *International Journal of Geographical Information Systems*, 1, 207-217.

Ho, H., Tien, K. M. T., Wu, A., & Singh, S. (2021). A sequence analysis approach to segmenting credit card customers. *Journal of Retailing and Consumer Services*, *59*, 102391.

Hood, N., Clarke, G. & Clarke, M. (2016). Segmenting the growing UK convenience store market for retail location planning. *The International Review of Retail, Distribution and Consumer Research*, *26*(2), 113-136.

Jenneson, V., Shute, B., Greenwood, D., Clarke, G., Clark, S., Rains, T., & Morris, M. (2020). Variation in fruit and vegetable purchasing patterns in Leeds: using novel loyalty card transaction data. *Proceedings of the Nutrition Society*, *79*(OCE2).

Kim, B. D. & Park, K. (1998). Studying patterns of consumer's grocery shopping trip. *Journal of retailing*, *73*(4), 501-517.

Kirby-Hawkins, E., Birkin, M., & Clarke, G. (2018). An investigation into the geography of corporate e-commerce sales in the UK grocery market. *Environment and Planning B: Urban Analytics and City Science*, 2399808318755147.

Lansley G, de Smith M J, Goodchild M F and Longley P A (2018) Big Data and Geospatial Analysis, in de Smith M J, Goodchild M F, and Longley P A (2018) *Geospatial Analysis: A comprehensive guide to principles, techniques and software tools*, 6th edition, The Winchelsea Press, Edinburgh

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

Leak, A. & Lansley, G. (2018). Geotemporal Twitter Demographics. In Longley, P., Cheshire, J., & Singleton, A. (Eds.). (2018). *Consumer Data Research*. UCL Press.

Lloyd, A., & Cheshire, J. (2018). Detecting Address Uncertainty in Loyalty Card Data. *Applied Spatial Analysis and Policy*, 1-21.

Longley, P. A., Adnan, M. & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, *47*(2), 465-484.

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

Lovelace, R., Birkin, M., Cross, P., & Clarke, M. (2016). From big noise to big data: Toward the verification of large data sets for understanding regional retail flows. *Geographical Analysis*, *48*(1), 59-81.

Martin, D., Cockings, S., & Leung, S. (2015). Developing a flexible framework for spatiotemporal population modelling. *Annals of the Association of American Geographers*, *105*(4), 754-772.

Melis, K., Campo, K., Lamey, L. & Breugelmans, E. (2016). A Bigger Slice of the Multichannel Grocery Pie: When Does Consumers' Online Channel Use Expand Retailers' Share of Wallet? *Journal of Retailing, 92*(3), 268-286.

Meyer-Waarden, L. (2007). The effects of loyalty programs on customer lifetime duration and share of wallet. *Journal of retailing*, *83*(2), 223-236.

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

Morris, M. A., Wilkins, E., Timmins, K. A., Bryant, M., Birkin, M., & Griffiths, C. CIRCA (2018). Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map. *International Journal of Obesity*, *42*(12), 1963.

Nevalainen, J., Erkkola, M., Saarijärvi, H., Näppilä, T., & Fogelholm, M. (2018). Large-scale loyalty card data in health research. *Digital health*, *4*, 2055207618816898.

Newing, A., Clarke, G. P. & Clarke, M. (2015). Developing and applying a disaggregated retail location model with extended retail demand estimations. *Geographical Analysis*, *47*(3), 219-239.

Norman, P. D., & Darlington-Pollock, F. (2017). The changing geography of deprivation in Great Britain: Exploiting small area census data, 1971 to 2011. In Stillwell, J. (Ed.). (2017). *The Routledge handbook of census resources, methods and applications: Unlocking the UK 2011 census*. Routledge.

O'Connell, R., Owen, C., Padley, M., Simon, A., & Brannen, J. (2019). Which types of family are at risk of food poverty in the UK? A relative deprivation approach. *Social Policy and Society*, *18*(1), 1-18.

Office for National Statistics (ONS), Department for Environment, Food and Rural Affairs. (2017). *Living Costs and Food Survey, 2015-2016*. [data collection]. *2nd Edition*. UK Data Service. SN: 8210, http://doi.org/10.5255/UKDA-SN-8210-4

ONS, NRS, NISRA, 2017. 2011 Census aggregate data. UK Data Service (Edition: February 2017). UK Data Service. https://doi.org/10.5257/census/aggregate-2011-2.

Park, C.H., Park, Y.H., Schweidel, D.A., 2014. A multi-category customer base analysis. International Journal of Research in Marketing 31 (3), 266–279. https://doi.org/ 10.1016/j.ijresmar.2013.12.003.

Reynolds, J., & Wood, S. (2010). Location decision making in retail firms: evolution and challenge. *International Journal of Retail & Distribution Management*, *38*(11/12), 828-845.

Royal Statistical Society. (2019). The Data Ethics Landscape. Retrieved from: https://www.rss.org.uk/Images/PDF/get-involved/2019/Data%20Ethics%20Landscape.pdf

Sarantopoulos, P., Theotokis, A., Pramatari, K., & Doukidis, G. (2016). Shopping missions: An analytical method for the identification of shopper need states. *Journal of Business Research*, *69*(3), 1043-1052.

Singleton, A., & Arribas-Bel, D. (2019). Geographic Data Science. *Geographical Analysis*. Doi: 10.1111/gean.12194

Smith, A., Sparks, L., Hart, S., & Tzokas, N. (2003). Retail loyalty schemes: results from a consumer diary study. *Journal of Retailing and Consumer Services*, *10*(2), 109-119.

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? Understanding and coping with declining response rates to mail surveys. *Society & Natural Resources*, *32*(10), 1139-1154.

Treadgold, A. & Reynolds, J. (2016). *Navigating the New Retail Landscape: A Guide for Business Leaders.* Oxford: Oxford University Press.

Vuorinen, A. L., Erkkola, M., Fogelholm, M., Kinnunen, S., Saarijärvi, H., Uusitalo, L., Näppila, T. & Nevalainen, J. (2020). Characterization and Correction of Bias Due to Nonparticipation and the Degree of Loyalty in Large-Scale Finnish Loyalty Card Data on Grocery Purchases: Cohort Study. *Journal of medical Internet research*, *22*(7), e18059.

Wright, C., & Sparks, L. (1999). Loyalty saturation in retailing: exploring the end of retail loyalty cards?. *International Journal of Retail & Distribution Management*, *27*(10), 429-440.

Wrigley, N. (2002). 'Food deserts' in British cities: policy context and research priorities. *Urban studies*, *39*(11), 2029-2040.

Wrigley, N., Wood, S., Lambiri, D., & Lowe, M. (2019). Corporate convenience store development effects in small towns: Convenience culture during economic and digital storms. *Environment and Planning A: Economy and Space*, *51*(1), 112-132.