Detecting outliers and incompleteness in store network patronage.

Tim Rains[*1] and Paul Longley[†1]

[1]Department of Geography, UCL.

November 20, 2017

Summary

Novel forms of data such as loyalty card or transaction data come with new challenges such as representation, messiness and incompleteness. These can cause uncertainty within analysis. The transactions for loyalty card holders of a major UK retailer are validated by investigating their completeness and spatial interactions in stores through comparison with Census travel to work flows and other neighbourhood patterns. A small percentage of cardholders (3%) are found to have atypical patterns, but almost half are estimated to have some degree of incompleteness, creating uncertainty when generalising to the wider population.

KEYWORDS: Consumer data, ground-truthing, retail, loyalty, uncertainty, validation.

1. Introduction

Novel forms of data generated automatically from transactions, clicks and sensors have become the bedrock of much analysis in the 21st Century. These data enable us to obtain deeper descriptions of what is happening (Miller and Goodchild, 2015); however they bring with them new and additional issues to overcome (Longley, 2012) such as representation, messiness, and incompleteness. One of the sources is consumer data, in the form of loyalty card and transaction data. These data, like all, are only a representation of reality, and necessarily come with the characteristic of incompleteness (Longley et al, 2005). Retailers rarely cover an entire market equally, choosing to target certain markets or particular types of consumers (Reynolds and Wood, 2010); moreover, customers do not shop in a single retailer (Gijsbrechts et al, 2008), nor do they always swipe their loyalty cards (Wright and Sparks, 1999). Thus, we do not know what has been retained and what has been discarded from the universe of purchases. These characteristics have implications when trying to generalise (spatial) patterns and flows.

Using retailer loyalty card and transaction data, an attempt is made to understand customer activity in relation to a physical store network by comparison with the UK census travel to work flow data. This aims to pick out those customers who frequently interact with the store network in places where we would only expect to see them occasionally. Establishing this helps deal with uncertainty within analysis or issues of specification as well as benefitting retail operations. It also aims to identify customers who have incomplete patterns, from whom it is harder to make generalisations about society when more broadly reusing their data.

2. Methodological framework

A consumer wishing to make a purchase in a store (rather than online) will more often than not consider where they are in relation to that store (Clarke et al. 2006; Jackson et al. 2006). Much spatial interaction theory (e.g. Huff 1964) focused on residential shopping, but more recent work has explored consumer spatial dynamics away from these, such as tourist shopping (Newing et al. 2013) and workplace linked shopping (e.g. Waddington et al. 2017; Berry et al. 2016). Combined, these create a framework of retail types: residential based shopping, workplace based shopping, and the occasional shop. The approach taken here is to construct location sets or store-sets that broadly

[*] Tim.rains.15@ucl.ac.uk

[†] P.longley@ucl.ac.uk

represent these locations for each neighbourhood. These store-sets are then applied to individual customer behaviour to create types of customers based on their interactions with the store network, in order to validate the observed geo-temporal patterns. The data used are transaction and loyalty card data from a major UK retailer for almost 12 million customers who transacted more than 5 times over a 52 week period.

2.1. Store sets

A series of home stores are first established using transaction frequency and customer counts. A catchment area of stores was created for each Lower Super Output Area (LSOA) by selecting stores that the most customers shopped in the most frequently. A cut-off for 50% of all transactions for the LSOA was used. This approach, rather than the nearest x stores, allows for consideration of store-network density and the choice-set for store patronage. Almost 80% of LSOAs attained the 50% cut off through two or less stores. However, other LSOAs had a wider spread of home stores, with 1550 having more than 5 home stores. Examples of home stores are illustrated in **Figure 1.**

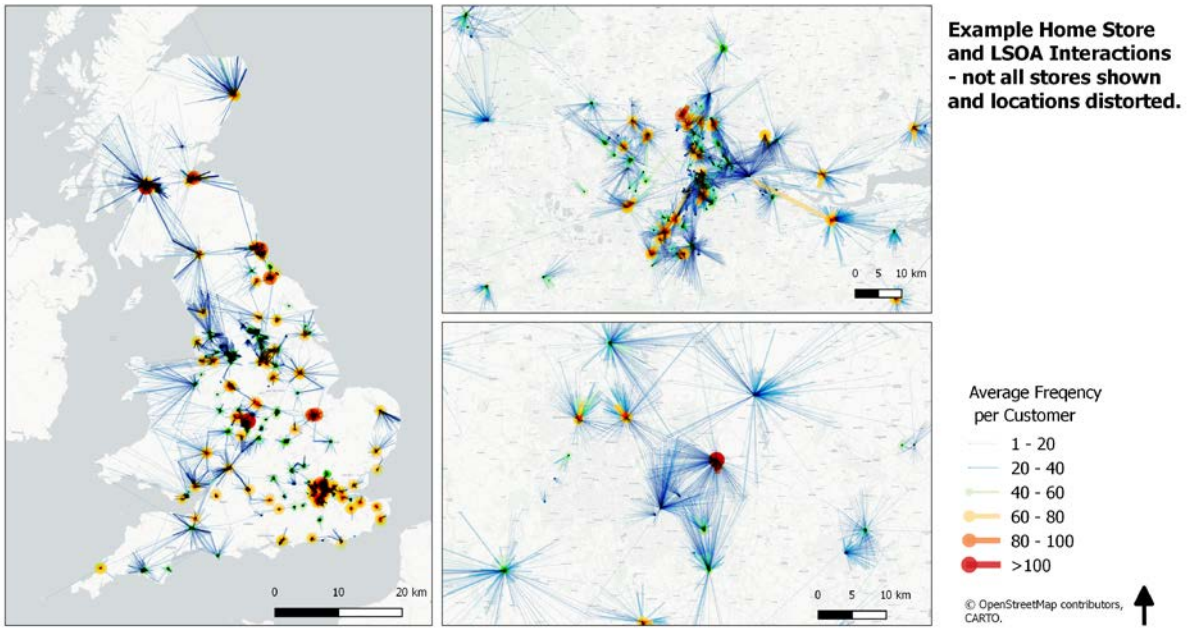

Figure 1 Home Stores for a sample of LSOAs.

Common activity area stores were identified using the 2011 UK Census1 Travel to Work flows representing origin LSOAs and destination OAs/SOAs (ONS 2011). These data, whilst collected for a different domain and population segment, are the most complete and openly available indicators of where people are when not at their place of residence. As consumers shop in continuous geographic space (Birkin et al. 2010), the destination flows were reconciled to the store network by aggregating worker counts for all OAs/LSOAs within a 5 mile radius of each store. If a store had workers within 5 miles, it was labelled as a common store.

Infrequently visited stores were made for each LSOA from all remaining stores. This resulted in a series of stores represented by the example store set extents in **Figure 2**.

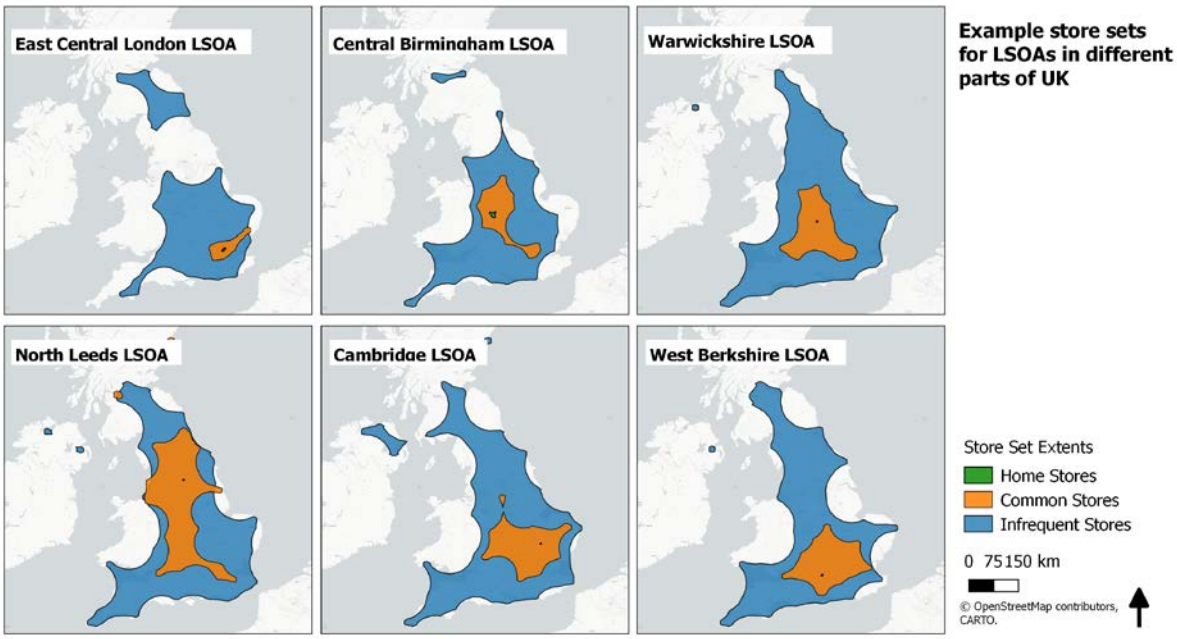

Figure 2 Store set extent examples.

2.2. Customer transactions and variables

For each customer (any active cardholder with more than 5 transactions within the 52 week data period), the LSOA-Store labels were joined to each transaction. The frequency of transactions conducted in each store set was calculated per customer. To get an understanding of incompleteness, transactions were grouped into fortnights (the average time between transactions for the most frequent visitors over the median frequency is 13 days). For each fortnight, a reduced classification or state was obtained, such as “Home” or “Common and Home”, representing the store sets visited in the fortnight by each customer. When a customer was not observed, an imputed value of “Not Observed” was generated.

The percentage of time spent in each state was then derived. The average distance travelled during weekdays and weekends, the number of different stores a customer was observed in and the number of times a customer shopped more than 5 times in a single day were also calculated as measures of promiscuity.

2.1. Clustering

With the variables created for each customer, a k-means clustering approach was taken owing to its relative simplicity (Alexiou and Singleton, 2015) and unsupervised nature. The focus of the clustering was to identify those who are mainly active over sustained periods of time within the infrequent store set, as well as those who have degrees of incompleteness within their transaction histories. To establish the number of clusters, bootstrapping followed by investigation of the adjusted Rand Index and Calinski-Harabasz index, as set out by Putler & Krider (2012) were used. This resulted in 6 clusters being retained for analysis.

3. Results

The clustering results indicate 6 clear distinctions between different types of customers. **Table 1** gives an overview of the characteristics.

Table 1 Cluster characteristics and counts

Cluster	Name	Count Customers (11,961,109)	Description/Characteristics
1	Home store shoppers	29%	Predominantly shop at home stores, smaller distances, and occasional shop at common or infrequent stores.
2	Atypical shoppers	3%	Shop in infrequent store set. Higher average distance travelled (80% over 50km weekday and weekend), median stores visited is 3.
3	Common store shoppers	12%	Shop in common stores, sometimes closer to home. Less than half fortnights observed in home stores. Some (6%) have average distance over 50km in both weekdays and weekends.
4	Mixed/switcher shoppers	9%	Switch between home stores and common stores (half of fortnights feature both). Higher number of stores visited on average (4). Shopping conducted closer to home.
5	Less frequent home store shoppers	25%	Similar to cluster 1 but observed in less fortnights, visit fewer stores and travel less. Up to 60% of fortnights not observed.
6	Infrequent shoppers	22%	Around ¾ not observed in half fortnights. Mix of average distances. Lower observations creates uncertainty within cluster.

To illustrate the types of locations that each cluster type is observed in when shopping, the example of the Birmingham local authority district is used. Figure 4 presents hex-bins showing the transaction intensity for stores within each cell for all LSOAs within Birmingham for the four more complete clusters (1-4). With each, there is a hotspot of activity around Birmingham. In the home stores cluster, common and mixed clusters, these three hex cells represent 98%, 97% and 86% of transactions respectively, highlighting the local dominance of shopping transactions. In contrast, these same cells represent just 5.5% of transactions for the atypical stores cluster. These customers transact across the UK, but are mainly concentrated in the wider Midlands area, London and the south east and other pockets across the UK.

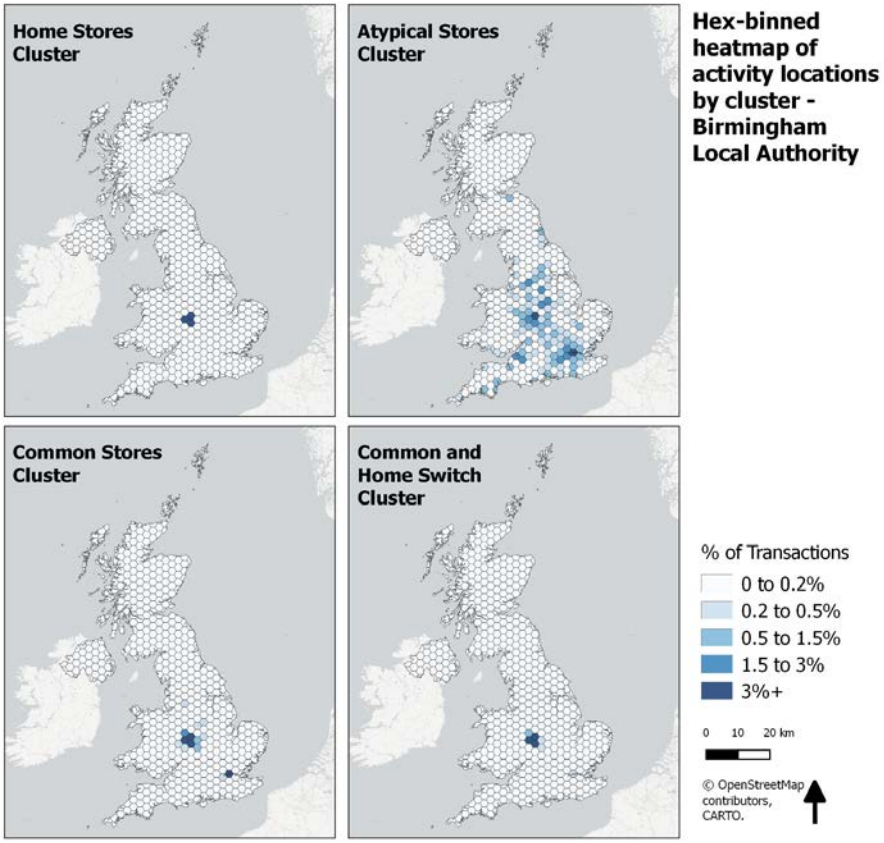

Figure 3 Example transaction locations for clusters in Birmingham

4. Discussion and Conclusions

The results show the detection of a small group of customers who have atypical patterns according to the location of stores that they visit compared to others within their neighbourhoods. This group typical travels further and are active mainly in the stores that we would expect to be visited only infrequently. The percentages of customers comes close to the percentage of customers that Lloyd and Cheshire (2017) flagged as uncertain (4%). In some cases, these may be genuine activity spaces that customers operate in, whilst in others the stated address details may be inaccurate. Flagging of these customers is useful for future re-use of the data to avoid incorrect generalisations, however this impacts relatively few customers.

Another outcome is the identification of around a quarter of customers for whom incomplete or unattributable purchasing behaviour is available. This incompleteness accords with the findings of Gijsbrechts et al (2008) that not all customers shop exclusively within a single retailer, nor do they always swipe their loyalty cards (Wright and Sparks, 1999). There is no way for this genuine uncertainty to be measured using a single retailer's loyalty card data. However, the approach taken shows that it is possible to distinguish incompleteness within the recorded loyalty card data to some degree. When reusing loyalty card data, consideration should be given to these aspects of the data in relation to the work being conducted. If using such data for more longitudinal work where complete histories are needed, a subset of customers may be more appropriate. However, if examining differences in spatial interaction, all customers may be more valuable.

Our research findings are of direct relevance to academics interested in consumer behaviour, as well as the retail industry itself. More broadly, the work can be seen as contributing to better understanding of the activity patterns that characterise different types of consumers, as part of wider investigation into spatial and social mobility.

5. Acknowledgements

This research was sponsored by a high street retailer, and was carried out at the ESRC Consumer Data Research Centre (ES/L011840/1).

6. Biography

Tim Rains is a 3rd year PhD student in Geographic Information Science at UCL. His research interests include spatio-temporal patterns of consumer behaviour and retail geography.

Paul Longley is Professor of Geographic Information Science at University College London, where he directs the Consumer Data Research Centre. His research interests are grouped around socioeconomic applications of Geographic Information Science and Systems.

References

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

Longley, P. A. (2012). Geodemographics and the practices of geographic information science. *International Journal of Geographical Information Science*, 26(12), 2227-2237.

Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. (2005). *Geographic information systems and science*. Chichester: John Wiley & Sons Ltd.

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.

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.

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.

Clarke, I., Hallsworth, A., Jackson, P., De Kervenoael, R., Del Aguila, R. P. & Kirkup, M. (2006). Retail restructuring and consumer choice 1. Long-term local changes in consumer behaviour: Portsmouth, 1980–2002. *Environment and Planning A*, 38(1), 25-46.

Jackson, P., Del Aguila, R. P., Clarke, I., Hallsworth, A., De Kervenoael, R. & Kirkup, M. (2006). Retail restructuring and consumer choice 2. Understanding consumer choice at the household level. *Environment and Planning A*, 38(1), 47-67.

Huff, D. L. (1964). Defining and estimating a trading area. *The Journal of Marketing*, 28(3), 34-38.

Newing, A., Clarke, G. & Clarke, M. (2013). Visitor expenditure estimation for grocery store location planning: A case study of Cornwall. *The International Review of Retail, Distribution and Consumer Research*, 23(3), 221-244.

Waddington, T. B., Clarke, G. P., Clarke, M., & Newing, A. (2017). Open all hours: spatiotemporal fluctuations in UK grocery store sales and catchment area demand. *The International Review of Retail, Distribution and Consumer Research*, 1-26.

Berry, T., Newing, A., Davies, D., & Branch, K. (2016). Using workplace population statistics to

understand retail store performance. *The International Review of Retail, Distribution and Consumer Research*, 26(4), 375-395.

Office for National Statistics, 2011 Census: Special Workplace Statistics (United Kingdom). UK Data Service Census Support. Downloaded from: https://wicid.ukdataservice.ac.uk

Birkin, M., Clarke, G., & Clarke, M. (2010). Refining and Operationalizing Entropy-Maximizing Models for Business Applications. *Geographical Analysis*, 42(4), 422-445.

Alexiou, A. & Singleton, A. (2015). Geodemographic Analysis. In Brunsdon, C. & Singleton, A. (Eds.), *Geocomputation: A Practical Primer*. (137-151). London: SAGE.

Putler, D. S., & Krider, R. E. 2012. *Customer and business analytics: Applied data mining for business decision making using R*. CRC Press.

Lloyd, A., & Cheshire, J. (2017). Challenges of Big Data for Social Science. *GISRUK 2017 Proceedings* (2017).