Challenges of Big Data for Social Science:

Addressing Uncertainty in Loyalty Card Data

Alyson Lloyd[*1], James Cheshire[†1]

[1]Department of Geography, University College London

December 8th, 2016

Summary

There is a fundamental need to better appreciate the dynamics and uncertainty of large consumer datasets, particularly if we are to utilise them to model social and geographical phenomena. This research aims to understand the potential and limitations of loyalty card data for research within the social sciences and humanities. Initial spatial analyses of customer postcodes and transactional behaviours suggested instances of consumer behaviour that may deviate from expectations, based on our existing knowledge of the UK population. A data-driven model was constructed to define and quantify this geographical uncertainty, drawing on knowledge and theory from multi-disciplinary domains.

KEYWORDS: Big Data; Spatial Analysis; Retail; Consumer Behaviour; Loyalty.

1. Introduction

Large-scale digital datasets have become increasingly ubiquitous in recent years and are routinely spatially and temporally referenced, leading to a fundamental shift towards data-driven geography (Miller and Goodchild, 2015). 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 metadata such as age, gender and postcode are collected, adding a dimension of demographic data that can be attributed to the transactional behaviours.

Big spatially referenced data have the potential to inform a broad spectrum of social, economic, political, and environmental patterns and processes (Graham and Shelton 2013; Kitchin 2014), by allowing us to capture spatio-temporal dynamics on a multitude of scales. However, many issues arise when applying these data in research, since many new forms of data are a by-product of alternative commercial agendas. Lack of researcher control means that these data may be susceptible to data error, which is of particular importance when considering data-driven geography applications. In the case of loyalty card data, locational attributes are entirely dependent on accurate human input and the motivation to update this information in the event of a location change. However, if inaccurate, these data have the potential to obscure, rather than reveal social and spatial processes (Graham and Shelton, 2013).

There is a fundamental need to better understand the dynamics and uncertainty of large

[*] Alyson.lloyd.14@ucl.ac.uk

[†] James.cheshire@ucl.ac.uk

consumer datasets, particularly if we are to utilise them to inform social and spatial phenomena. This research formed the preliminary stage of understanding the potentiality and limitations of loyalty card data for applications within the social sciences and humanities. An inductive approach was taken to understand data dynamics of a unique loyalty card dataset, obtained from a major UK retailer. Initial spatial analyses of customer postcodes and transactional behaviours suggested instances of consumer behaviour that may deviate from expectations, based on our existing knowledge of geographic phenomena. For example, customer store visiting behaviours being substantially far in proximity from their reported place of residence. Therefore, a data-driven model was constructed to define and quantify this uncertainty, drawing on knowledge and theory from multi-disciplinary domains. Development of such methods are not only important for the reliable adoption of Big Data in research, but also for retailers if utilising this information to inform location-based marketing strategies.

2. Data

Loyalty card transactional data was provided by a major UK high street retailer with a national network of stores, accounting for every transaction that had occurred between April 2012 and March 2014. Customer metadata consisted of demographic information including gender, date of birth and postcode for over 18 million customers. Transactional data included store of purchase, product type, amount spent and a timestamp for over 400 million records.

3. Method

Drawing on principles from research domains such as travel behaviour and spatial consumer behaviour, a methodology was constructed to define plausible and implausible store visiting behaviours based on a customer's reported geographical location. Figure 1 provides an overview of the methodology applied. Firstly, origin-destination flows were observed per small area using a trip distribution matrix, allowing analysis of the distribution of journeys from each small area across Great Britain. This was created by aggregating postcodes to Middle Layer Super Output Areas (MSOAs) and obtaining MSOA (origin) to store (destination) flows with the sum of unique customers (*T*) that had performed each pair.

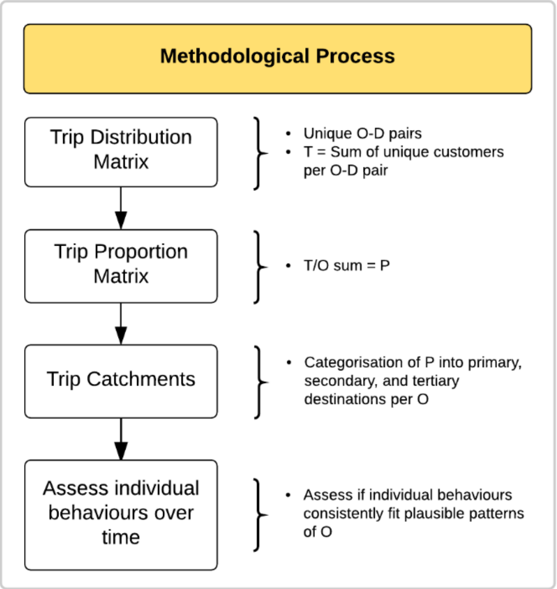
Methodological Process
Trip Distribution Matrix
• Unique O-D pairs
• T = Sum of unique customers per O-D pair
Trip Proportion Matrix
• T/O sum = P
Trip Catchments
• Categorisation of P into primary, secondary, and tertiary destinations per O
Assess individual behaviours over time
• Assess if individual behaviours consistently fit plausible patterns of O

Figure 1. Overview of methodological process.

Subsequently, trip distributions were converted into trip proportions, by dividing *T* values by their *O* sum (total number of customers per MSOA) to understand distributions in the context of each small area. Figure 2 demonstrates the distribution of visits between stores for customers from 3 example MSOAs in close proximity. Unsurprisingly, there are many overlaps in patronage across the store network, such as for large, city centre flagship stores (store 180). However, there are also unique local distributions of patronage per area.

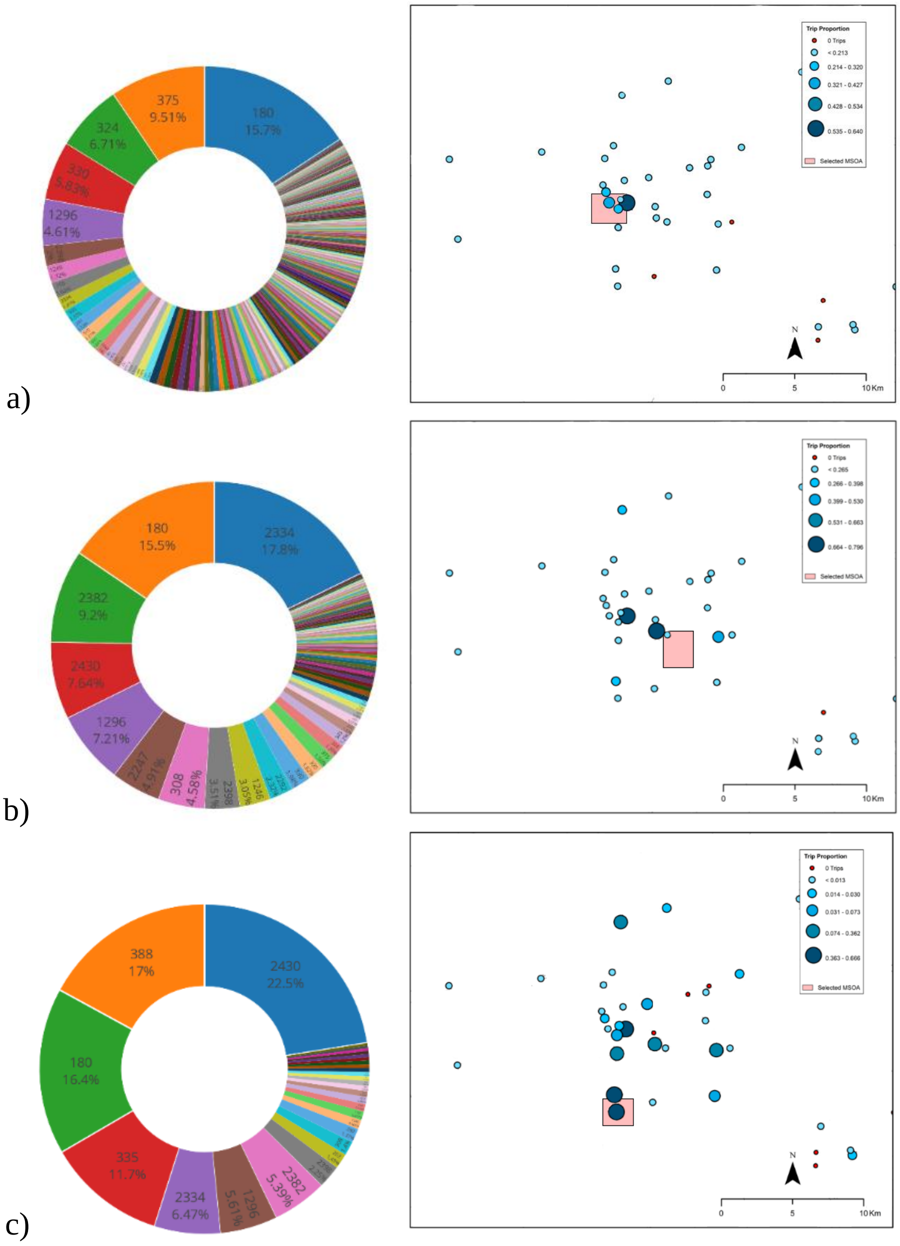

Figure 2. Example flow distributions for 3 MSOAs in Great Britain (further contextual information has been deliberately omitted to ensure the anonymity of the data provider).

These data were then categorised into the most and least likely store destinations per geographical area, by defining threshold values that represented the highest 70% (primary), 70-90% (secondary) and > 90% (tertiary) of flows. An algorithm was designed to detect abnormal behaviours based on customers' stated origin location and their transactional histories. Two fundamental patterns of error could be identified. *Postcode errors* were defined as customers who had never transacted at a primary store location, and *postcode changes* demonstrated a change in patronage behaviour within the time span of the data. These could typically be identified as a shift to a new network of stores that was outside of their registered area's primary destinations.

4. Results

Results revealed that 1.5% of the customer base demonstrated uncertain behaviour. Figure 3 shows travel flows from customers' origin MSOAs to store locations for one store type, using a) the uncleaned postcode data and b) the cleaned data.

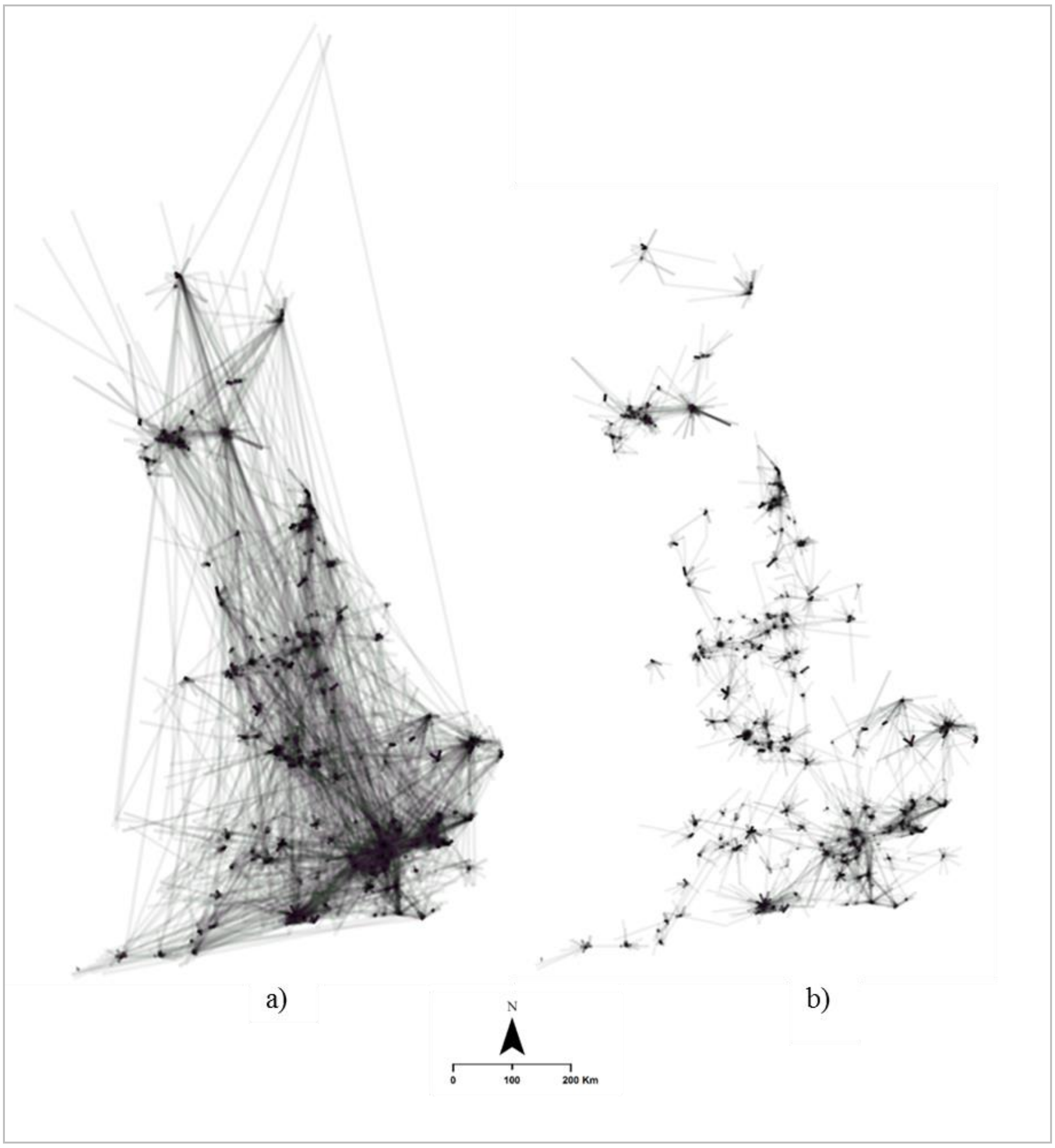

Figure 3. Flows from customer origin MSOA to store (one store type), using a) the uncleaned data and b) the cleaned data.

Applying this cleaning method produced flows that were consistent with what we may expect for this store type, which primarily serves local surrounding communities. In comparison to the uncleaned data, the majority of patterns that were inconsistent with our existing knowledge of spatial behaviour were identified and removed. However, this subset of data could be further utilised to estimate and analyse potential areas of relocation using their current store visiting behaviours. Figure 4 shows regional flows of these customers, demonstrating patterns consistent with migration flows outlined by the 2011 Census.

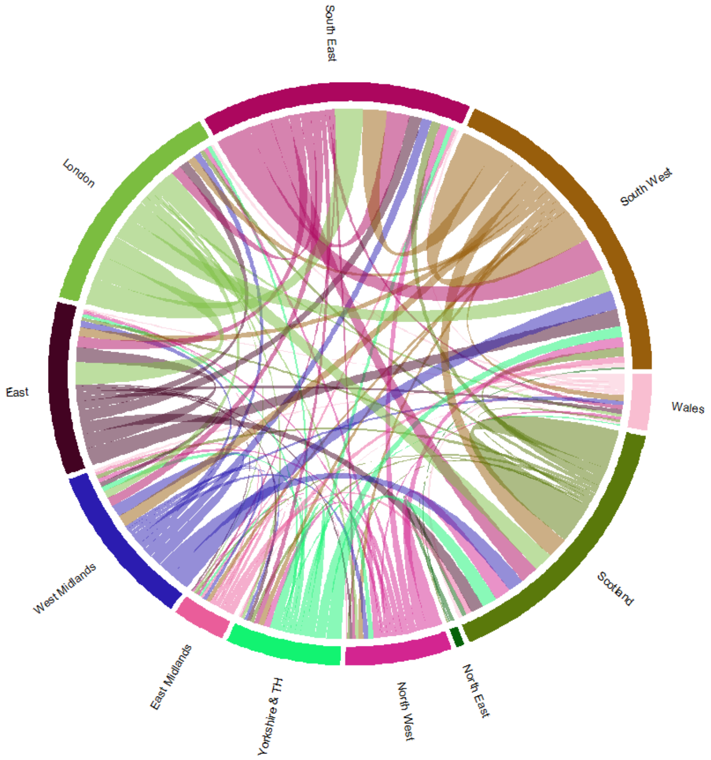

Figure 4. Regional flows of *postcode error* and *postcode change* customers.

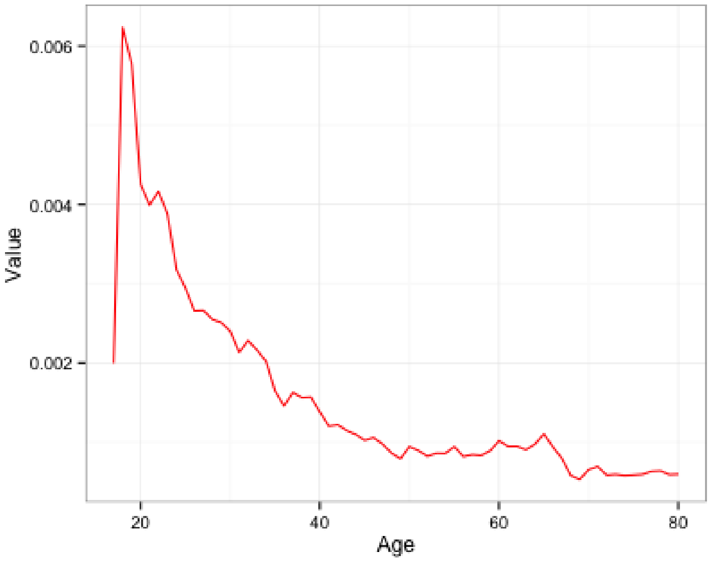

Figure 5. Age distributions of *postcode changes*, recorded at time of change point (normalised by total customers per year of age).

In addition, demographic analyses suggested that the risk of location change may be considerably skewed towards younger cohorts, primarily between the ages of 18-20 (see Figure 5). This could be indicative of the differing behaviours of life stages, such as leaving a family home or student migration. Younger cohorts may also typically exhibit more transient residential locations, increasing the risk of these age groups exhibiting postcode uncertainty.

5. Discussion

These analyses were able to utilise knowledge from multi-disciplinary domains to understand and model potential geographical uncertainty based on customers' stated origin location and their transactional histories. Results suggest that a segment of the population within postcode referenced data may be unrepresentative of a current place of residence, demonstrating the applications of data-driven methods to offer insights that would not be practically obtainable using traditional methods. Furthermore, these data indicated that they may be able to provide a level of insight into population characteristics such as migration flows – an important source of information typically only provided by traditional methods such as national Censuses. Development of such methods are not only important for the reliable adoption of Big Data in research, but also for retailers if utilising this information to inform location-based marketing strategies.

6. Future Work

This research formed the preliminary stage of understanding the potential and limitations of loyalty card data for social science research. Work will continue to improve the proposed methodology, by investigating optimum threshold values for categorising flows across small areas. However, future objectives will aim to understand the potential contributions of these data for informing social and spatial population dynamics.

7. References

Graham, M., & Shelton, T. (2013). Geography and the future of Big Data, Big Data and the future of geography. *Dialogues in Human Geography, 3*, pp. 255-261.

Kitchin, R. (2013). Big Data and human geography Opportunities, challenges and risks. *Dialogues in human geography, 3*, pp. 262-267.

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

8. Acknowledgements

This research is funded by the Economic and Social Research Council (ESRC).

9. Biography

I am a second year PhD student in the area of Retail Sustainability and Resilience. My research aims to understand the uses of Big Data for informing social and spatial population dynamics, consumer behaviour and retail insights.