Consumer Data Research Centre

An ESRC Data Investment

Challenges of Big Data for Social Science: Detecting Postcode Error in Loyalty Card Data

Alyson Lloyd and James Cheshire
University College London

Impacts

- Presents a novel data-driven methodology for detecting geographical uncertainty in loyalty card data.
- Demonstrates that a segment of customer postcodes within loyalty card data may be unrepresentative of a current place of residence.
- Insights may inform retailers utilising this information to inform location based marketing strategies.
- Contributes to the reliable adoption of Big Data in social science research.

Project Background

Large-scale digital datasets have become increasingly abundant in recent years and many have turned their attention to harnessing these for insights within the social sciences. 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 geo-demographics that can be attributed to transactional behaviours.

Despite this, these new forms of data are often produced as a by-product of commercial activities, leading to issues when attempting to apply them in research. Importantly, lack of researcher control over data collection means they may be susceptible to unidentified data error, which is of particular importance when considering spatial applications. In the case of loyalty card data, customer postcodes are entirely dependent on accurate human input and the motivation to update this information in the event of a location change.

It is vital that we explore the characteristics of these novel forms of data before using them to inform social and spatial phenomena. Working in partnership with a major UK high street retailer, this case study presents the preliminary stages of understanding the potentiality and limitations of loyalty card data for applications within social science research. Exploratory analyses of locational and behavioural data revealed instances of postcode uncertainty – where customer travel patterns were logically/geographically inconsistent with their reported place of residence. A novel methodology is presented that aims to quantify plausible and implausible store visiting behaviours through the linkage of their postcode and behavioural attributes.

Data and Methods

Loyalty card data were provided by a major UK high street retailer, representing two years of transactional data (2012-2014) across an expansive national network of stores. Transactional information included store of purchase, product type, amount spent and a timestamp for over 400 million records. Customer metadata included gender, date of birth and postcodes for approximately 18 million UK customers.

True interpretation of irregular behaviour in this context required knowledge of complex travel behaviours. For instance, travel patterns may not always fit with what appears geographically logical, due to incorporating store visits into daily obligations which can vary according to purpose (i.e. work, leisure or tourism).

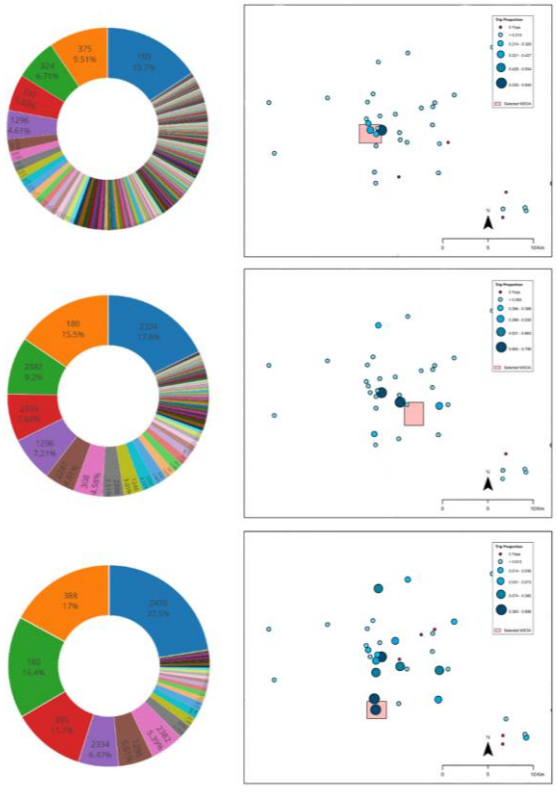

Figure 1. Example trip distributions for three MSOAs in Great Britain. Circular charts show store visiting proportions on a national scale.

The proposed methodology utilised data-driven techniques to quantify observed travel patterns, by calculating trip distributions between origins

Consumer Data Research Centre
An ESRC Data Investment

Challenges of Big Data for Social Science: Detecting Postcode Error in Loyalty Card Data

Alyson Lloyd and James Cheshire
University College London

(customer locations, aggregated to Middle Layer Super Output Area or MSOA) and destinations (store locations). This allowed analyses of the interactions between locations and therefore the most and least performed journeys per customer origin. Figure 2 demonstrates examples of the local and national trip distributions to store locations from three MSOAs in close proximity. Further contextual information has been deliberately omitted to ensure the anonymity of the data provider.

Using these distributions, unique thresholds to describe principal catchment areas could be defined for each MSOA across Great Britain. An algorithm was then designed to assess individual customer transactional histories over time and identify those who consistently performed behaviours outside of their catchment area. This method was also able to estimate a likely time of location change using transactional time-stamps, and estimate areas of potential relocation using their current store visiting behaviours.

Key Findings

Results suggested that a segment of the population within loyalty card data may be unrepresentative of a current place of residence.

Figure 2. Flows from customer MSOA to stores using a) the raw data, b) the cleaned data.

Figure 2 illustrates travel flows for one store type, that typically serve local surrounding communities, using a) the raw data and b) the cleaned data.

Outputs indicated that the method was able to detect travel patterns that were logically inconsistent with a reported place of residence, with approximately 4% of the sample demonstrating uncertain behaviour. Whilst this may be a small proportion, it represents a significant number of customers in relation to the sample size. Furthermore, Figure 3 shows the regional migration flows of these customers using their estimated areas of relocation, demonstrating patterns consistent with migration flows outlined by the 2011 Census.

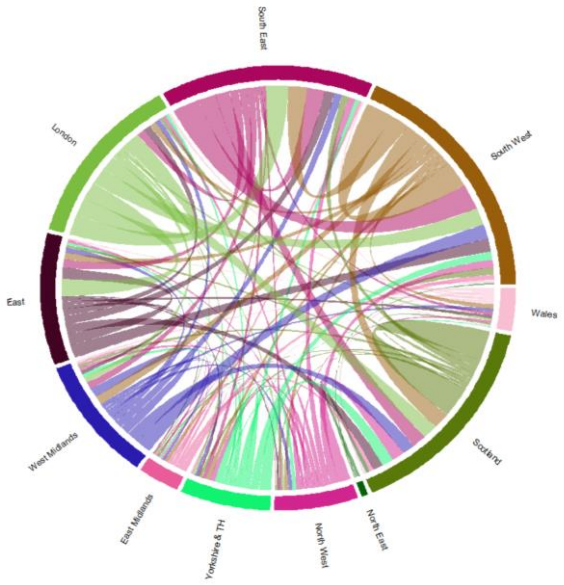

Figure 3. Regional flows of moved customers.

Demographic analyses of this segment also suggested that the risk of postcode uncertainty may be considerably skewed towards younger customers. This could be indicative of the differing behaviours of life stages, such as leaving a family home, student migration or younger individuals typically exhibiting more transient residential locations.

Future Directions

Future work aims to continue to optimise this methodology, understand the results in terms of geo-demographic characteristics, and continue to investigate the potential contributions of these data for informing social and spatial population dynamics.