Consumer Data Research Centre

An ESRC Data Investment

CONSUMER DATA RESEARCH CENTRE

Masters Research Dissertation Programme

Case Studies 2017

Edited by Guy Lansley

13 October 2017

www.cdrc.ac.uk

Consumer Data Research Centre

An ESRC Data Investment

Foreword

The Consumer Data Research Centre's Masters Research Dissertation Programme instigates several student led research projects which seek to tackle topical problems put forward by industry. Each year we invite representatives from major retailers and organisations which handle consumer data to propose research ideas to Masters level students. Students from all UK institutions are eligible to apply to undertake the projects via the CDRC website. Successful students complete the research over the summer with joint supervision from their academic tutors and their industry sponsor.

This year we welcomed 16 students from a diverse range of academic disciplines. Between them they used a wide range of software packages and analytical techniques in their research.

A selection of short summaries of their research have been provided in this document.

If you are interested in becoming an industrial partner or are a student wishing to find out more please visit https://www.cdrc.ac.uk/retail-masters/ or contact Guy Lansley g.lansley@ucl.ac.uk.

An exploratory analysis of the temporal fluctuations in footfall around Hammersmith town centre in relation to Business Improvement District events

Adjoa Adu-Poku[1], Michael G. Epitropakis[1] and Livia Caruso[2]
[1]Lancaster University, [2]HammersmithLondon BID

Project Background

HammersmithLondon BID seeks, as part of its aims as a Business Improvement District, to promote the attractiveness of the Hammersmith area. A result of the promotion includes an improved visitor count, which is understood to offer local businesses greater selling opportunities. One way it aims to increase the appeal of the area is through locally-hosted events that occur mainly in the summer and winter periods. This project is responsible for organising temporal fluctuations in footfall, in such a way that HammersmithLondon may recognise how their events could be contributing to the cause of pattern changes. This project is an exploratory analysis that applies a range of tools in order to make inferences from a set of time series data. The tools used are implemented effectively in the existing literature in a wide range of applications.

Data and Methods

The data consists of an hourly count of pedestrians, captured throughout one year, across 6 wi-fi sensors located across the central and western areas of the BID. The data can be presented as a time series, which sees fluctuations, as well as observations that may be considered anomalies.

The 6 tools applied to the data are 1. Time series decomposition, 2. Change point detection (via a binary segmentation algorithm), 3. Anomaly detection, 4. Point data mapping and animation 5. Point data interpolation and 6. Principal component analysis. These tools were employed through packages on R Studio software and ArcGIS.

The first 3 of these tools compose a time series analysis, which allowed the comparison of positive anomalous footfall events to a calendar of HammersmithLondon events, in order to analyse for correlations. This allowed the project to make inferences about which events, or types of events, served as a part of the cause of high footfall counts. The final 3 tools made up the spatial analysis element. These 3 methods allowed us to see the relative importance of the sensor data in each location, and justified the omission of 3 sensors from the time series analysis.

Key Findings

Figure 1 presents an example of those positive anomalies detected from the time series data of one sensor. Through joining these results with corresponding specific events falling within the anomalous dates and times, and change point detection results to specific weeks and seasons in the year, the project was able to determine the most important events that HammersmithLondon hosts.

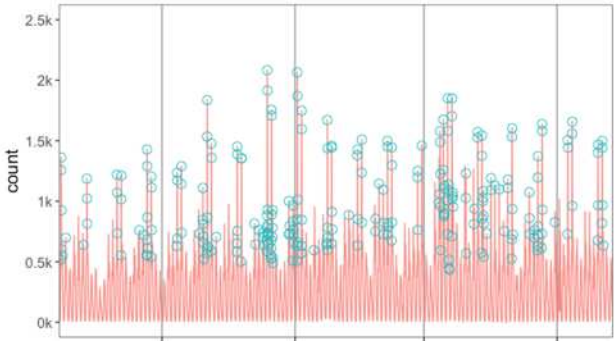

Figure 1. Anomaly detection graph for the 'Lyric Square' sensor

The empirical results indicate that BID-hosted 'Yoga in the Square' and 'Wimbledon in the Square' events are most likely to cause abnormally high footfall rates. Results are also explained in the context of categories. Of these categories, 'Big Screen' and 'Fitness, Health and Wellbeing' contained the most significantly correlated events to positive anomalous footfall results. 'Virtual Reality Cabin' and 'Hammersmith Movie Hub' events, in particular, were found to correlate the least with anomalous results. Change point detection in the Lyric Square trend level also indicated improvements in the footfall counts during the BID Summer Festival period.

Value of the Research

The results of this study have given the partner some evidence to suggest that certain events are drawing in more visitors than others, helping to justify spending in specific areas. HammersmithLondon can also justify investment into an increased number of footfall sensors in the area to improve the spatial analysis. The study also provides a positive case for the implementation of electronic footfall sensors for industries which are concerned with visitor counts as a response to promotional activities or events.

Consumer Data Research Centre

An ESRC Data Investment

Sainsbury's

Customer pathway and in-store activity

Ulzhan Bissarinova[1], Daniel Hulme[1], Constantinos Gavriel[2] and Emilie Lhomme[2]
[1]University College London, [2]Sainsbury's

Project Background

It is desirable for retailers to understand the pathways customers use to move around the store while purchasing their products. Pathway information gives insights on the space occupancy rate and might allow store plans to be reallocated more efficiently. The most promising state-of-the-art indoor tracking methods involve usage of closed-circuit television cameras (CCTV), Wi-Fi, Bluetooth, ultra wide-band, Long-Term Evolution, inertial measurement units and near-field communication, whose work principle is based on electromagnetic fields. These methods come with a high cost of implementation, as well as inaccuracy problems related to the distortion of signals. Unlike the signal-based methods, the method proposed in this work does not require involvement of any additional hardware and relies purely on transactions and floor-plan data.

Data and Methods

The floor plan of one specific store along with its transaction data recorded during a four week period were analysed for this research. The analysis included obtaining x, y-coordinates of customers along their predicted paths, creating clusters of similarly visited sections and generation of association rules for the products frequently bought together. The family of shortest path algorithms was considered to estimate the potential path of clients, among which included Dijkstra's algorithm. A K-means algorithm was implemented to form clusters based on the numbers of "views" and sales volume of each section. Number of "views" indicates how frequently a particular section was passed by and therefore was viewed by a visitor. Finally, the association rules were used to identify strongly related product pairs and to observe how distant those products are located from each other.

Key Findings

Aggregated pathways show that overall, mornings were less busy and the store was busiest during the lunchtime rush. A clear path that was followed by the majority of the visitors during lunchtimes on weekdays included sections "Snacking & Sharing", "Fresh Pizza Bread & Pasta", "Fresh Fish", "Fresh Poultry", "Bacon", as well as "Yogurts" and "Cheese". Figure 1 illustrates the strength of related products against the distance between them. The product pairs that have high confidence and high distance values could potentially be placed more closely together to improve their sales. The top 3 examples with highest confidence and traveled distance values (distance>164m, confidence>0.75) include sections "Fresh Fish" -"Veg & Salad", "Fish Counter"- "Veg & Salad", "Potatoes" – "Veg & Salad".

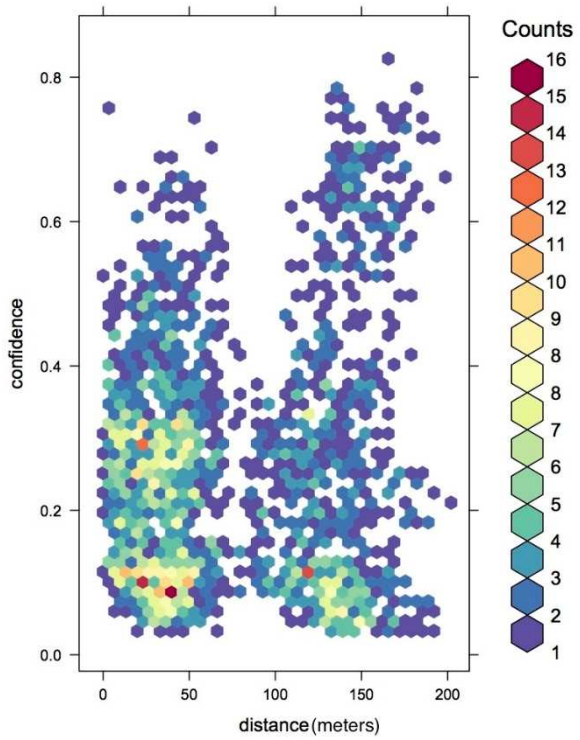

Figure 1. Confidence of associated products (>0.01) against distance between them

Value of the Research

This methodology has several benefits. First of all, it presents a tool to visually represent the customer pathways within a given store. This can be used to estimate how customers navigate a store and occupancy rate of areas according to seasons, days of week and times of day. This insight could be used to create recommendations on better store layouts and to understand the effectiveness of promotions. A better understanding of customer pathways may allow stores to increase the sales of complementary products and impulsive purchases. In addition, it could be used to minimise congestion and improve customer satisfaction.

Generally, the methods discussed in this work are applicable to all types of retail stores where consumers browse items across the store floor. The core advantage of the approach is that the retailers don't need to invest in expensive (and potentially undesirable) tracking technologies and can instead take advantage of data which are routinely collected by transactions.

Consumer Data Research Centre

An ESRC Data Investment

SHOP DIRECT

Customer value modelling – how sensitive is the model to change?

Aidon Blong[1], Simon Maskell[1], Nicola Dunford[2] and Tony Birch[2]
[1]University of Liverpool, [2]Shop Direct

Project Background

This dissertation explores how sensitive different implementations of a customer value model within the retail industry can be. Existing linear stepwise regression models that are currently in place are able to predict how valuable a segment of customers are going to be worth in the future, more specifically in terms of a monetary aggregate value. The purpose of the project is to better the existing models in terms of lowering the overall error score of the predicted values.

Data and Methods

The project used the same data set that the partner's existing customer value model uses, which ensured fairness when comparing different models. The dataset consisted of many different categories of data such as purchase history, browsing history and other key statistics regarding customers. Hypothesis testing was used as a means of proving and disproving certain statements of belief about the project. Initial hypothesis outlined that experiment 3, which used a random forest version of customer value model, would yield a lower overall error score compared to the original customer value model. Statistical significance was used to assess whether each customer value model held a statistical significance over the other.

Key Findings

The initial hypothesis was proved true, and the random forest implementation of the customer value model was able to outperform the existing linear regression model in terms of RMSE (Root Mean Square Error) score. Other error scores such as MSE (Mean Squared Error) have also been applied to the different algorithms, with no significant difference in terms of the clear winner.

However in terms of the accuracy of the models; linear regression outperformed random forest. Note that accuracy is not the same as 'classification accuracy' and simply defines how close the predicted target value aggregate is to the actual target value. The accuracy and error impact is something which is commonly known as a bias-variance trade-off; which given more time to investigate could explain why the two algorithms score differently for both error and accuracy. Experiment section 2 detailed multiple different attempts to select the 'best' variables from the data set, relative to the target variable. Some of the methods included Pearson/Spearman correlation, allowing for linear/non-linear relationships, as well as different entropy measures. Each of the experiments conducted are outlined within Figure 1.

Experiment	Average of RMSE	Average of Accuracy
1 - Stepwise	163.37	99.59%
2.1 - Information Gain	163.79	100.68%
2.2 - Info Gain Max	167.54	100.36%
2.3.1 - Pearson 90%	163.36	100.20%
2.3.2 - Pearson 80%	165.57	100.35%
2.3.3 - Pearson 70%	169.40	100.24%
2.4.1 - Spearman 90%	163.89	100.23%
2.4.2 - Spearman 80%	164.70	100.22%
2.4.3 - Spearman 70%	170.76	100.70%
2.5 Top 13 Variables	173.13	100.05%
3.1 Random Forest 77k	161.10	93.00%
3.2 Random Forest 232K	153.22	93.32%
Average	**164.99**	**99.08%**

Figure 1. The overall average scores per 10-fold cross-validation on the test data sets

Value of the Research

The research conducted provided some valuable feedback for Shop Direct to review; even to consider changing their existing model. The project provided Shop Direct with an insight as to how a random forest can outperform linear regression in terms of error score, however, there are several different methods such as BART (Bayesian Additive Regression Trees) which have the potential to yield even better results.

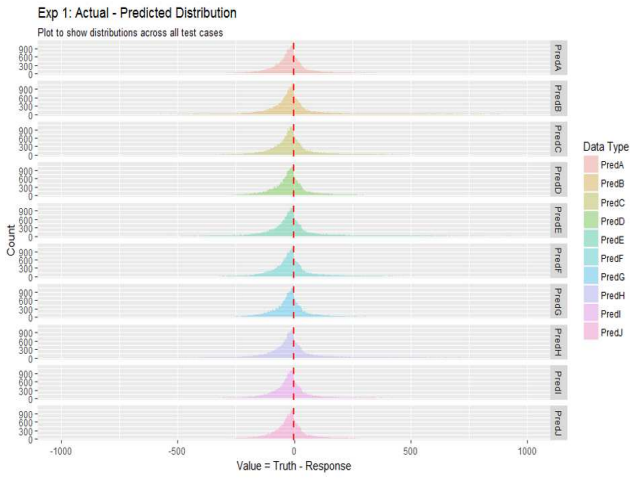

Figure 2. Experiment 1: Stepwise linear regression

Consumer Data Research Centre

An ESRC Data Investment

Sainsbury's

Estimate impact of slot availability on customer demand by using choice-based demand models

Yifan Cao[1], Arne Strauss[1] and Matthew Pratt[2]
[1]The University of Warwick, [2]Sainsbury's

Project Background

The increasing sales through online channels has made large supermarket retailers seek new strategies to meet this surging customer demand in online grocery shopping. However, no matter what new service retailers aim to provide on e-fulfilment, it is important to strike the balance between delivery efficiency and customer satisfaction. A suggested approach is to adopt revenue management techniques to help better understand customer choice behaviour towards delivery time selection considering slots availability and their prices. This research sought to demonstrate that it is useful to analyse the impact of slot availability and price on customer demand for delivery services. It also presented methods to assist the planning of resource allocation and capacity management policies to improve the efficiency of services.

Data and Methods

Data used for this research was collected from Sainsbury's nationwide online transaction records over a period of 26 days, from March 12th 2017 to April 6th 2017. It included information considered relevant for customer decisions, such as slots availability and their prices. The provision of slot information for non-purchasing customers gave a better theoretical basis for the analysis since it avoided the approximation for the non-purchase behaviours.

Two independent multinomial logit (MNL) models were used to estimate the complete delivery slot selection process, one focused on the day whilst the other considered the slot selection. Although both models are independent, their results can be combined to calculate the likelihood of a customer choosing any slot throughout the week. Factors that were used as variables for the delivery day selection MNL model were capacity ratio (% of available slots on a delivery day), price average and customer arrival day. Factors that were assumed to affect slot choices were slot availability and their prices.

Two validation methods were used to assess the predictive powers of the models. The first method compared the modelled results between the estimation data set and validation data set, while the second approach compared the computed model results with real data of slot selections.

Key Findings

The results from both MNL models provided significant insights on customer choices. The results from the delivery day selection model indicate that when all slots are available, most demand fall on the first three days (78%) upon customer arrival, and there is a 13.72% chance that a customer will not pick any slot. As the capacity ratio of each delivery day is lowered in turns, this probability rises but does not exceed 20%. However, it is indicated that a decrease in capacity ratio of delivery days closer to customer arrival results in a greater negative impact.

The delivery slot selection model results comprised two parts. Firstly, without considering price, findings identified that customer behaviour towards delivery slots varies significantly between weekdays (Monday to Thursday) and weekends (Friday to Sunday). Morning slots were found to be more popular than the others during the weekend, while early afternoon sessions were more popular through weekdays. After introducing differentiated prices as a leverage to smooth demand, the choice behaviours were found to be altered and more balanced. Another interesting finding was that with and without price impact, half hour slots (e.g. 13:30 – 14:30) were found to be usually less popular than their closest hourly slots (e.g. 13:00 – 14:00).

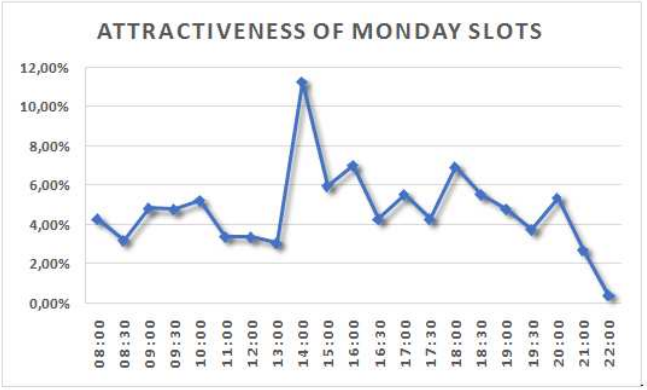

Figure 1. Estimated attractiveness of Monday Slots

Value of the Research

The impact of slot availability on demand can be inferred by combining results from the delivery day and delivery slot selection models with demand forecasts. These insights are valuable in daily delivery capacity management and for improving service delivery. For instance, the results can be used for an effective resource allocation policy focusing on client satisfaction and delivery capability. The results also provided further evidence on MNL models suitability in estimating consumer behaviour, especially in the field of e-fulfilment where validation tests are rarely published.

Consumer Data Research Centre

An ESRC Data Investment

How competitive is propensity score matching using 'address embeddings' with supervised classification for record linkage tasks?

Sam Comber[1], Dani Arribas-Bel[1] and Ronald Nyakairu[2]
[1]University of Liverpool, [2]Local Data Company

Project Background

The growing availability of Open Data has opened the opportunity for businesses to enrich their existing insight from data sources. In particular, a competitive advantage exists in the analysis of integrated data compared to analysing data in isolation. Yet, in reality, most real-world databases are noisy, inconsistent or replete with missing values; these issues complicate the integration of data. The process for disambiguating record pairs that represent the same entity into matches and non-matches is known as record linkage.

In this project, an innovative technique is compared with a supervised method for predicting the match status of address pairs.

Data and Methods

Our record linkage methods are tested on commercial address databases maintained by the Local Data Company (LDC) and Valuation Office Agency (VOA). Given the Cartesian product of the 700,077 and 1,978,830 records for the LDC and VOA databases creates a 1,311 comparison space of candidate record pairs, we require the addresses to be 'blocked' due to computational complexity. This means we partition the number of comparisons for linkage to within mutually exclusive blocks by postcode value.

Once partitioned, two methods for linkage are employed. Our innovative technique uses propensity score matching on the vector representations of addresses learnt from a neural-based language model. Here a 100 dimension vector is learnt using the word2vec algorithm to create an 'address embedding' for every address. Propensity score matching on these vectors is then used to facilitate the linkage.

Our second method begins with using the libpostal address parser to segment each address string into feature columns – for example, business name, street number, street name. These features become the basis for generating 'comparison vectors' for each candidate record pair. Here, each element in the vector is a similarity score between the features columns. This score might, for example, indicate how similar two street names are from one another. The comparison vectors are then classified as matched or non-matched using decision tree ensemble learners. These models are trained on address pairs with a known match status obtained from a previous round of matching between LDC and VOA addresses.

Key Findings

Precision, recall and F1 scores in Table 1 are interpreted as percentages. Our ideal outcome is to minimise misclassification error – i.e. the number of false positives and false negatives. For this reason particular attention is paid to the F1 score which balances the trade-off between false negatives and false positives. As shown, the random forest and XGBoost model far outperform propensity score matching on address embeddings.

Classifier	Precision	Recall	F1 Score
Logit	0.993	0.989	0.991
Random forest	0.994	0.996	0.995
XGBoost	0.994	0.996	0.995
PSM	0.000	0.000	0.000
PSM (blocking)	0.144	0.245	0.181

Table 1. Quality metrics for classifications

Value of the Research

The untidiness of Open Data complicates the process of linkage. Innovation in the methods employed for record linkage problems has the potential to improve the richness of knowledge discovery from data sources when this data can be successfully linked.

Our proposed method for learning the vector representations of addresses for linkage via propensity score matching was only able to match 18.1% of address pairs. Nevertheless, our supervised classification workflow was able to match up to 99.5% of addresses correctly. We attribute this to libpostal's segmentation of addresses into accurate feature columns for comparison. In summary, the availability of open source libraries such as libpostal represents an opportunity for companies to achieve high quality match rates for record linkage tasks.

An evaluation of the relationships between construction developments and retail vacancy

Matthew Grove[1], Les Dolega[1], Ronald Nyakairu[2], Adam Valentine[3] and Steve Shelley[3]
[1]University of Liverpool, [2]Local Data Company, [3]Barbour ABI

Project Background

Vacancy rate is considered to be a key indicator when assessing retail centre performance, yet very little literature has been produced evaluating the relationship between construction developments and retail vacancy. Suggestions have been made that developments in surrounding areas may impact the performance of a retail centre, yet no previous studies have undertaken quantitative analysis of this. An increase in the proportion of vacant retail units is generally linked to economic decline within an area, therefore establishing the factors that influence vacancy is valuable. This study explores this novel concept and attempts to establish the strength of the relationships present between data provided by Barbour ABI and the Local Data Company.

Data and Methods

The relationships between construction developments and retail vacancy within the area of Greater Manchester were assessed using multiple linear regression and geographically weighted regression. Fixed effects were included within the linear model to account for spatial heterogeneity. Four contrasting drive distance catchments were computed using road networks and retail centres and their performance evaluated across four different time-span groupings. From the options evaluated, it was determined that variable drive distance catchments relative to retail centre size best represented the data over the course of a bi-yearly twelve-month timespan. Four categories of construction developments were then evaluated in relation to Comparison, Convenience, Leisure and Service vacancy, alongside a variable that accounts for retail vacancy collectively. Model fit for both linear regression and geographically weighted regression was assessed throughout, in order to establish the strength of the relationships present.

Key Findings

The results obtained within this study are complex, allowing for generalisations to rarely be made across each of the variables and areas. Residential developments predominantly yield a positive relationship with reduction of retail vacancy, with Stockport being the only area to display a negative relationship. Results from human amenity, industry and transport service developments vary considerably, emphasising that the relationship between construction developments and retail vacancy is not unanimous across time and space. The level of detail and complexity within the results of this study suggest that the relationship between construction developments and retail vacancy cannot be generalised to all areas, and should be considered on a case by case basis in future studies.

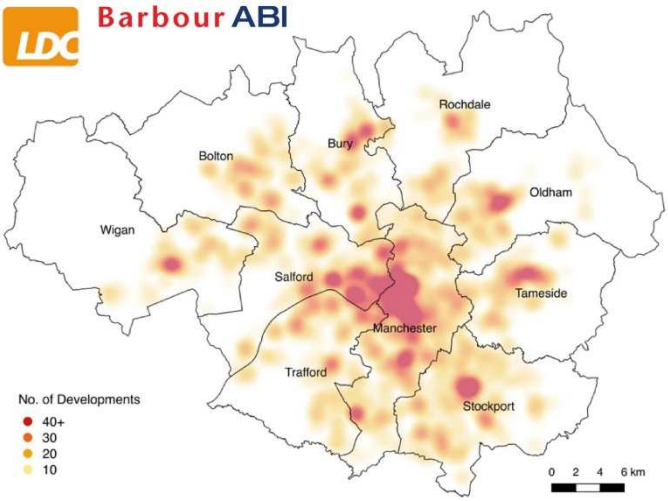

Figure 1. Construction developments between 2011 – 2017 within the area of Greater Manchester

Value of the Research

The exploration of this novel concept has given insight into an area that has received relatively little attention within previous literature, allowing for a better understanding of the factors that influence such a key element of retail health. Contrasting two datasets from inherently different sectors has allowed for relationships to be established that were not achievable before. This not only allows for better insight into the relationship between construction developments and retail vacancy, but also allows for a more pragmatic approach to be undertaken when assessing future developments.

Electric vehicle charge point placement optimisation

Jon Haycox[1], Guanpeng Dong[1], Jeeah Hwang[2] and Katie Hendry[2]
[1]University of Liverpool, [2]E.ON UK plc

Project Background

There is increasing demand for plug-in and hybrid vehicles in the UK as they are becoming more affordable due to decreased battery costs and lower manufacturing costs. In order for the growth in electric vehicles to continue, consumers need to be reassured that there is adequate charging provision, with a recent survey showing almost half of UK consumers are worried that they would not be able to find an available, working or compatible charge point (CP). E.ON is an international, privately-owned energy supplier. As part of its sustainability programme, coupled with the increasing need to focus on the environment, E.ON is looking to assess the market for new public charging points across E.ON sites and the feasibility of incentivising its business customers to operate public charging stations.

Data and Methods

A regional-level optimisation model is considered for Great Britain, with a full analysis performed for the London and North East regions. Points of Interest (POIs) data are exploited in the analysis of CPs because they are potential locations where charging is needed, i.e. where people are likely to drive to and spend time. The number of POIs within walking distance of existing CPs have been used to rank the POI categories in terms of importance. Workplace population and night time population from the 2011 census, and traffic data and EV registration locations from the Department for Transport were also considered as potential demand indicators for CPs. The data for these variables was resampled to a 1 km grid so they could be joined with the POI data to build a box factor measure representing demand for CPs in a cell. The maximum coverage location problem was solved for the demand function to obtain candidate cells that maximise the total demand coverage such that the overall demand is covered in the most effective way. This was performed using the lpSolve package in R and displayed using ArcMap. A sensitivity analysis was then performed to determine the best variables at matching existing CPs.

Key Findings

From our statistical models, a positive correlation was found between workplace population and CP intensity, which is consistent with the theory that people are likely to charge their cars in places where lots of people work, as well as at their workplaces. The model combining workplace population and POI in the analysis, with the weighting two-thirds to one-third respectively, has shown to be quantitatively the best for the London region (Figure 1) and North East region. In both areas, the model selecting candidate sites appears to match the actual CP locations best in the area with highest workplace values, i.e. in the dense urban areas. In the suburbs of London the pattern is very similar but there is a poor match for the rural areas in the North East. A simple proximity analysis has shown that even in a dense, urban environment like London there are candidate cells that are nearly 3 km away from an existing public site.

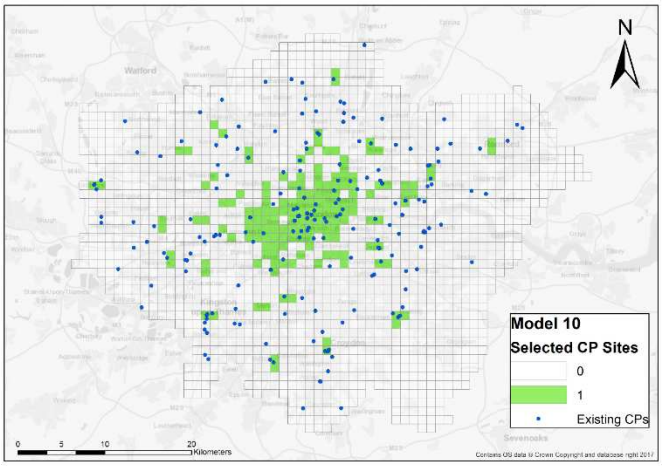

Figure 1. Optimised CP locations (green cells) and existing CP locations (blue points) in London.

Value of the Research

This study has successfully produced a proof-of-concept method for optimising the location of CPs at a regional level and builds upon previous work by including new datasets, such as EV registrations, in order to generate better insight. The methods presented in this study provide a way for E.ON to identify candidate B2B sites that are suitable for further local- and street-level investigation. This research also supports a major energy company in improving the future suitability of the UK infrastructure to electric vehicles and, therefore, supports carbon footprint reduction initiatives.

Consumer Data Research Centre

An ESRC Data Investment

SHOP DIRECT

Pricing optimisation - a key tool in successful retailing

Erin Knochenhauer[1], Graham Clarke[1] and Christian Setzkorn[2]
[1]University of Leeds, [2]Shop Direct

Project Background

Few decisions impact the success of a retailer more than product pricing optimisation. It requires balancing a multitude of tradeoffs which together determines a company's success. The decision maker has the difficult task of selecting the right price when faced with multiple, competing objectives such as maximising profits and revenue whilst minimising the price distance between competitors and maintaining proper inventory levels. Optimal prices are essential to the marketing mix and to the business as a whole. An inaccurate price could lead to significant forfeited profit, inflated inventory management costs from excess stock or low consumer satisfaction due to stock shortages. Thus, the main objective for this project was to explore methodologies that produce sets of optimal price alternatives for a sample of products at a given time at the individual product level.

Data and Methods

Non-dominated Sorting Genetic Algorithms (NSGA) were used for this project for their efficient, flexible methodology and limited input data requirements. The project first required calibrating the initial demand forecasting models using daily historical sales data at the individual product level to better predict revenue and profits in the ultimate goal of using their estimated parameters as revenue and profit maximising objective functions in the pricing optimisation. The competing objective entailed minimising the distance from the average competitor price. Sets of time-dependent optimal price solutions were produced for 13 electronic products within Shop Direct's 'Very' fascia utilising various functional forms to model revenue and profit. The pricing optimisation was completed using Liger, an open-source optimisation environment, and Python. The demand forecasting was completed with R. This study utilized both NSGA-II and NSGA-II-PSA algorithms, the latter is a partition-based selection algorithm.

Key Findings

For the Electronics trading department as a whole, the linear-log demand model, including price, promotion, a price-promotion interaction term and seasonal dummy variables as predictors of demand, had the best out-of-sample prediction accuracy. The log-log demand model consistently showed evidence of overfitting, and the additional terms of competitor price ratio and substitutes price ratio as well as their measures of variance did not greatly reduce out-of-sample prediction error. Linear models are not the most robust choice of demand modelling, as predictions errors are inflated, however, their low complexity is a useful choice as objective functions in pricing optimisation for scalability. Depending on business needs, prices could be updated daily, weekly or monthly by changing which historical data are used to forecast demand. Finally, there was no significant difference between the performance of the NSGA-II-PSA algorithm compared to NSGA-II when deriving optimal prices for the electronic goods.

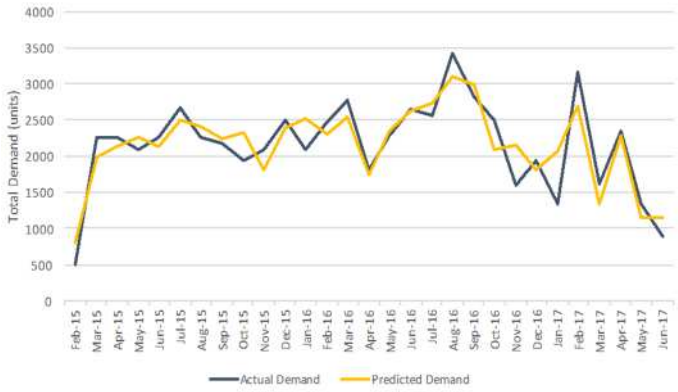

Figure 1. Actual (blue) vs. predicted (yellow) demand for a particular product

Value of the Research

This project served as a Proof of Concept utilising Genetic Algorithms for the multi-objective optimisation problem of setting optimal prices. This methodology produces a range of optimal solutions the Decision Maker can choose from promoting competitive advantage within the dynamic electronic market. Though a small sample of products was assessed, the results can be generalised to the greater electronics department. Further study is required to generalise to other product categories before scaling up to the entire product portfolio. Overall, this project was a valuable base on which to build Shop Direct's in-house pricing optimisation workflow.

Consumer Data Research Centre

An ESRC Data Investment

CLUTTONS

Exploring submarket trends in London: a constrained spatial interaction modelling analysis

Neelam Kundi[1], Adam Dennett[1] and Faisal Durrani[2]
[1]University College London, [2]Cluttons

Project Background

This study investigates the characteristics of the London housing market and the effects of variables such as house price and accessibility on commuter distance through spatial interaction modelling. Observations from Cluttons data show that the housing market in London is characterised by high prices and turnover. Furthermore, it may be that on closer inspection of the capitals' regions there could be more desirable areas for people to live in to commute to work. This is a complex market to delve into. A new geography is formed to provide an improved outlook to niches within London, offering another perspective on areas and allowing for Cluttons and census data to be utilised in new ways to uncover and explore microtrends for a deeper understanding of a remarkable urban centre.

Data and Methods

The submarket categorisation utilised by Cluttons for real estate grouping and analysis form the basis of the methodology. As this geography has not been represented before, a map has been created using thiessen polygon analysis and overlaid with Census boundary data at the medium super output area level.

To explore the relationship between house price and migration flows, constrained spatial interaction models are used. Variables for the interaction model are arranged to ensure the coefficients obtained from the model are informative. Variables determined to be most useful include an accessibility index, relative medium house price and mean net weekly income (destination to origin ratio).

Attraction constrained models allow insight into the number of people that have left a given area, and aims to explore which variable explain people's destination choice. Production constrained models allow insight into the number of people arriving at a destination, and aims to explore which variables explain their origin. The two constrained models form one set - three sets of these models are run, the first explores the entire research area which consists of Submarkets and Travel to Work Areas (TTWAs). The second explores only submarket areas using all variables. The third is also based on the submarket area, but uses Cluttons data on property prices.

Key Findings

The first set of models looking at movement between submarkets and TTWAs show that although the most important variable in dictating flows for the production constrained model is house price to net annual income ratio, people have an aspiration to live in more expensive areas rather than commute into London from surrounding TTWAs. The attraction constrained model shows that people living in submarkets have a desire to move to more affordable places such as Slough and Heathrow, both of which hold the highest estimates. In other words, living in London is an incentive to move out towards the TTWAs.

Similar trends are seen in the next two sets, where submarkets such as Hammersmith are an attractive destination for those looking to enter the London housing market. However, those already residing in these submarkets will be looking to move into a more attractive, and often higher priced submarket such as Holland Park. These results link in with existing literature on the property desirability which suggests such trends, where an individual's desirability changes based on their changing financial position.

The results suggest the presence of a property ladder within London whereby South Eastern and South Westerns submarkets are more affordable and therefore are more attractive as destinations for those who move from TTWAs or cheaper submarkets. Prime Central London appears to be the most favoured origin, with residents less likely to move away despite ever increasing property prices.

Value of the Research

The study allowed for a new geography to be created based on Cluttons property price data per submarket. From this data visualisation could be carried out to provide an insight into the net flows of people between submarkets and then using spatial interaction models to find variables which best explain these flows. Furthermore, modelling origin emissiveness and destination attractiveness valuable trends in the Submarkets gives evidence of a clear housing ladder effect within the capital, which is becoming increasingly difficult to move up due to high property price to income ratios

Consumer Data Research Centre

An ESRC Data Investment

Parsons bake the good stuff

Where next for Parsons Bakery? A preliminary quantitative location assessment exploring the link between demographics, competition and other variables.

William Laing[1], Richard Harris[1] and Tom Curling[2]
[1]University of Bristol, [2]Parsons Bakery

Project Background

Parsons Bakery competes in a dynamic market and each year it needs to adapt to new trends, deal with new threats and connect with the local community to provide the best chance of being successful. Store location assessment has become a critical element of retail strategy. Location and finding the right community is key to company profit. Parsons Bakery, a Bristol based family run business of 46 bakeries plans to expand its network and needed better spatial intelligence in order to make the best decisions.

The aim of the research was to determine the optimal sites for future expansion using location assessment analysis. The adoption of an optimal location strategy has the potential to provide retailers with a competitive advantage over their rivals. The research itself, focused on South West England and South East Wales due to the coverage of Parsons' store network.

Data and Methods

The initial part of developing a quantitative location assessment is determining what a company requires of its locations. Firstly, Geographic Information Systems (GIS) were used to identify the geodemographic groups geographically associated with Parsons' best performing bakeries. This information is used to target new locations and produced 53 unique Lower Super Output Areas (LSOAs).

A catchment-cum-analogue approach was then developed in order to systematically identify the factors that are associated with high performance bakeries. Two catchment areas were created for each existing bakery. Using Parsons' 2016 financial performance figures from existing bakeries combined with catchment area variables calculated from census statistics, six performance-forecast models were then designed. Overall eight hotspots matching the identified geodemographic subgroup were revealed for further performance testing.

Key Findings

Using six performance models a number of postcodes within Gloucester were revealed as the areas where a high profitability and customer number could be expected. In this setting, a bakery based in Gloucester is calculated to produce about 35% more sales than the sales expected from the other locations within the study area investigated by performance modelling. Therefore, Gloucester was recommended as a prospective location for further investigation into the unique requirements of Parsons Bakery.

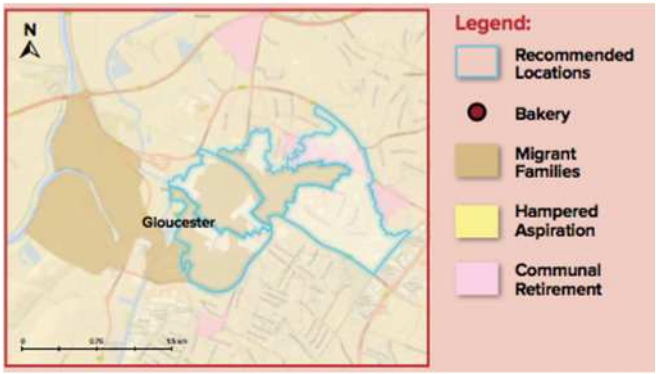

Figure 1. The geodemographic hotspots of LSOAs around Gloucester

Value of the Research

Parsons, like many other small and medium sized businesses, has not historically been of sufficient size to have the technical resources and capital investment required to carry out specialist location planning research. A comprehensive location research strategy requires many other factors to be taken into consideration.

Quantitative analysis provides a useful starting point. The analysis conducted in this research is a specific, measurable and reproducible approach to planning, which is affordable and reduces subjectivity. The catchment-cum-analogue models designed in this research avoided analytical bias as the independent variables are given uniform importance. Whilst their functionality was limited by the static nature of census data, the models are designed so that they can be updated when new data becomes available.

Other retail companies that currently do not conduct their own quantitative location assessment are recommended to do so. Whilst other retailers may follow a similar methodological process, the importance of corresponding retailer specific location requirements to their overall corporate strategy and corporate goals cannot be overemphasised.

Consumer Data Research Centre

An ESRC Data Investment

SHOP DIRECT

Propensity modelling – moving beyond standard regression models & explorations into online engagement

Mary E. Lennon[1], Guanpeng Dong[1] and Tony Birch[2]
[1]University of Liverpool, [2]Shop Direct

Project Background

Online retailers (e-tailers) have not only cast aside brick-and-mortar locations but also established new methodologies for customer focused initiatives. In their place they are exploring Big Data and associated statistical techniques and machine learning algorithms, to understand their customer's engagement. Broadly, engagement is a customer's spend (transactional) and emotional (non-transactional) commitment to a brand that can be quantified for customer acquisition and retention efforts. However, engagement tends to be ethereal and is difficult to define and leverage. This dissertation has consequentially situated itself between theoretically defining and using engagement for statistically driven customer focused initiatives. To achieve these goals, this study had two aims:

1. Explore random forest, a machine learning alternative to traditional regression framework, for propensity modelling in the context of online engagement.

2. Improve current theoretical models on engagement through the addition of individual contextual information. Specifically, through the inclusion of demographic and geo-demographic variables.

Data and Methods

This study primarily used data from Shop Direct (SD), a prominent UK e-tailer, specifically, data from their current engagement model. Each observation was an active customer account between 1st January – 1st April 2017, for their largest brand, totaling 2.2 million observations. This study also considered the age and gender of the customer's from SD's customer account file. The Internet User Classification (IUC) and Retail Attractiveness Scores supplemented SD's internal data to test the effects of geographic contexts, such as proximity to retail centres, on customer engagement. The data were included in a derivation of the random forest algorithm entitled 'Bag of Little Bootstraps' (BLB). BLB was coded in R and employed to tackle the memory and computational issues that arose when analysing SD's Big Data. The technique saves time and memory by running optimised bootstrapping across parallel cores on a computer while producing results that are as robust and reliable as the traditional random forest algorithm. To assure robustness, the results were cross validated.

Key Findings

This study had three key findings. First, BLB random forest offered notable improvements to the logistic regression baseline, with an average increase in performance of 1%. Second, the results were robust. The model error rates had a fluctuation of +/- 0.15% between each validation test. Third, the top performing model was that which included all available demographic and geo-demographic variables indicating that individual context matters. Beyond model performance, logistic regression's coefficients and random forest's variable importance shed light on the relationship of the variables to engagement. Key among these are that proximity to retail centres affects a customer's propensity to engage with an e-tailer, such as SD. Further, there is complicated relationships between IUC and the SD data. Customers in all IUC groups except for the most unlikely, elderly and unconnected, had a negative relationship with customer engagement. The results for IUC may be muddled due to the varying nature of the relationships caught by both the classification system and engagement model. For effective use of IUC additional exploration of these relationships is necessary. Regardless of the nature of the relationship, IUC and primary retail catchments were important contributions to the top performing model.

Logistic Regression v. Random Forest

Variables	*Model No.*	*Training*	*Testing*
SD Variables	M_1	0.2284	0.2273
SD Variables + Demographics	M_2	0.2281	0.2273
SD Variables + IUC	M_3	0.2289	0.2278
SD Variables + Retail Catchments	M_4	0.2289	0.2276
SD Variables + Demographics + IUC + Retail Catchments	M_5	0.2282	0.2270

Table 1. Comparing Model Performance

Value of the Research

Online engagement is critical to many e-tailers. Despite this, theory on how to measure it is limited and non-theoretical examples of its implementation are currently unknown. This project offers a starting point for crucial discussions into defining and predicting engagement. Further, it has proven the potential of machine learning techniques as well as the benefits of demographic and geodemographic contextual information to engagement models. They provide notable improvements in model performance and increase insights into consumer behaviours.

Consumer Data Research Centre

An ESRC Data Investment

movement strategies

An exploration of mobile data: towards proximity based passenger sensing on public transport

Christian Tonge[1], Dani Arribas-Bel[1] and Daniel Stockdale[2]
[1]University of Liverpool, [2]Movement Strategies

Project Background

Urban public transport systems have traditionally been planned with limited understanding of their customers travel patterns. Ticket gates and fareboxes yield disaggregate counts of passenger activity on specific stations, vehicles and individual passengers. However, greater information about origins, destinations and interchanges of individual passengers are typically acquired through much smaller samples of costly and therefore infrequent passenger surveys. Substantial research has been conducted on more intelligent systems for passenger monitoring, although little work has sought to capitalise on the wireless mobile sensors that transport customers carry. As such this study aims to evaluate the potential that a large geolocated mobile device data set has for gaining a better understanding of urban transit users.

Data and Methods

A methodology is developed for proximity based passenger sensing on the London bus network using location data collected from the Global Positioning System, and to a lesser extent Bluetooth. As the primary purpose of the data is for providing hyper-contextual adverts and not to estimate public transport patronage, a significant amount of pre-processing is required on the raw data.

The raw data is taken from 5 consecutive weekdays from Monday 3rd to Friday 7th July 2017, capturing a total of 66,179,367 data events composed of 393,360 distinct users. Initial investigation rendered Bluetooth and speed less useful determinants of ridership despite capturing daily means of 0.01% and 21.4% of total events as bus users respectively. Automated vehicle location data is used to enrich the GPS data, which is used to show 13,144 unique bus journeys on 3 London bus routes with origins and destinations for each.

Key Findings

This study has offered the data filtering and methodologies required to understand the boarding and alighting points using passenger proximity on the TfL network, offered at aggregations higher than previously permitted using traditional automated data collection systems alone. Pre-processing techniques have highlighted the issues of horizontal accuracy and users with multiple devices within the data.

100% of journeys using the combined data have origins and destinations, a figure higher than any previous GPS study in the literature. The busiest points of boarding and alighting are predominantly at stations which infers polymodal travel within the network. The distance travelled by users to the Oxford Street region can be used to indicate that the area attracts users from over 15km away, highlighting high levels of retail centre attractiveness. The successful cleaning processes and methodology developed in this study will allow more conclusive results to be drawn on OD matrices and other important insights. Such insights include temporal variability, daily and seasonal ridership patterns, whole network comparisons and data penetration levels. However, due to the novelty of the methodology and the spatiotemporal resolution at user level it is difficult to justify the success of the study against another form of data.

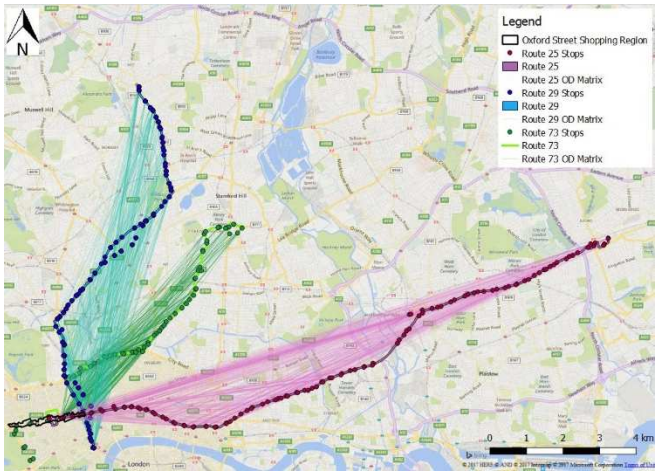

Figure 1. Origin Destination Matrix of Trajectories at Bus Stop Level

Value of the Research

As the first attempt at defining a methodology for this data, the time constraints prevent more interesting and certain insights from being gained than OD ridership levels. While the novelty of this methodology makes it difficult to compare to other data sources, it has been shown to infer aspects of public transport activity that are not immediately available from automated data collection systems. Implementing this methodology over a whole network for a longer period would help to provide more useful insights for transport planners and marketers alike. Despite this data being London centric, the pervasiveness and ubiquity of GPS and the relative ease of implementing SDK applications, makes this method reproducible wherever vehicle location and passenger GPS data are available.

Consumer Data Research Centre

An ESRC Data Investment

Knight Frank

Identifying Socio-Spatial Inequalities in Student Housing in London

Terje Trasberg[1], Guy Lansley[1] and Thomas Murphy[2]
[1]University College London, [2]Knight Frank

Project Background

This project aimed to analyse existing spatial patterns in student accommodation and, based on student's preferences, suggest the most suitable areas for students in London. Previous research on students' residential decision making has shown that students often cluster into areas around the university, ignoring better options available on the market. So, this study proposed a computing multi-criteria attractiveness index in order to select the most suitable areas for student housing. The results of the study should support students in making more informed residential decisions and serve as guideline for real estate consultancies in targeting areas for purpose built student accommodation.

Data and Methods

Existing spatial patterns and criteria for the index were identified based on Student Accommodation Survey results provided by University of London Housing Department. The dataset included 4,699 survey replies from students who rent from the private sector (shared flats, studio, and landlord's house). The criteria for the attractiveness index were weighted using pairwise comparison after analysing the spatial inequalities of student preferences across London.

An open-source tool MCDA4ArcMap, designed for spatial multi criteria analysis in GIS software, was applied to assign an attractiveness score to all the areas in the dataset. The areas with the highest score were seen as the most suitable locations.

Key Findings

The attractiveness index was built upon the following criterion: proximity to university, cost of the area, safety and attractiveness of the area (deprivation measure). Criteria were weighted based on the spatial analysis of student preferences. It was observed that there are significant spatial inequalities in factors, which students hold important about the location of their accommodation.

To accommodate the spatial differences, the attractiveness index was first calculated using the examples of University College London in central London and St George's, University of London, in south London. Those universities were selected due to their high volume of respondents that were geographically clustered near their campuses. For the UCL sample, proximity to university was ranked by students to be more important than cost. So, most suitable areas were expected to be close to the university. Surprisingly, this was not the case and instead Paddington and Maida Vale, which are around 4km away from UCL, received the highest attractiveness score. Those areas have lower weekly rents than areas with a high concentration of students (for example Camden). Similarly, areas in the immediate surroundings of St George's University, where student concentration is high, received lower suitability scores compared to the other surrounding areas due to lower area attractiveness and higher crime rate.

The study concludes that to receive better living conditions students should consider areas with low student concentrations and not the immediate surroundings of the university. Most suitable areas are around 3-6 km away from the university.

Figure 1. Percent of students that ranked each of the factors as important

Value of the Research

This paper introduced a methodology to define the most suitable areas for student accommodation based on student surveys on the example of 2 universities. Similar analysis could be conducted for all London universities to observe, if any areas stood out as generally the most suitable. This would be valuable input for real estate consultancies which are planning to establish purpose built student accommodation in London and are looking to find the best suitable locations and learn more about student preferences and expectations for the accommodation.

Consumer Data Research Centre

An ESRC Data Investment

City of Westminster

The impact of the night tube on Westminster's Night Economy

Paula White[1], Steven Gray[1] and Dominic Baker[2]
[1]University College London, [2]City of Westminster Council

Project Background

The City of Westminster has the largest evening and night time economy (ENTE) in the UK; it is larger than the ENTE's of Edinburgh, Birmingham and Manchester combined. The highest concentration of Westminster night time provisions is found in the West End, an area comprising of Soho, Leicester Square, Piccadilly Circus and Covent Garden – it is a major tourist destination within the UK and internationally. In addition, with the introduction of Night Tube in August 2016 to December 2016 there are now 8 stations operating in and around the vicinity of the West End for 24 hours Friday and Saturday nights. There is widespread interest from businesses, residents and policy makers to understand the ENTE in the London borough of Westminster and the West End in particular

Data and Methods

First, work was undertaken to understand the problem domain, next data were collected, cleaned and reformatted, after which exploratory data analysis were undertaken to give a solid foundation for detailed discovery steps. The analysis for this study covered three keys areas; footfall and incidents, commercial premises applications and social media analysis. These were undertaken using a number of techniques from GIS, quantitative methods, data mining and social media analytics.

Service data and commercial premises application data has been provided by Westminster City Council. The Consumer Data Research Centre provided footfall data. This is from the Smart Street Sensor project which is a partnership project between UCL and Local Data Company (LDC). A variety of social media was sourced via IBM Watson Analytics, a data exploration and visualisation tool.

Key Findings

Overall this study, has shown that the West End is still undoubtedly a massive, highly dense, hub of activity.

There were four geographical areas of analysis in the following areas; The West End, Victoria and surrounding area, Marble Arch and surrounds and Bayswater. Comparisons were made from a reference footfall before the Night Tube and the various changes through and after its deployment.

There was a significant increase in footfall on Friday and Saturdays in parts of the West End by 202% and 219% respectively. By contrast, Thursdays only have a 164% increase. In Bayswater, Friday and Saturdays saw an 80% increase in comparison to 50% on Thursdays. Note that these are proxy values only so are indicative of areas of change rather than showing complete accuracy.

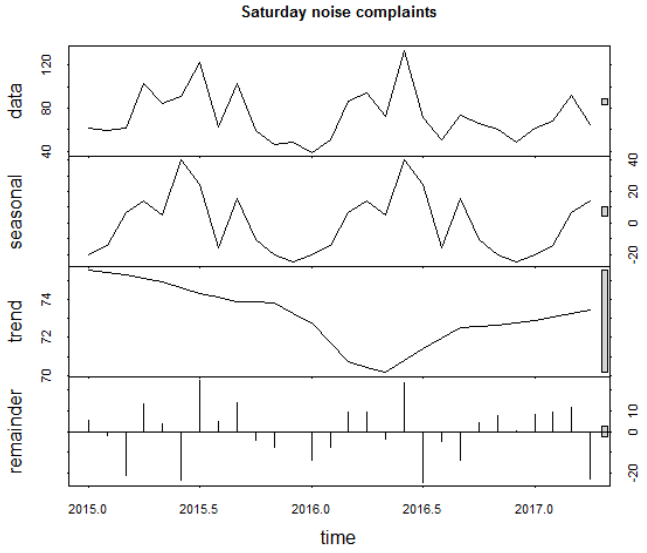

Figure 1. Saturday Night Noise Complaints

As can be seen from Figure 1, it was also observed that street noise complaints are on the increase. Analysis showed that many of these complaints have a higher relative density in the West End but there is also a relative increase in the Victoria area.

Since the Night Tube deployment, the data indicates some shifts in the locations of applications for commercial premises of hotels and pubs which are now more dispersed across the borough. There was also a significant growth in Casino applications, as well a short-term trend toward applications for smaller capacity pubs.

London's nightlife seems to have a positive sentiment on social media as does the Night Tube from a worldwide audience which reflect London's global status as a vibrant and exciting world class city.

Value of the Research

This dissertation found some answers to important questions about the direction the ENTE is taking in Westminster and starts to allow for better conversations and thinking about the city ENTE. It also provides some thoughts on future research data and approaches to further the understanding of this fascinating topic in greater detail.

Consumer Data Research Centre

An ESRC Data Investment

M&S EST. 1884

Can harnessing the demographic characteristics of store catchments improve the planning of promotions and pricing strategies?

Daniel Winder[1], Nik Lomax[1] and Muhammad Adnan[2]
[1]University of Leeds, [2]Marks and Spencer

Project Background

The overarching aim of the study was to examine the influence of demographics on price sensitivity and deal proneness in the UK grocery retail industry. If Marks and Spencer (M&S) fail to adapt to changing market pressures and consumer behaviour this could lead to a decrease in market share. The company now places a strategic focus on its grocery element and using their 'Sparks' loyalty card data to increase the efficiency and relevancy of promotions. It is within this strategic direction that the study resided: by firstly attempting to show how demographics may affect the uptake of promotions and could therefore be used to improve promotional and pricing efficiency; and secondly to evidence the valuable insight that can be obtained from loyalty card data.

Data and Methods

The research was enriched by commercial data that is not openly available to academia. POS data for ten grocery products over the course of a year were provided from four stores across England. Expenditure on 'Sparks' loyalty cards at each store, attributed to a specific output area (aggregated from postcodes) was used for the estimation of the store catchment areas with a GIS. The estimated catchment areas were then joined with 2011 Census data to obtain the key demographic characteristics of the stores' catchments that have been argued to be effective determinants of deal proneness and price sensitivity.

The effect of price on the sales of four products at the different stores was explored by examining sales data alongside price histories. Linear regression models were applied for two purposes: firstly, for hypothesis testing to establish whether the impact of price was seemingly influential on sales; secondly, to explore the extent of the price-effect, if present, by estimating the price elasticity of demand from the outputs of the regression models. The results were then collated with the demographic characteristics of the stores to infer potential demographic influence on price sensitivity and deal proneness.

Key Findings

The use of loyalty card data for catchment area estimation was shown to be an applicable method but it did overestimate areas for stores where customers were likely to exhibit complex customer patronage behaviour. Most poignantly, the extent of the calculated catchment area of a convenience store located in a travel hub evidenced the intricacies associated with estimating a catchment area for a store of this type.

The findings displayed distinct differences in how price impacted sales at stores. Price and deals were argued to affect customers of the travel hub convenience store differently to the other more traditional outlets. In tandem with the differing price-effects, the estimated catchment area demographics also varied (see Figure 1). Customers at Store A were argued to be the most price sensitive and thus deal prone. The comparatively high proportion of households where 5 or more people reside led to the inference that households of this size were potentially more deal prone. In addition, the findings were argued to contradict the common belief that elderly people are more price sensitive due to lower opportunity costs and that more educated people are less price sensitive because of higher opportunity costs.

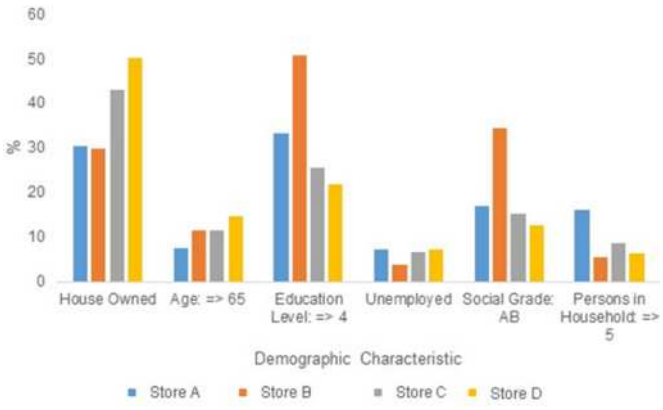

Figure 1. Estimated composition of store catchment areas with key demographic variables from the 2011 Census

Value of the Research

The academic and commercial benefit for M&S resides firmly in the explicit explanation of the catchment area estimation technique which is brought forward for the first time in the existing literature. Furthermore, the insight asserts that strategically M&S, and other retailers, should not offer homogenous pricing and promotions across their network; these should be customised for increased effectiveness. The inferences in the findings show that catchment area demographics should be used to support the planning of these.

Consumer Data Research Centre

An ESRC Data Investment

In collaboration with a national telecommunications provider

An application of statistical and spatial analysis techniques to engineer callout data from a UK telecommunications company

Patrick Wood[1], Ruth Hamilton[1] and Andy Simpson[2]
[1]University of Sheffield, [2]CDRC industry expert

Project Background

This research aims to understand patterns in engineer callouts for a telecommunications & media company. The research analyses data concerned with the frequency and nature of engineer callouts during the year 2016, within Bristol, Bath, Bradford and Newcastle. It combines engineer callout data with geodemographic data in order to assess whether different social groups are more or less likely to experience issues with internet services. The research also presents an original enquiry into the effect of weather temperature on internet services as measured by engineer callouts. Analysis focuses on customer-caused spliced cable callouts testing a hypothesis from an engineer interviewed during a company 'ride-out' day.

Data and Methods

To investigate geodemographic trends, a number of statistical tests were conducted. Chi-square test results indicate whether the distribution of engineer callouts by geodemographic groups are random and a binary logistic regression produces odds ratios quantifying how much more or less likely ACORN and Internet User Classification (IUC) groups are to experience engineer callouts than selected reference groups. Bivariate linear regression and Pearson's R tests detail the strength and direction of trends between engineer callout data and explanatory variables such as 'final mile' distance and weather temperature.

Mapping and spatial analysis methods were carried out using ESRI ArcGIS software. Dasymetric mapping techniques visualise engineer callout data by shading building footprints, instead of administrative area polygons, to better reflect the geography of urban areas, customers and population density. Global and local spatial autocorrelation statistics analyse whether the data showed statistically significant geographic clustering of percentage change in customer-caused spliced cables between cold and hot days at the postcode sector level.

Key Findings

'Chargeable visit' engineer callouts are coded when nobody is present at the property upon the engineer's arrival or when the problem is caused by faulty customer equipment. These callouts are more likely amongst younger, more e-engaged neighbourhoods. 'Clarification of service' callouts, indicating a customer has falsely reported a fault due to a misunderstanding of their service, are more likely amongst older-age groups. However, socio-economics amongst older age groups is also influential, with higher socio-economic groups having proportionally less callouts.

This research identified a positive and statistically significant relationship between the average 'final mile' distance and engineer callouts, although the regression models indicated that it only explains roughly 4% of the variance. Mean daily weather temperature also exhibited a similar influence on television-related and internet-related callouts. However, weather temperature was found to explain 22% of the variation in customer-caused spliced cables. Studying the percentage change between cold and hot days across space revealed that built-up semi-periphery areas have the greatest increases in callouts on hot days.

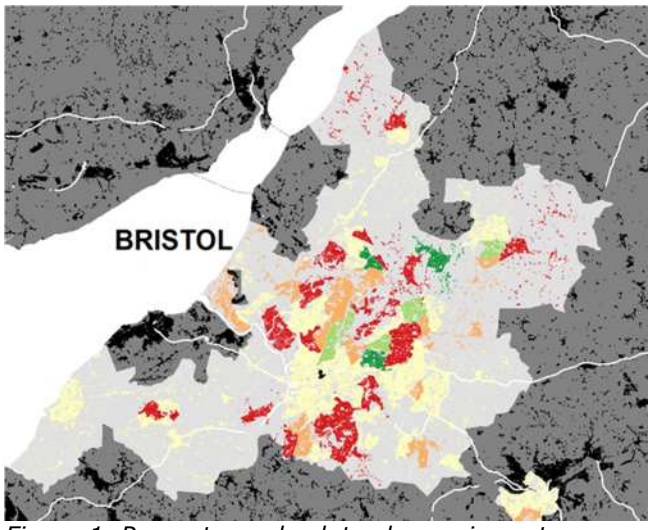

Figure 1. Percentage absolute change in customer-caused spliced cable callouts in Bristol (Red = increase, Green = decrease)

Value of the Research

While the population is becoming more E-engaged, this research has demonstrated that there are still variances in miscommunications and understandings of telecommunications technology between different age groups and also different socio-economic age groups. This has subsequently led to geographic trends in engineer callouts within cities. This insight can therefore support telecommunications providers and local governments by enabling them to communicate more effectively with consumers with particular geodemographic characteristics to improve service delivery and reduce barriers to internet access.