Consumer Data Research Centre

An ESRC Data Investment

CONSUMER DATA RESEARCH CENTRE

Masters Research Dissertation Programme

Case Studies 2015

Edited by Guy Lansley

7 October 2015

Identifying fuel and poverty characteristics through e.on consumer records and geo-demographic segmentation data

Alexander Bland[1], Andy Newing[1], Ben Keown[2]
[1]University of Leeds, [2]E.ON UK plc

Project Background

Fuel poverty is known to be a distinct social problem, which can occur across a wide array of household demographics. The labelling of households through the associated definition has been under recent review and now encompasses two main elements of fuel poverty: energy cost and stability of household income. The Hills review definition (2012) provided a much-needed update, to more accurately identify the fuel poor in the UK. E.ON wanted to utilise this review in effort to align their obligations set by the government to provide support to disadvantaged customers. As a distinct social problem the occurrence of fuel poverty and its consequential impacts has been identified by many different studies, using a range of data sets. This dissertation presents a unique identification process, combining household resolution data with commercial grade geo-demographic customer segmentation classifications (CAMEO), not previously used in the identification of fuel poverty.

Data Methods

3.9 million E.ON customer data records were received with CAMEO classifications attached. Constraints associated with selecting a usable customer samples reduced the data set to 2.1 million records. Record eligibility was selected by the following criteria; reasonable consumption values and being a dual fuel customer reduced the data set significantly. Carrying out the filtering by use of the IBM software: SPSS, enabled the justified selection to be made, allowing for subsequent fuel poverty analysis.

Key Findings

From the 2.1 million dual fuel customers, the fuel poor were identified using the Hills definition (2012). This presented 211,441 households who were regarded as being below the fuel poverty threshold; a population representing 9.73% of the E.ON customer sample. This record has provided the high-resolution data sample displaying the characteristics of the fuel poor population as a whole. Figure 1 shows the distribution of the fuel poor population across England and Wales, using density mapping, areas representing high levels of fuel poverty prevalence include Liverpool, Manchester and the West Midlands.

Proportionally the fuel poor record possesses larger quantities of old terraced housing, whilst more recently built residences, particularly detached and semi-detached are less likely to house the fuel poor.

The associated depth of fuel poverty has been examined, presenting evidence that the composition of the fuel poor population is diverse within itself. Most notably the increase in the proportion of council tax-band A households, the highest disparity being over 8,000 households.

To validate the selection of fuel poor made statistical comparison to a previous study was made. A Pearson's correlation co-efficient of 0.584 ($p<0.05$) with an independent study by the Centre for Sustainable Energy, on the same geographic resolution was found. Predictive modelling presented no statistically significant correlations between variables in predicting fuel poverty, reasserting the complexity of the social issue whereby variables can negate one another.

Attempts to generate a predictive model from the most influential characteristics in determining fuel poverty were conducted.

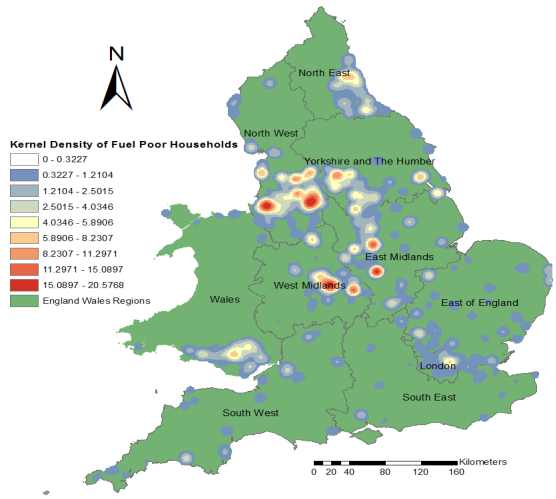


Figure 1: Kernel density map of fuel poor customers across England and Wales

Value of the Research

Results presented in this study can be utilised by E.ON through production of customer specific marketing strategies aiming to alleviate fuel poverty amongst its customer base. Information produced will also enhance the understanding of the complexities of fuel poverty by displaying the distribution of the fuel poor and highlighting the most prevalent characteristics, supporting previously published literature.

Can we identify vulnerable energy customers from smart meter data?

Anastasia Ushakova[1], Slava Mikhaylov[1], Andy Simpson[2]
[1]University College London, [2]British Gas

Project Background

For companies like British Gas issues of supporting vulnerable consumers and fuel poverty together with carbon emission reductions have been important over recent years, especially with the introduction of Energy Company Obligation (ECO) and the Green Deal by the UK government. With the availability of streaming data from smart meters we are able to develop simple and reliable methods of identifying vulnerable energy customers and as a result develop targeted policy interventions. The research question of this study is thus how can vulnerable customers be identified from natural gas consumption data.

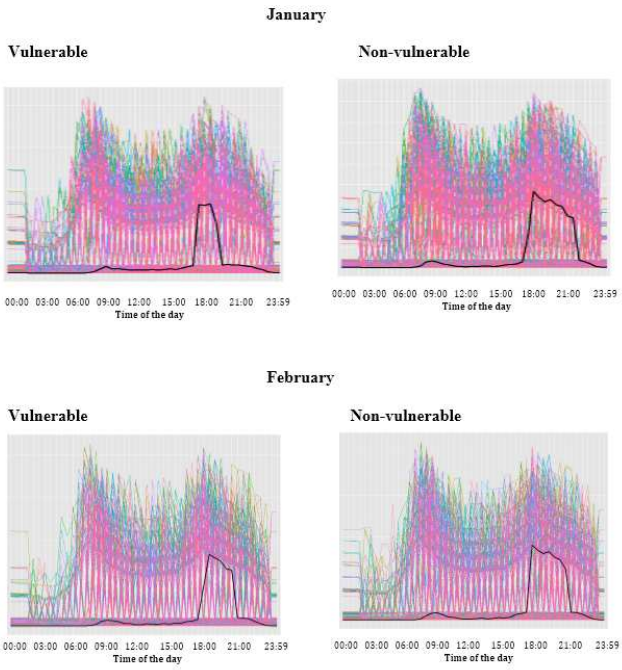
Data & Methods

The study was based on the sample of 2,000 smart meters from Scottish region that recorded half-hourly consumption across the year 2014. A binary vulnerability flag was created to indicate if one or more support measures for vulnerability were applied to the customer, which indicated that around twenty five percent of total sample had a vulnerability flag present. The methodology was based on supervised machine learning techniques used for targeted classification and prediction. Total yearly consumption was analysed for the whole sample together with randomly selected customers from different life-stage groups. To see overall grouping of the data and differences in consumption patterns for January and February the simplest method such as k-means clustering was applied to data corresponding to typical Tuesday-Thursdays mid-week profile. K-means procedure was further complemented with hierarchical and Gaussian mixture models analysis. Lastly, as a major part of the research neural network model was trained to see if we can predict vulnerability flag using only data on half-hourly consumption.

Key Findings

Cluster classification presented evident difference in vulnerable consumption in comparison with those individuals who have no vulnerability flag. Slight differences in patterns was also evident between the two months of consumption. In fact if we look at overall median consumption patterns these differences are hardly observed. This has indicated importance of using more profound methodology to describe energy usage for different types of customers.

Neural Network model prediction was designed and trained for both January and February to recognise associations with vulnerability on new unseen data. Results have shown that model has prediction power yet it may depend on number of hidden nodes that we selected. Generalizing if predicted value is below 0.5 as zero and above as 1 gave a prediction accuracy of 76% on average regardless of the number of nodes. Further research may consider inclusion of variables on climate and geographic characteristics that may well contribute to predication power of the model.



Average daily energy profiles of vulnerable and non-vulnerable households in January and February

Value of the Research

The extent to which analysis of gas consumption from smart meters was analysed in current academic literature is quite limited and for British Gas, it is also the first time that gas consumption data has been attempted for predictive analysis. Thus, this paper opens up a clear possibility to use machine learning techniques not just for operational research but also for public policy research that aims at informing policy interventions in the energy sector.

Consumer Data Research Centre

An ESRC Data Investment

The co-operative food

An investigation into the effects of a relaxation of the current Sunday Trading Legislation on the Co-operative's convenience stores

Brendon Edwards[1], Graham Clarke[1], Mark Coleman[2]
[1]University of Leeds, [2]The Co-operative

Project Background

The 2015 July Budget proposed a devolution of control of Sunday trading hours to city mayors and local authorities. The effects the changes could have on the Co-operative convenience stores were unknown. This research sought to forecast the financial impact the changes would have on their convenience stores and give advice on how to best mitigate losses.

Data and Methods

The approach to forecasting employed two different models; a Geospatial Model and a Multi-linear regression Model. Scotland has no restrictions on Sunday trading so was used as a control variable in both models.

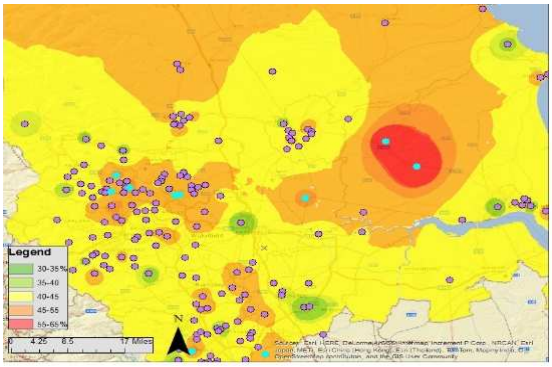
Geospatial Model: This model employed various buffers between 500m and 10km to examine the impact of supermarkets of varying size on the 'Post 4pm Sunday turnover' in both Scotland and Yorkshire. The model forecasted turnover for Yorkshire, based on the assumption the stores would take the same as equivalent stores in Scotland that have the same proximity to supermarkets of varying sizes, if there was a relaxation in the laws. The model was then applied to the rest of the UK.

Multi-linear regression Model: This model provided a store by store forecast using a combination of three predictive variables 'Beers Wines and Spirits sales', '10-11am turnover' and 'Grocery sales' to provide an accurate model with an adjusted R-squared value of .856 to predict a turnover loss to the convenience sector of -£31.78M. The results gave a store by store forecast. The forecasted stores to suffer the largest loss of takings are the stores that currently take a higher percentage of Sunday takings post 4pm.

Key findings

Geospatial Model: The results suggested that the closer a given Co-operative convenience store is to a supermarket competitor (particularly supermarkets sized 10,000+ square feet) the higher the takings were during the times the supermarket was closed. This is due to customers who would have used the supermarket then shopping at the Co-operative. The effect is different over different sized buffers and against different sized competitors. The buffer and competition size that represented the largest losses was a 10km buffer with a 10,000+ square foot competitor intersect. National turnover losses to the Co-operatives' convenience stores were forecasted as £33.25M. Following the geospatial analysis it was decided a more in depth statistical review was needed in order to make forecasts for individual stores.

Multi-linear regression Model: This model provided accurate store by store forecasts. The national forecasts were very similar to the first model further validating both forecasts. The forecasted annual losses for the Co-operatives convenience stores were -£31.78M, however when the model was applied to the Co-operative stores over 280 square meters there were forecasted gains in turnover of £90.64M. This forecast is likely to be an overestimation due to the model not being purpose built for forecasting for stores over 280 square meters. Although possibly inaccurate the predicted gains for the stores over 280 square meter highlights one of a number of ways the impact of the changes can be mitigated against.



"Post 4pm" Inverse Distance Weighting of Yorkshire

Value of the Research

This research effectively forecast the potential losses to the Co-operative convenience stores of £31-£33M. The possible gains have also been demonstrated with a forecast for stores over the size of 280 square meters. The method to mitigate against such losses were fully examined to form the advice given to the Co-operative as a result of the potential changes in the Sunday Trading Laws. Whilst it appears the new laws will not be brought in nationally rather there will be a devolution, the multi-linear regression model enabled a store by store forecast that will enable the Co-operative to assess the impact on the individual stores where the laws are brought in to place.

Consumer Data Research Centre

An ESRC Data Investment

easyJet

Identifying the main drivers of customer satisfaction and dissatisfaction by mining customer verbatim feedback

Clemens Zauchner[1], Thierry Chaussalet[1], Mario Streng[2]
[1]University of Westminster, [2]easyJet

Project Background

As competition is getting fiercer, it is vital for companies to improve their products and services constantly to satisfy their customers. Only if a company understands and acts upon drivers of satisfaction and dissatisfaction, can it survive in the competitive markets. EasyJet conducts customer satisfaction surveys, where customers rate and comment on various touch points of their customer experience. So far, only the rating has been considered. If customers comment on their experience, more valuable information can be extracted, as text allows more valuable insights than ratings alone. However, the complexity of analysing the data is higher with text data, especially when the amount of information is not manageable by a human.

Data and Methods

The data used in the analysis was collected over a 14 month period from April 2014 to May 2015 and was provided by easyJet. It contains customer and flight data and ratings and comments on touch points during the customer journey. For the analysis of the text, two approaches are used: topic models and aspect based sentiment analysis. The topic model algorithm used is *Latent Dirichlet Allocation* (LDA), which is an unsupervised algorithm to identify common topics from a set of documents, where it is assumed that a document is formed of many topics and each topic is a distribution over words. Aspect based sentiment analysis describes the task of identifying frequent aspects from the text and classifying the related sentiment. It is therefore more granular than sentiment analysis on document level, assuming that the document bears only one sentiment.

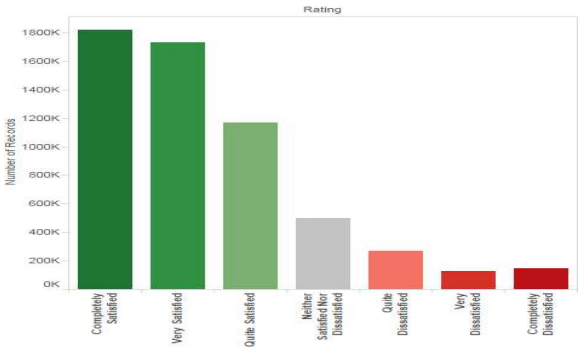
Key findings

The topic model of comments after a positive rating and the topic model of comments after a negative rating were considerably different. In the positive comments, the topics were the *usability* and *ease of use* of the website, also efficiency of processes, as well as the staff or customer service. In negative comments, customers complained about delays, queues or inefficiency of processes. Further, the handling of cabin baggage seemed to cause dissatisfaction and so did the information/communication on delays. Service encounters seem to be one of the most important drivers for customer satisfaction. A friendly and helpful crew drives customer satisfaction, whereas a crew that is rude, does not smile or does not help drives dissatisfaction. Announcements of the cabin crew have room for improvement. In case of delays or changes, customers are not satisfied with the information they get and the style of communication of the crew. The cabin crew plays an important role in safety and security.

The cabin bag guarantee policy seems to be one of the main drivers of customer dissatisfaction, as it is discussed mostly in negative sentiment context.

Customers appreciate processes that run smoothly and efficiently and dislike disruptions, delays and waiting times. The longer customers have to wait, the less satisfied they are. The facilities at the boarding gate seem to be very important. Customers dislike it if they are in the wrong terminal or have to go long ways.

Many customers state that they are neither interested in food and drink, nor did they pay attention to the menu or purchase anything. This means that cross-selling potential is not realised. The selection and the quality of the food are discussed controversially. Customer segmentation might give more insights.



Frequency of satisfaction ratings of all rated touchpoints

Value of the Research

The analysis has found main drivers of customer satisfaction as well as dissatisfaction that could be used to improve service quality, which could lead to customers purchasing again or recommending the service. The thesis is at the interface of academic research and application in business; it describes and discusses both, managerial implications and the applicability of topic models and sentiment analysis in industrial context. The way of presenting and visualising the results was carefully chosen to allow managers to understand the findings easily.

Consumer Data Research Centre

An ESRC Data Investment

Relations between structure and performance of retail centres in England and Wales and demographics of their catchment areas

Karlo Lugomer[1], Guy Lansley[1], Ronald Nyakairu[2]
[1]University College London, [2]Local Data Company

Background and motivation

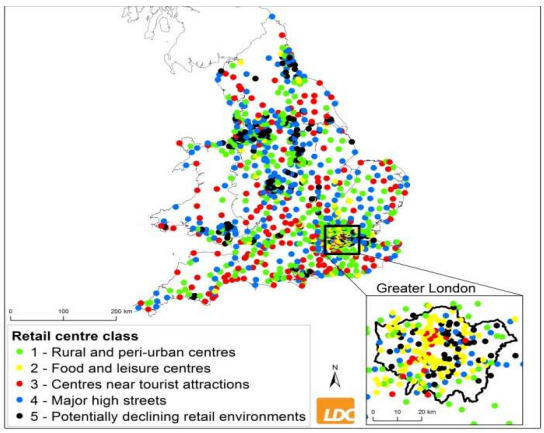
While there has been plenty of research on the population structure and how population data can be used to predict retail turnover both within the academic community and amongst retail analyses. Yet little has been done to establish how the characteristics of retail centre catchments vary across England and Wales, and how these characteristics may influence the health of retail centres. This paper explored these relationships through the estimation of retail catchment areas of 1206 retail centres in England and Wales, and then measuring the population characteristics of those residing and working within each catchment. The retail centres were classified based on the composition of stores from different retail categories so that relationships between the retail centre characteristics and local consumers could be easily identified. Finally, this research produced a list of key demographic variables which were found to link most closely to what the Local Data Company defines as a healthy retail centre.

Data and methods

Demographic data acquired from 2011 Census of population was used along with attributes of retail centres in England and Wales obtained from Local Data Company. Latter comprised areal boundaries of retail centres, percentages of 42 different store categories within each of the retail centres, corresponding vacancy rates, retail health indices and retail diversity indices. Huff's probabilities of residents and workers shopping in each retail centre were calculated using Lower Super Output Area level population data through a spatial interaction modelling procedure. A K-means clustering algorithm was employed to derive distinctive clusters of retail centres based on the percentages of different store categories present. Location quotients were calculated for each of the demographic variables to establish the extent of which demographic characteristics were represented in each class of retail centre. Finally, regression methods (bivariate, multiple, geographically weighted) were used to detect the most important demographic indicators of the retail performance.

Key findings

Five classes of retail centres were identified: rural and peri-urban retail centres, food and leisure centres, centres near tourist attractions, major high streets and potentially declining retail environments. Each class was found to have unique average demographic characteristics across their catchments. The latter was most associated with deprived neighbourhoods and many of the high streets within this category are some of the least healthy in England and Wales. Residential and workplace groups of variables were found to be important predictors of retail performance based on health indicators, especially when combined. The most important predictors were found to be percentages of: semi-routine workers, workplace population with no qualifications, workplace population with level 4 qualifications, population employed in manufacturing sector and population employed in information/finances/real estate/professional sector; together accounting for 23.8% of the variance of the retail performance. Across England and Wales socioeconomic and educational variables were found to be the most influential on retail health, while age structure, family composition and ethnicity proved to be of a minor significance. Although local variances were also identified across the country.



Classes of retail centres, England & Wales

Value of the research

The classification of 1206 retail centres in England and Wales gave a deeper insight into how local population traits are spatially associated with retail centre characteristics, and this information could be used to help retailers make better informed locational decisions. While this paper suggested that the demographics of retail centre catchment areas alone cannot fully predict retail performance, a set of the most influential variables has been detected. This data could be used evaluate the suitability of retail centres for particular stores, and also to provide information to assist the regeneration process of struggling high streets.

Topic modelling online customer reviews

Radoslaw Kowalski[1], Slava Mikhaylov[1], Tom Cusak[2]
[1]University College London, [2]Argos

Project Background

Many published studies inform how to use topic modelling to understand customer review data, but no prior studies attempted to implement a single topic model across a wide range of product categories. Meanwhile, retailers need a solution that would allow them to analyse customer review data on all of the products they sell simultaneously. Retailers' customer reviews may span hundreds of product categories and tens of thousands of products, and topic modelling for individual product categories or products would be very costly. Moreover, it would be hard to compare topic model results between product categories as each model produces a different set of topics. Therefore, a single topic model for all product categories can give retailers more insight than many topic models for individual product categories or products.

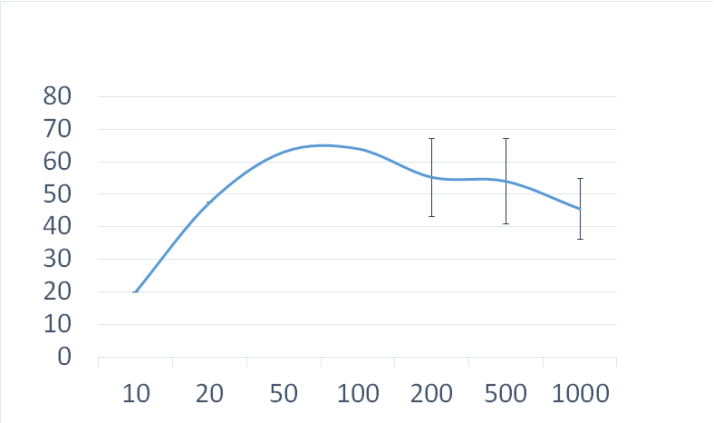
Data and Methods

Latent Dirichlet Allocation (LDA) model with Gibbs sampling was used in this study to see whether topic model results can be a reliable tool for summarising customer reviews across product categories. The study used 650,000 customer reviews and user-generated sentiment scores obtained from Argos plc. The analysis answered the research question: How to implement topic modelling on reviews of a wide range of products or services simultaneously? The tested hypotheses were:

1) A combination of select parts of speech is best for identifying distinct product features and benefits
2) There is an optimum number of topics to be found in customer reviews

Key findings

The first two hypotheses were evaluated through inspection of the top 5 words to see if the words signified identifiable product benefits and drawbacks. Both hypotheses were confirmed. Optimum variants of LDA have been identified for the choice of parts of speech, n-grams and target number of topics. The third hypothesis was not confirmed. Weighted average of sentiment scores for topics cannot effectively measure sentiments associated with each topic. Nonetheless, LDA topics representing product benefits have the same meaning across product categories because LDA takes words' context into account.



Accuracy of topics (%) depending on the number of topics generated by an LDA topic model

Value of the Research

Future topic modelling analyses need not be constrained by arbitrarily defined product categories. Topic modelling can be carried out across product categories without sacrificing the quality of modelling results. Research findings suggest that retailers can use topic modelling to obtain new insights. For example, topic model composition can be compared between products and product groups to discover possible correlations with sales as well as with customers' demographics and online behaviour patterns. Moreover, analysts can freely investigate the topic structure from customer reviews according to their own preferences (e.g. organizing reviews according to posting date as opposed to membership in a product category) without need to recalculate the topic model again.

Exploring the utility of the 2011 Work Place Statistics to help The Co-op better understand transient new store locations, worker flows and worker demographics

Thomas Berry[1], Andy Newing[1], Kirsty Branch[2]
[1]University of Leeds, [2]The Co-operative

Background and Motivation

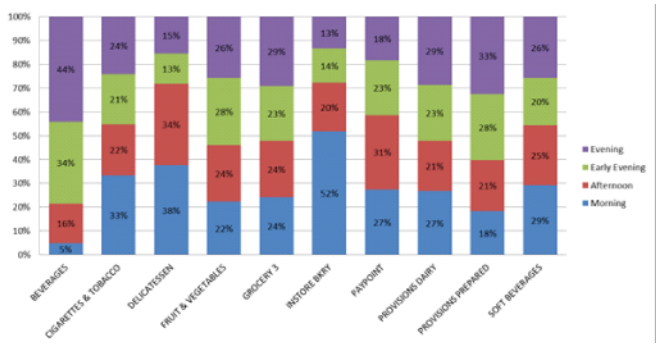
We consider the utility of census-derived Workplace Zone (WPZ) population statistics to supporting location-based decision making in the convenience grocery retail sector. Using Co-op trading data for highly transient Central London stores, we explore the link between workplace population composition and the observed temporal store trading patterns. The construction of WPZs following the 2011 census enables us to explore small-area spatial variations in highly transient workplace populations not previously captured by the residential geography used for the dissemination of population statistics. This can afford new insights into the factors driving observed consumption behaviours at Co-op stores in Central London, where trade is dominated by highly transient workplace populations.

Data and Methods

Key Data: WPZ population statistics: Freely available via the ONS website and developed using 2011 Census statistics related to self-reported workplace location. Co-op store performance data focussing on Time-of-Day purchase records for 10 product departments generating the greatest proportions of revenue across key transient stores within Inner London. Inferences about workplace population characteristics within these WPZ statistics included: *Age, sex and key socio-economic characteristics *Travel-to-Work method *Residential origin (by OA) We used these statistics to identify the characteristics of workplace populations within predetermined 500/700m buffers around five Inner London case study Co-op stores. Inner London's dominance as the central hub for workers identified it as the area requiring further analysis while the five case studies were chosen to provide a variety of worker demographic characteristics and store performance data for investigation. Flow data analysis was developed to identify inflows of commuters to workplaces within these buffers to enable an understanding of where workers live.

Key Findings

This research has resulted in considerably improved analytical detail of population composition around key Inner London stores. Findings show that workplace populations are heavily concentrated, most notably around the City of London and Canary Wharf. The majority of these people have a residential origin within Greater London and form the highest socio-economic groups. They have greater disposable incomes and will be looking to purchase more premium options. Average distances travelled by customers to the most central Inner London stores can be expected to exceed 20km. Areas such as Sevenoaks were found to be key commuter towns outside of Greater London, with highly affluent populations. Stores can expect the vast majority of sales to originate during three periods of the day: before work (8-9am): during lunchtime (12-2pm) and after work (5-7pm), with the key sales period being the purchase of lunch. Stores in areas with a lower workplace population density experience a gradual increase in trade throughout the day, until a major spike after 5pm. They experience a greater proportion of their sales in the evening in the form of more residential expenditure. The most successful transient stores are located directly next to/opposite major transport nodes, probably driven by greater footfall and visibility. They should focus on providing a time-efficient shopping experience to allow for large customer flows of which the average basket spend is relatively low. Key goods are sandwiches and snacks while incentives to encourage repeat purchase such as a discount on dinner when purchasing lunch could be adopted to increase sales.



Time of day revenue analysis of Top 10 product departments at transient Inner London Co-op stores

Value of the Research

This research is based on use the of novel population statistics which have provided the Co-op with enhanced insight into workplace population characteristics around key transient stores in Central London. This includes a detailed break-down of customer spending habits by product category within stores in the case study locations. This will enable them to make informed decisions about developing revenue prediction models for stores in highly transient workplace locations. It can also be used to support the identification of potential new store locations by providing an indication of the volume and composition of non-residential trade.

Shopping centre's turnover estimation using microsimulation: an exploratory research in Inverness

Yiqiao Huang[1], Guy Lansley[1], Chase Farmer[2] and Alex McCulloch[2]
[1]University College London [2]CACI UK

Project Background

Turnover estimation of retail outlets plays a vital role in market analysis. Traditional approaches such as spatial interaction modelling and its varieties are still commonly favoured as an approach to predict the extent of which local residents will patronise individual stores or centres. The approaches assume that the decision to spend money at a particular store is driven by the consumer's characteristics, the store characteristics and spatial distribution of stores, and these models often take advantage of aggregate population data. Retail consultancies are currently seeking new techniques that can take advantage of recent increases in computational power and rich data sets on consumers and retail environment to predict consumer interactions on an individual level. This exploratory research was conducted in the area of Inverness and its outskirts; it aimed to investigate if an agent-based technique to estimate turnover from the bottom up could be a viable alternative to traditional methods.

Data and Methods

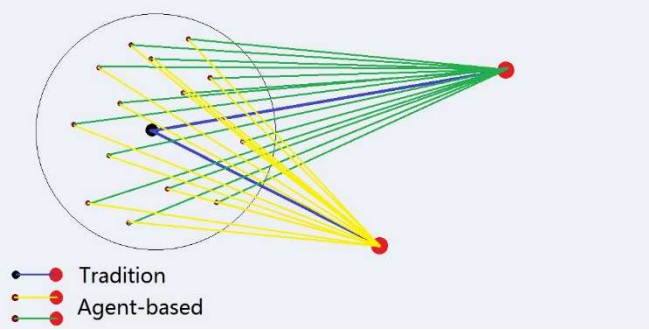
The data for this research provided by CACI included the Acorn consumer classification at the postcode level, data on 5 shopping centres in Inverness (including the number of categories of stores located, price level for each store), and also aggregate shopping behaviour data distinguished by Acorn group type. In addition, household composition by car or van availability and age by sex from 2011 census were also used to gain more of an understanding about households. The Integrated Transport Network (ITN) was used to build the road network in the study area and gain a better understanding of travel times between consumers and shopping centres. With this data combined this research was able to explore population synthesis and microsimulation techniques.

Key findings

This exploratory research was conducted in a hierarchy of steps. A simple micro-inverse-drive-time model which only takes drive time into account when choosing which shopping centre to go to was constructed first to set up the frame for a more sophisticated model. A pseudo-validation can be conducted by comparing the predicted estimation with calibrated money flow from a CACI performed spatial interaction model. As a result, both in-town and out-town turnovers were over predicted. Two main causes were identified: a rather simplistic structure and latent demands. Latent demands cannot be estimated given the limits of datasets.

A micro-multi-factor model was constructed then. In this model, shopping behaviour based on different Acorn group, household composition and car availability as well as travel time were considered to differentiate people's choice of shopping destinations. By comparing with the previous output and output from the spatial interaction model, the output here was an improvement, but it still over predicted turnover. The main cause here was that the weights for each variables were set to 1 which was not realistic. The weights can be set reasonably only if there were data to identify how each variable contributes to individual decision or previous years' real turnover data could be an input for Genetic Algorithm (GA) to optimise the weights.

In conclusion, microsimulation can be a useful technique for store performance prediction. However, in order to be a viable alternative to spatial interaction modelling, more data on consumer behaviour needs to be available in order to estimate how micro characteristics are distributed, and how they influence consumer behaviour.



Comparison between traditional and agent-based turnover estimation.

Value of the Research

This research has presented a potential alternative or even a supplement to spatial interaction modelling for predicting store turnover. Microsimulation allows researchers and analysts to achieve a more detailed prediction of consumer behaviour through the inference of individual characteristics. By taking advantage of new techniques and big data sets produced by the retail industry, more intricate models of consumer interactions can be developed.