Consumer Data Research Centre

An ESRC Data Investment

Retail Research Masters Dissertations 2014

Edited by Guy Lansley

7th October 2014

www.cdrc.ac.uk

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Can smart meters save consumers and British Gas money and carbon by pinpointing which consumers are most likely and best placed to install insulation in their homes?

Nicholas Samsoni[1], Guy Lansley[1] & Andy Simpson[2]
[1]University College London, [2]British Gas plc

Project Background

British Gas, as with other energy providers, have certain energy-savings targets to fulfil as part of the Government's ECO (Energy Company Obligation) policy. These include providing wall cavity and loft insulation to customers' homes who could benefit from it. With no existing dataset regarding how energy-efficient consumers' homes are, this has to be inferred from other sources. One such source is smart meter data, a new dataset which is both extremely new and offers much higher-temporal resolution insight into customer's energy use than was previously available. As the aim was to focus on customers living in homes with cavity walls, only customers who are homeowners, not living in flats or terraced houses and whose homes were built between 1919 and 1976 were included in the study.

Data and Methods

The smart meter gas consumption data consisted of 48 half hourly daily readings averaged across Tuesday-Thursdays in winter 2013-14, excluding the Christmas and New Year period, to create a 'typical' midweek profile. The cleaned dataset represented approximately 40,000 customers. Customers were segmented into 4 load profile groups formed of 12 individual clusters using the K-means clustering algorithm, according to their average weekday energy use. An index of energy use by affluence for each distinctive load profile shape, which incorporated data about already insulated homes, was created and used to rank neighbourhoods at the Output Area level according to their suitability for targeted marketing of energy-efficiency measures. Areas were ranked according to their Output Area Classification (OAC).

Key Findings

The K-means algorithm was found to be an effective method of segmenting customers by their energy use. The 12 clusters were all distinctive and identified insightful large scale descrepencies between patterns in daily gas consumption between customers. The clusters were also amalgamated into 4 load profile groups (figure 2) which were named; Twin Peak, Flat Profiles, Evening Peak (figure 1) and Afernoon Dwellers. An additional 5th group of continuous high engery use was also identified but it only represented 0.5% of smart readers. Attributes of customers within each of these clusters were compared against the mean and were found to provide insightful information regarding domestic energy use. The data verified the idea that, for example, elderly customers tend to possess more homogenous daily profiles and that more affluent customers use more energy.

The research identified neighbourhoods that;

- contain customers who exhibited higher energy use per their affluence, thus have a higher energy-saving potential
- have customers who have shown willingness to accept energy-efficiency measures in the past, meaning they should be more likely to accept them in the future
- contain a high proportion of residents who are:
  - Homeowners
  - Living in homes built with cavity walls
  - Living in homes where measures can be installed with minimal disruption.

These neighbourhoods could be ranked using the OAC, (mapped in figure 3).

Value of the Research

British Gas were shown a means by which their smart meter data can be used to segment their customers using data other than solely their household composition attribute data. A method for identifying consumers who are most suitable for targeted marketing of energy efficiency measures was created and could be used to pinpoint these customers. I benefited greatly by gaining experience working with a new and real dataset on a meaningful and relevant project, receiving expert assistance and guidance from industry professionals.

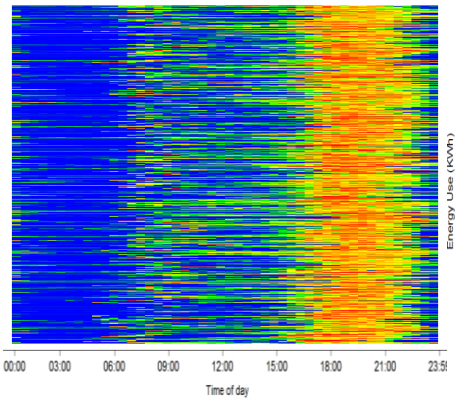


Fig. 1. Heatmap for group 'Evening Peak'

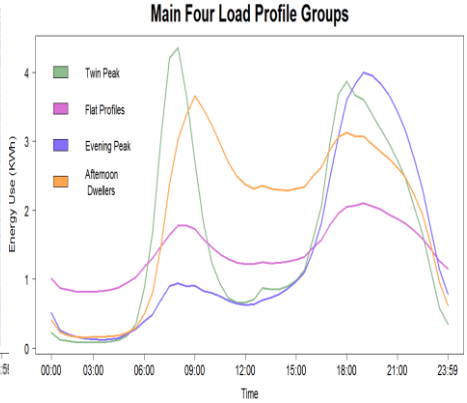


Fig. 2. Four load profile groups

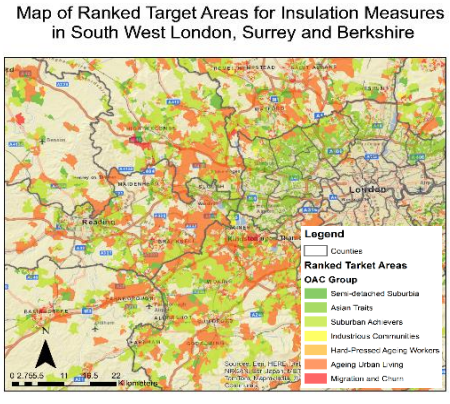


Fig. 3. Ranked output area classifications

A Geodemographic Segmentation Analysis of EasyJet's UK Customers

Elizabeth Townley[1], Guy Lansley[1] & Ben O'Brien[2]
[1]University College London, [2]EasyJet plc

Project Background

The objective of this study is to analyse EasyJet's customer and booking data to further understand who the airline's primary customers are. A range of different geographical analysis will be used to assist within this study project along with the use of K-means Clusters, this technique was able to highlight customers with similar characteristics based on their historic booking data and segment customers into clusters. Demographic analysis within the airline industry has been a rather recent subject due to the dramatic changes to the industry. Many of the past studies have been on a personal level which has limited the volume of customers included within the research, resulting in unrepresentative conclusions. This research is based on data provided by EasyJet to aid in producing clear and representable results for all EasyJet customers within the United Kingdom in conjunction with previous research and analysis.

Data and Methods

The data being used has been provided by EasyJet and covers a three year period from May 1st 2011 to May 1st 2014. This consisted of all bookings made within the UK and customers which have used EasyJet and departed from one of the UK airports. The data used within the project includes customer data, booking data, airport data, postcode data, boundary data and 2011 census data. The project was implemented using a variety of differing analysis approaches. To start with Population pyramids were constructed to visualise the age distribution of EasyJet customers. Customers address locations was also analysed using the postcode data provided with the booking data to create dot density and choropleth maps highlighting the density of customers within the UK. A Location Quotient (LQ) was carried out to highlight 'unique' regions or groups. This method was used against the Output area classifications to highlight which socio-economic groups have the highest percentage of customer using EasyJet. The next stage was aggregating destinations by winter/summer season to highlight a seasonal variance. Finally, a K-means analysis was carried out to cluster customers into distinctive groups based on destination and booking variables; which included the frequency of flights by a destination categorisation and season.

Key Findings

Firstly EasyJet customers were classified based on their date of birth to create a population pyramid (fig. 1). The results highlighted that 50 % of all female customer where aged between 21and 38, with the average age of the female customer being 40 and 43 for the male customers, resulting in the average age being 42. The results also identified that there was a slightly higher number of men than woman using the airline with 55% of customers being male. This could be due to the male association of business trips. To establish whether a customer living closer to an airport is more inclined to use air travel. A 30 mile radius around airports was established. Those living within a 30 miles radius of an airport on average have taken 1.9 trips a year compared to those living outside the radius taking 1.5 trips a year. The results to the OAC Location Quotient (LQ) displays the top three categories of whom are EasyJet Customers; Inner City Students, Comfortable Cosmopolitans and Aspiring Affluent (fig. 2). The OAC group least likely to be EasyJet's customers included the Migration and Churn, Industrious Communities and Hard Pressed Ageing Workers. Finally, the K-means detected 9 distinctive groups based on the EasyJet customer data.

Value of the Research

The benefits of performing this study have provided EasyJet with valuable information on their current customers. This will assist them to target the correct audience to gain and retain customers, keeping one step ahead of rival low cost airlines.

Fig. 1. Population pyramids of EasyJet customers

Ageing City Dwellers	59.73	Ethnic Family Life	106.76
Ageing Rural Dwellers	69.24	Farming Communities	84.15
Ageing Urban Living	128.29	Hard-Pressed Ageing Worke	40.63
Asian Traits	146.42	Industrious Communities	39.48
Aspirational Techies	218.40	Inner-City Students	388.12
Aspiring and Affluent	358.60	Migration and Churn	37.27
Challenged Asian Terraces	53.35	Rented Family Living	81.07
Challenged Diversity	47.16	Rural Tenants	73.68
Challenged Terraced Worke	58.46	Semi-Detached Suburbia	86.16
Comfortable Cosmopolitans	362.69	Students Around Campus	208.41
Constrained Flat Dwellers	222.55	Suburban Achievers	142.45
Endeavouring Ethnic Mix	139.38	Urban Professionals and Far	111.92
Ethnic Dynamics	86.46	White Communities	70.57

Fig. 2. Representation of EasyJet customers by 2011 OAC groups (100 = average penetration)

How Open Data Resources can be used to define a Composite Measure of Cultural Identity & Heritage for England and Wales

Yiran Wei[1], Guy Lansley[1] & Tim Rains[2]
[1]University College London, [2]J Sainsbury's plc

Project Background

Ethnicity has been a major subject in the realm of social research. Issues related to the definition, measurement and classification of ethnicity have attracted on-going debate in the literature. The multidimensional, subjective and complex nature of ethnicity has created barriers for researchers to measure and interpret the characteristics and distribution of various ethnic groups in the United Kingdom where the population has become more culturally diverse as a result of high levels of immigrants from all over the world. The purpose of this project is to identify the major spatial traits in ethnic identity in England and Wales using a range of cultural and heritage indicators that are available as open data sources. It will research ways in which these datasets can give a better understanding of overall cultural diversity at small area geography levels.

Data and Methods

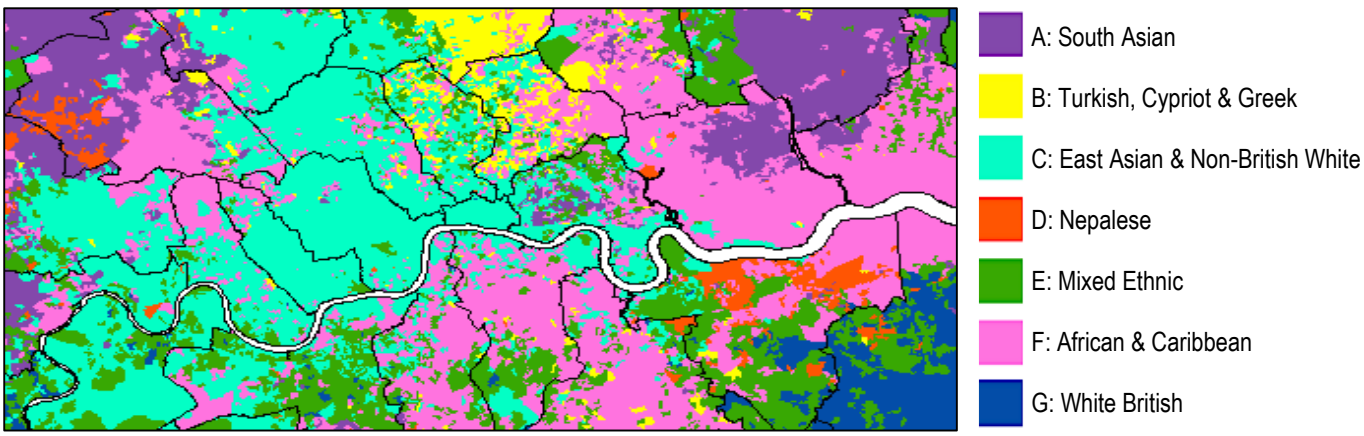
A total of 134 variables were selected from 2011 Census for England and Wales at Output Area level, covering 7 dimensions of cultural identity including country of birth, ethnic group, religion, main language, proficiency in England, age of arrival in UK, and length of stay in UK. The core method applied to create the ethnicity classification is known as cluster analysis; specifically, the k-means clustering approach was adopted for this study. This technique constructs groups of the most similar areas based on the overall similarities and dissimilarities between the variables. Prior to the clustering analysis, Pearson correlation analysis and principal component analysis were applied to reduce the list of 134 variables as much as possible without losing the essential information they contain. Afterwards, a total of 60 variables were identified as the best suited for the cluster analysis.

Key Findings

Every Output Areas in England and Wales were grouped into 7 clusters. Each of the groups is identified with distinctive characteristics of ethnicity components and spatial distribution pattern, including Group A South Asian, Group B Turkish, Cypriot and Greek, Group C East Asian and Non-British White, Group D Nepalese, Group E Mixed Ethnic Group, Group F Black African and Caribbean, and Group G White British. In addition, a London-specific Output Area classification has also been created alongside the England and Wales classification using an identical methodology. Comparison between the two reveals some interesting patterns. Both outputs are similar in general, however the London-specific version further separates the South Asian group into two subgroups: one reflecting more Muslim African ethnicity whilst the other is far more Bangladeshi or South-East European. Furthermore, Nectar data on sales of ethnic food is provided by Sainsbury and used to benchmark the clustering results. Cross-referencing the cluster analysis with retail sale data has suggested the validity of the classification through positive association. For example, the consumption of halal meat in African and Caribbean neighbourhoods is found to be approximately 30 times higher than the White British group. Finally, the examination of distance from the cluster centre for measuring the uncertainty indicates that variables chosen for classification appear to have been reasonably sound, but it also highlights the issue of subjectivity regarding the variable selection process.

Value of the Research

Retailers that successfully understand and target the shifting cultural makeup of society will outperform those using antiquated research and customer stereotypes. This research provides valuable insight for Sainsbury's ethnic food sale planning. The classification results retained from this study can provide a sound reference for Sainsbury's ethnic food allocating planning. By knowing the spatial trait of different ethnic group, Sainsbury can then adjust and further improve their ethnic food product range to meet the need of their ethnic minority customers.



2011 Cultural, Ethnic and Linguistic Output Area Classification for England and Wales – Central London

Click & Cannibalisation: does the move to convenience online entice new customers or simply move old ones?

Hoi-lam Chan[1], Prof. Graham Clarke[1] & Helen Parker[2]
[1]University of Leeds, [2]Tesco plc

Project Background

This report will analyse if the opening of Tesco Click & Collect grocery operations has the desired customer impact in England and Wales. GIS is used as a tool to visualise the spatial pattern of changing customer shopping behaviour. The aim is to identify how each Click & Collect grocery collection format appeals to various types of consumers in different market places. How Tesco's current strategy for the location choice of different grocery collection formats affect the sales and quality of its existing home delivery service is related to the varying demographics of the consumers across England and Wales is explored in detail

Data and Methods

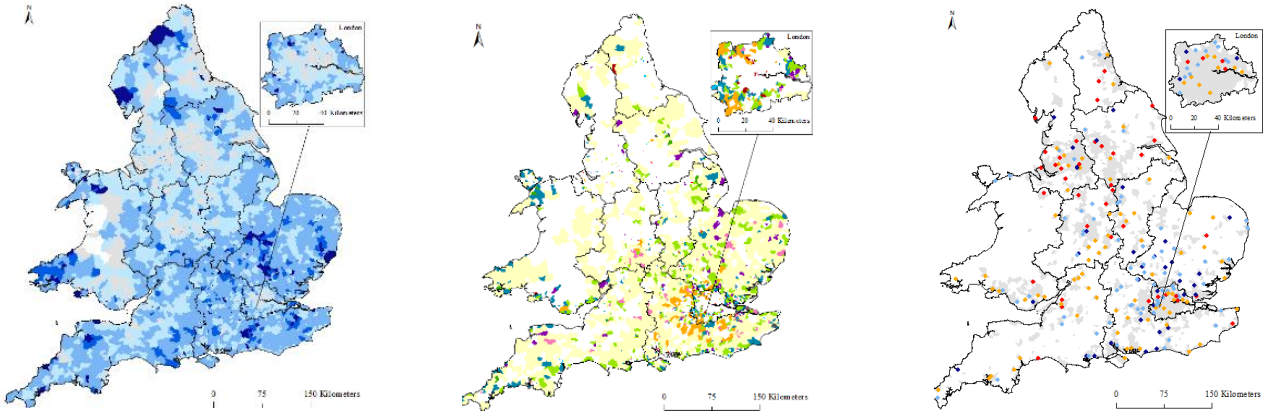
Tesco grocery home shopping data, open source data including 2011 UK census statistics and 2012 Family Expenditure Survey, consumer marketing data – CAMEO UK consumer segmentation data, and GIS data are used. Quantitative methods including measurement of market penetration and change in sales performance of both Tesco home delivery and Click & Collect grocery service are used to understand the spatial trend of changing customers shopping behaviour.

Key Findings

Both the existing home delivery and the emerging Click & Collect grocery service performed better towards southern regions and affluent neighbourhoods. However, the cannibalisation effect is not as bad as expected. The existing home delivery service tends to be favoured by the rural market while the adoption of the new Click & Collect service is more likely to be skewed to urban consumers. Six grocery collection formats are currently offered by Tesco. More new customers are attracted to drive-through pods, followed by integrated collection, and collection from temperature-controlled vans. Affluent consumers who live close to and with direct road connection to the collection points tend to be attracted to the new Click & Collect service. Less affluent consumers do not adapt to the new online service easily. For store collection, moving existing customers is achieved where the picking store is close to the collection point. The frequent Click & Collect shoppers usually do not live near the non-store collection points and collection points at Tube stations. The neighbourhoods near non-store collection points are increasingly aware of Tesco home delivery service instead.

Value of the Research

The findings would be useful for Tesco to understand its current position and to build a better Click & Collect grocery marketing strategy to boost the sales by maximizing the targeted customer satisfaction.



a) 2013-14 Tesco home delivery penetration b) Geodemographics of online consumers c.) Click & Collect performance

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
Consumer Data Research Centre
An ESRC Data Investment
AXA PPP HEALTHCARE
redefining / healthcare

European Union Government Open Data in Market Research: Exploring Current Discussion among Stakeholders on Opportunity, Accessibility, Usability and Desired Outcomes

Olga Tregubova[1], Justin O'Brien[1] & Martin Bellingham[2]
[1]Royal Holloway School of Management , [2]AXA PPP Healthcare ltd

Businesses navigating European markets are enriched by new ways open government data can create new opportunities and improve existing processes. Open data feed key insights for market segmentation, the creation of new markets and improvements in productivity. Similarly, government services can be improved with the input of proprietary data, which can provide honest and instantaneous feedback of success in government campaigns and demand for public services. Opening data encourages the growth of new services and products, thereby innovatively increasing profitability and creating value for not only government but also business and the wider economy.

The European Union is experiencing steady population growth, both organically and via accession of new member countries. Diversity within European populations, in terms of demographics, economic power, business practices and consumer culture, means that optimal strategy for organisations wishing to operate in these countries will likewise differ. Despite common trends in socio-economic and political thought, the EU is not homogenous with respect to consumer behaviour and shopping patterns. It follows that there are diverse markets therein, of varying size and market potential. To illustrate this point, these markets have individual purchasing powers ranging from 40% above to 60% below the European average.

In addition to varying economic and wealth indicators, markets within Europe possess differing societal and social contexts that must be considered when looking at the rapidly evolving demographics and their effect on retailing. Furthermore, distinctive national behaviours—rules, norms and beliefs surrounding economic activity that define or enforce socially accepted economic behaviour—contribute to differences in values surrounding governance and bureaucratic inertia, privacy risks, intellectual property rights, data quality, standards in collection and dissemination of data, as well as investment in data analysis.

A picture of these markets can be painted with the emergent tools of the open data movement. Open data originate from myriad sources including individuals, government, business and other institutions. More data is being used and reused for purposes other than those they were originally intended, and this paper focuses on public sector information (PSI) that governments collect, produce and ideally release in performing their public task. PSI can be used in synthesis with other types of data for innumerable purposes, most relevantly strategic decision-making and market research.

To understand the role of open data as a source of competitive advantage, this essay explores its use in market research, location decisions, market segmentation, merchandising and post-sales interaction. By interviewing stakeholders in a variety of roles in government, business and data environments, as well as reviewing various country progress reports and attempting to collate comparable data sets, this research begins to identify barriers to the realisation of value from open data within the EU.

This data is an asset with the potential to revolutionise business, government, research, crisis response and day to day activities, but it demands innovative processing before it can be useful as 'value-added' or 'derived' data. Already, companies are using complex algorithms in business flows to strategically expand into new markets, create insights, perform market research and detect fraud, among other uses. As the availability of data grows, so does the impetus to harness its power, and all parties must make difficult and responsible choices about how much and which data should be released.

In the dawn of the 'data deluge', the wealth of information is in the hands of the private sector. The shoe will soon be on the other foot—the historic exclusion of profit-making organisations from public data may backfire. Future research could explore the potential framework for sharing of information between the public and private sectors that takes into account due diligence to privacy, economic growth and the greater public interest and also factors in governments' strategic view of the future of open data. Such an exchange is less likely to be successful if NSIs continue to upset and exclude other data using communities and more likely if the parties enter into a partnership early.

Harnessing the power of social media to determine moving house flows and customers' energy needs.

William Trimble[1], Dr Paul Norman[1], Dr Nick Malleson[1] & Adam Rawson[2]
[1]University of Leeds, [2]E.ON UK plc

Project Background

The extraction and analysis of social media data has enabled researchers to derive new and intriguing datasets on a vast scale. This study extracted text strings (Tweets) from the micro-blogging platform Twitter to form a dataset. Twitter has been used in studies since its conception in 2006, yet it remains as a little used instrument for academic research. Existing research using Twitter datasets include determining the sentiment towards a certain subject matter, such as films and political preference. Other studies have involved epidemiology and earthquake location. The aim of this study is to determine if there is a correlation between Twitter data and E.ON data on the subject of moving house, to encourage the development of using Twitter as a data source. Secondarily, E.ON's online marketing strategies will be assessed.

Data and Methods

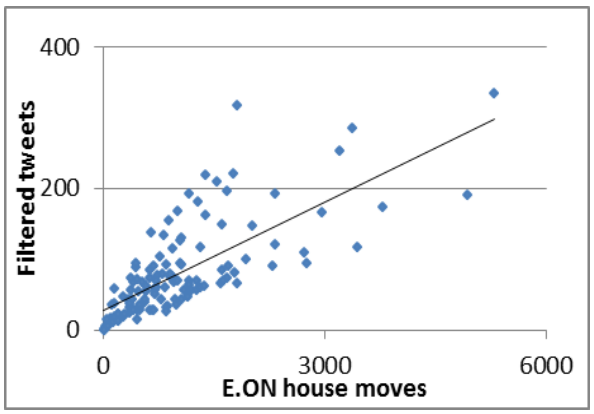
E.ON and Twitter datasets were received covering 3rd June – 25th July 2014. The moving house data supplied by E.ON originated from known customer moves in Great Britain. Geo-located Tweets, derived from a concise search related to moving house, were supplied by the University of Leeds. Both datasets were received and processed in Microsoft Excel. This involved discarding expendable metadata, and geo-locating postcode and place name locations to a coordinate system. A Geographical Information System (ArcGIS) was used to visually present the data on a postcode area scale, used to enable choropleth mapping of comparable regions. Correlation analysis, over time and by area, between the E.ON and Twitter datasets was conducted to fulfil the main aim of this study. Sentiment analysis using an online programme assessed customer attitudes towards E.ON's Twitter page.

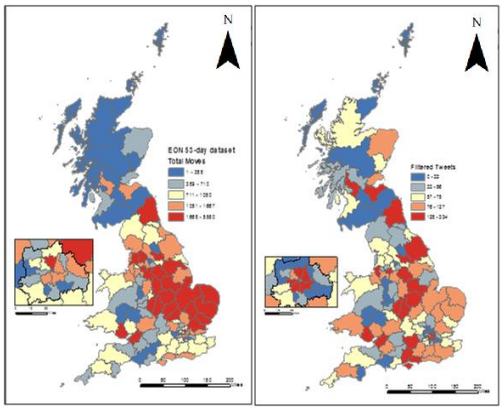
Key Findings

The correlation between the datasets was analysed in two ways. Firstly the correlation over time was calculated by segregating the datasets into eight, one week blocks. The Twitter dataset was separated into raw Tweets – derived from the initial search, and filtered Tweets – derived from manual elimination of Tweets not relating moving house. The Pearson's correlation coefficient found a positive correlation of 0.852 between the raw Tweets and the E.ON house moves data ($p<0.05$). Filtering the Tweets strengthened the correlation to 0.922 ($p<0.05$). The improvement of 0.07 due to filtering demonstrates that more advanced methods of processing could yield greater improvements. Correlation by postcode area involved segregating the datasets by their geographical location. This type of correlation sought to identify if there was a correlation between the datasets on a small geographic scale. The results found a strong positive correlation of 0.73 and 0.74 between the filtered Tweets and raw Tweets respectively ($p<0.05$) (figure 1). The choropleth map (figure 2) shows, despite some fuzziness within the middle quantiles, a broad degree of similarity. Duplicate analysis reveals that 52-54% of areas in the lowest quantile, and 68% of areas within the highest quantile from both datasets matched. Thus this analysis has shown that geographical and time variations both exhibit strong positive correlations between E.ON's real world dataset and a dataset derived from Twitter. Sentiment analysis showed that E.ON ranks third in terms of customer satisfaction on Twitter in relation to the 'big six' UK energy companies. Improvements such as homepage improvements and greater personal interaction were recommended.

Value of the Research

The outcome of this research will allow E.ON to use this study as a platform to develop the use of Twitter datasets. The use of simple methods in this study demonstrates the potential to utilise more advanced methods of data collection, such as a support vector machine (SVM). This study has also developed a personal understanding of how Tweets can be extracted as a data source, and how geographical elements play an essential role.





Figures 1 and 2 – comparing the spatial distribution of home movers with filtered Tweets

Segmentation of Clinical Commissioning Groups in England based on the prescribing patterns of their constituent General Practices

Ransome Mpini[1], Philip Worral[1] Helen Dickens[2] & Mark Bass [2]
[1]University of Westminster, [2]GlaxoSmithKline plc

Project Background

This project aims to enable GlaxoSmithKline to develop an understanding of how drugs are prescribed within England's Clinical Commissioning Groups (CCGs). The project groups CCGs based on the prescribing patterns they exhibit. The first part of the project aimed to cluster GP practices based on the drugs they prescribed (18 clusters in total). For each CCG, the cluster each of its practices fell into was identified and could be calculated as a relative proportion. Following this, CCGs were then clustered.

Data and Methods

GP practices are put into cluster groups based on their prescribing patterns for respiratory ailments. The resulting practice clusters are then analysed at CCG level to better understand the distribution of practice clusters within the CCGs. This gave an indication of the cohesiveness of a CCG where a cohesive CCG would have evidence of fewer practice clusters and CCGs with a wider variety of practice clusters is evidence of a less cohesive CCG.

Openly available datasets were used and these included the GP Prescribing dataset and the Quality and Outcomes Framework (QOF) dataset. Variables of interest in the GP prescribing dataset included the British National Formulary, which uniquely identifies each prescribed drug and it's Net Ingredient Cost of each prescribed drug at a GP practice. Another variable of interest was the Practice List Size, which was obtained from the QOF dataset. The practice list size values were used to normalize the Net Ingredient Cost figures so as to make them comparable between practices of different sizes.

The drugs prescribed at each GP practice were treated as the variables to describe the particular practice. The original prescribing dataset contains just over 10300 unique drugs across the GP's, although a reduced dataset of 551 drugs (focusing only on the respiratory drugs) was used for final analysis

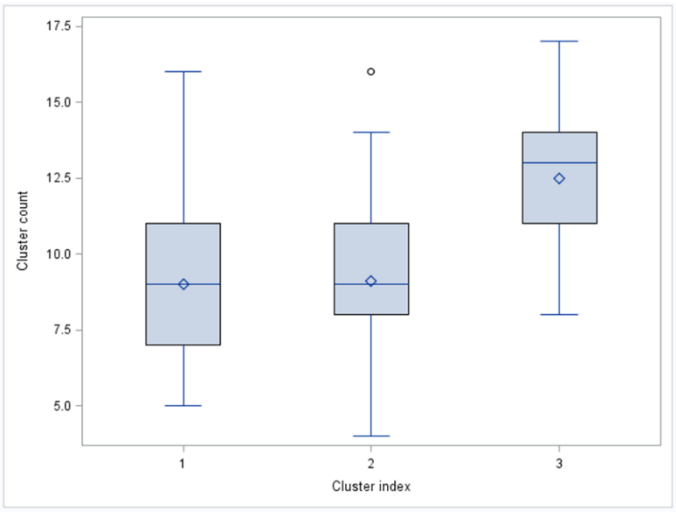
Key Findings

A cluster analysis using the 'Ward' clustering technique revealed that CCGs fall into 3 main clusters. The first is made up of smaller CCGs. The most common practice cluster in this group constitutes around 37% or below of the total practices in the CCG. The second cluster identified is made up of smaller CCGs but these tend to have higher ratios for the most common practice cluster. The third cluster identified is composed of larger CCGs where the most prevalent practice cluster groups have very small ratios and these tend to have the highest ratios for less prevalent practice cluster groups.

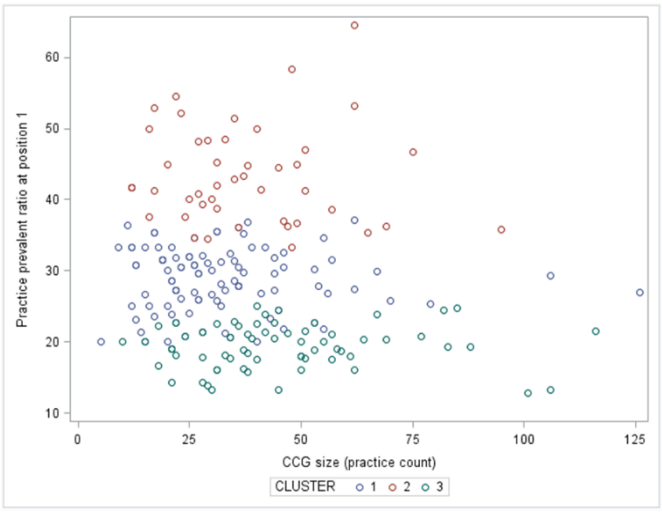
Complications were faced during the analytic phases of this project mainly caused by the sheer size of the GP Prescribing dataset that constituted just over 9383 GP practices (observations) described by just over 10,300 unique drugs prescribed at each GP practice which were treated as the variables describing the GP practices. A decision was made to focus on the respiratory chapter (BNF Chapter 3) reducing the total number of drugs (variables) to 551.

Value of the Research

This kind of work has not to the best of my knowledge been attempted and the result of this project will be a big contribution to GSK analytic process. It has the potential to enable them to better understand CCG cluster groups and also enable them to use this result as input to their core analytics process. Insight gained from this project will enable GSK to increase their competitive advantage through a better understanding of CCGs. GSK will be able to use evidence based, tailored messages when engaging CCGs thus maximizing returns on investment.



Practice cluster count distribution in CCG clusters



Plot of 1st positioned practice cluster ratio in CCG by CCG size