Official

ONS Big Data Project – Progress report: Qtr 4 October to Dec 2014

Jane Naylor, Nigel Swier, Susan Williams, Karen Gask, Rob Breton *Office for National Statistics*

Background

The amount of data that is generally available is growing exponentially and the speed at which it is made available is faster than ever. The variety of data that is available for analysis has increased and is available in many formats including audio, video, from computer logs, purchase transactions, sensors, social networking sites as well as traditional modes. These changes have led to the big data phenomena – large, often unstructured datasets that are available potentially in real time.

Like many other National Statistics Institutes (NSIs) the Office for National Statistics (ONS) recognises the importance of understanding the impact that big data may have on our statistical processes and outputs. So ONS established a 15 month Big Data Project to investigate the potential benefits alongside the challenges of using big data and associated technologies within official statistics. This is due to complete at the end of March 2015. In taking forward this work ONS is upholding all relevant legal and ethical obligations.

Summary

This report provides an overview of progress on the ONS Big Data Project during the fourth quarter (Oct – Dec 2014) and builds on the work that was documented in the first, second and third quarter progress reports[1]. An update is provided on the practical elements of the Big Data project: the four pilot projects covering economic and social themes. Each pilot uses a key big data source, namely Internet price data, Twitter messaging, smart meter data and mobile phone positioning data. Their objectives will collectively help ONS to understand the issues around accessing and handling big data as well as some of their potential applications within official statistics. Alongside the pilot projects a significant activity within the Big Data Project will be stakeholder engagement and communication.

[1] http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/the-ons-big-data-project/index.html

Contents

1 Introduction

The high level aims of the ONS Big Data Project are to:

- investigate the potential advantages that big data provides for official statistics; to understand the challenges with using these sources; and to establish an ONS policy on big data and a longer term strategy incorporating ONS's position within Government and internationally in this field; and
- make recommendations on the best way to support the ONS strategy on big data beyond the life of this project.

A major component of the project is to include some practical applications of big data, to both assess the role they might have within official statistics and to help understand the methodological, technical and privacy issues that may arise when handling them.

Four pilot projects have been chosen, covering economic and social themes. Each pilot uses a different big data source, namely Internet price data, Twitter messaging, smart meter data and mobile phone positioning data.

Although ONS is researching only samples of these data, even these can be too large and complex to process efficiently using standard ONS computers. The solution is to use the ONS innovation labs, a private 'cloud' based environment, for analysing them.

This report briefly introduces the ONS innovation labs, then provides an overview of progress on the four pilot projects in the fourth quarter (Oct – Dec 2014). In addition a summary of progress around stakeholder engagement for the project is provided, an important activity for the project. This report builds on the work that was documented in the first, second and third quarter progress reports[2].

In all these activities ONS is committed to protecting the confidentiality of all the information it holds. In order to produce statistics using big data sources we are interested only in trends or patterns that can be observed, not in data about individuals. However, we recognise that accessing data from the private sector or from the internet may raise concerns around security and privacy. The Big Data Project is therefore accessing only publically available, anonymous or aggregated data and this data will be used only for statistical research purposes. In addition all of our work fully complies with legal requirements and our obligations under the Code of Practice for Official Statistics.

2 Innovation labs

The ONS innovation labs have been set up to help facilitate research into new technologies and open source tools, new sources of public data, and to develop associated skills. The innovation labs are a key enabler for the ONS Big Data project because they allow us to handle large and complex data sets and to test new big data technologies.

[2] http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/the-ons-big-data-project/index.html

These labs consist of a number of high-specification desktop computers with some additional network storage. The hardware is configured using OpenStack[3] technology. This provides a very flexible environment to deploy different 'virtual environments' depending on the processing and storage requirements of different projects. In particular, this approach will provide a flexible framework for experimenting with big data parallel computing technologies such as Hadoop[4]. The innovation labs have been designed to provide a route for accessing open source tools.

We have placed restrictions on the data that can be accessed in the labs. In the Big Data Project these are currently confined to the Twitter and internet price data pilots, which are using publicly available data, and the analysis of anonymous smart-type meter information.

Within the labs, a lot of focus has continued to be on data analysis this quarter. Methods for processing large datasets have been explored - such as utilising parallel processing. Processing time has been improved for some analytical tasks by huge margins - from 12 hours down to 30 minutes in some cases using smart techniques and multiple processor packages available in Python. There is continued exploration of technologies such as MongoDB and Hadoop. Initial implementation of a trial Hadoop cluster has been completed although it has yet to be fully tested.

3 Prices pilot

Background

Web scrapers are software tools for extracting data from web pages. The growth of on-line retailing over recent years means that many goods and services and associated price information can be found online. The Consumer Price Index (CPI) and the Retail Price Index (RPI) are key economic indicators produced by ONS. Web scraping could provide an opportunity for ONS to collect prices for some goods and services automatically rather than physically visiting stores. This offers a range of potential benefits including reduced collection costs, increased coverage (ie more basket items and/or products), and increased frequency.

Supermarket grocery prices have been identified as an initial area for investigation because food and beverages are an important component of the CPI and RPI basket of goods and services.

Research objectives

The objectives are to:

- Set up and maintain prototype web scrapers to test the technical feasibility of collecting price data from supermarket websites.
- Develop methods for quality assuring scraped data.
- Compare scraped data with data collected using current methods, explore methodological issues with scraping prices from supermarket websites
- Establish whether price data could be sourced directly from commercial companies and if so, how these compare with data scraped by ONS prototypes.
- Evaluate the costs and benefits of these alternative approaches to collecting price data

[3] http://www.openstack.org/
[4] http://hadoop.apache.org/

Progress

Data collection

The daily web-scraping operation for a selection of 35 item categories and three online supermarkets continued during the quarter with some further development and refinement of exploratory data analysis to investigate anomalies in the data.

The experience of this pilot so far suggests that a web-scraping operation of this scale is a cost-effective means of collecting large volumes of data. However, it is important to stress that this pilot has involved scraping only three websites; maintenance overheads would increase with an operation scraping a larger number of websites.

Data analysis and methodology development

Good progress has been made with the initial analyses and evaluation of these data. The team has been building on the analysis completed by the University of Huddersfield. Exploratory data analysis is revealing several interesting patterns. Some key examples of the analysis are given below:

- For the months of October and November 2014[5]: Around 23% of prices were on discount – and around half of these were estimated to be a multi-buy discount. Discounts are of interest to index creation as they can affect inflation estimates. The high volume data the scrapers are collecting permits detailed analysis into the distribution and prevalence of discounts – which is not possible with current collection methods. While discounts are included in the index, multi-buy discounts are not, hence the particular interest in this aspect. The analysis demonstrates that multi-buy discounts are common. However, the analysis cannot provide an estimate of take-up by consumers - hence it only presents a partial picture into the impact of multi-buy discounts on the 35 items.

- Bi-modal and multi-modal price distributions appear to be common amongst the CPI/RPI item classifications (indicating more than one price grouping within each item category). The high volume of data allows detailed price distribution analysis at the CPI item level.

We have started to produce early prototype indices using web scraped data. The timeliness, high frequency and volume of the data allow many index specifications that are not possible with the current collection. The volume of the data also allows the creation of item level indices - with much greater numbers of individual items. These item level indices could be interpreted as being more representative - due to the higher item volumes.

Commercial data purchase:

The pilot team have purchased three years of data from MySupermarket.com. The data covers the 35 CPI/RPI items currently being collected by the team's web scrapers. The data is being used to

[5] The analysis is restricted to October and November 2014 as discount information was a late addition to the web scrapers (due to website complexity); prior to October discount information was only collected for one supermarket

assess whether it is more cost efficient to purchase the data scraped by MySupermarket.com than web scrape it (in-house).This data will form a major part of the next phase of research.

Future work

The main focus is the completion of the prices report bringing together all elements of the prices pilot, conducted over the 15 month project, to conclude phase one of the research. The key elements being:

- Collection (web scraping)
- Cleaning and manipulation of the data (wrangling)
- Analysis of the data

It is likely that part of future research will investigate machine learning, MySupermarket.com data and the methodology of index number creation using web scraped data.

4 Twitter pilot

Background

Twitter is a micro-blogging site which has become one of the leading social networking platforms. Most tweets are public data and Twitter provides open source tools for accessing these data (albeit with some limits). Twitter provides an option for users to identify their current location. This means that tweets from a subset of users can be tied to specific locations over time. This data can then be used to track mobility patterns (eg Halwelka et al 2013).

A historic weakness of England and Wales mid-year population estimates has been capturing the internal migration of students. Students typically move to different parts of the country when they commence studies and then move to a new location again when they graduate and find employment. The main source for estimating internal migration is the GP patient register but young people, especially young men, are often slow to re-register when they move. These populations are more likely to use Twitter than other populations (Koetsier 2013).

The primary aim of this research is to determine whether geo-located data from Twitter can provide fresh insights into internal migration within England and Wales and whether these insights could be used to improve current estimation methods.

Even though these data are all publicly available, the pilot team is very conscious of the ethical issues around how these data will be used and will handle the data appropriately. Although we are working with data at the individual level (which is publicly available) our research question and ultimate interest is around patterns and trends in internal migration at the aggregate level, eg for groups within the population such as students in a particular city.

Research objectives

The objectives are to:

- Develop an application to harvest geo-located tweets from the live Twitter stream.
- Develop a method for processing these data by user to identify clusters and to derive different cluster types (ie home, work, study, and commutes).
- Develop a method for detecting changes in cluster patterns over time that could be interpreted as internal migration.
- Compare these results with current internal migration estimates and census data to understand their coverage and any resulting bias, and to establish whether these data are useful.
- Identify any big data technologies that may be needed if this research is to be taken forward over the longer term.

Progress

Data collection

As outlined in the third progress report, the Twitter harvesting application developed by the pilot was stopped on 15 August due to compliance issues. Additional data was purchased from Twitter covering the period from 15 August to 31 October 2014 as well as for the period from 1 April to the 10 April 2014. This data has now been downloaded, processed and combined with the data collected internally (for which we have permission to use) to give seven complete months of data, just under 100 million tweets.

Data processing

The pilot has made good progress in developing methods for clustering the raw data for each user to identify frequently visited locations, or anchor points using a DBSCAN algorithm (Backlund et al, 2011) written in Python. These clusters are later used as the primary units of analysis for identifying mobility patterns over time. Using this type of algorithm with large amounts of data is heavily processing intensive because the number of calculations increases exponentially with the number points to be clustered (Tsai & Wu, 2009).

Excellent progress has been made to improve the efficiency of this processing algorithm through a series of incremental steps:

1. A pre-processing step to identify high-frequency users and process them separately

2. A simple implementation of parallel computing that has proved very effective in boosting overall processing times.

3. Creating more efficient code

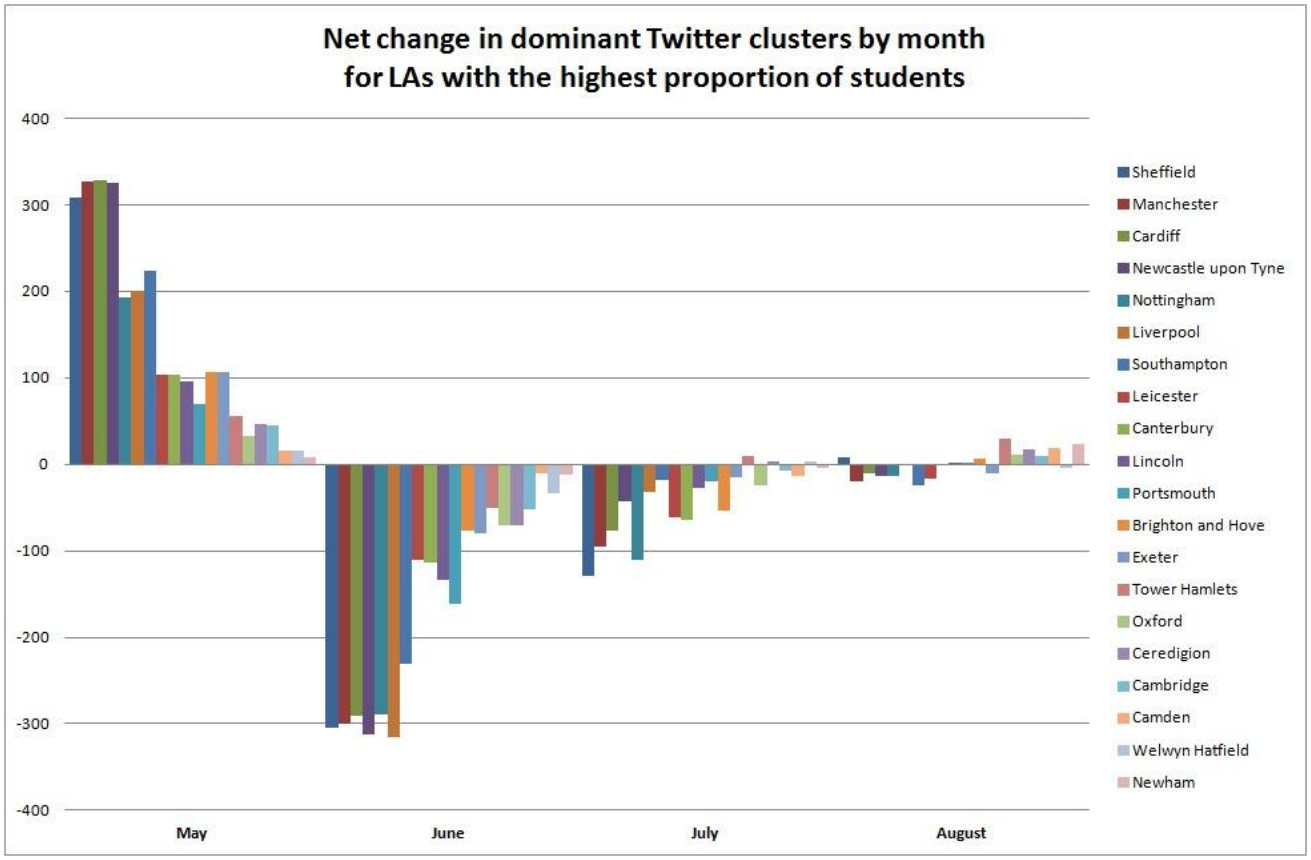
In combination, these steps have delivered huge gains in processing efficiency. Initially it was taking about a week to process 60 million data points but this was reduced to 30 minutes.

The pilot has successfully developed a method of using AddressBase[6] to classify clusters by type (eg residential or commercial). The main purpose of this processing step is to help distinguish residential addresses from other types of address. This step would enable commercial addresses to be removed and to focus on residential addresses, which are more likely to be where the user actually lives.

Analysis

Initial analysis has focused on the data collected directly by the pilot between 10 April and 15 August 2014. Figure 1 shows a monthly comparison of the net change in the number of dominant clusters for the 20 local authorities with the highest proportion of students based on the 2011 Census. A dominant cluster is the residential cluster for each user with the most data points in each time period. A change in the number of dominant clusters by local authority reflects changes in the predominant locations of where people are for each month. These can be viewed as indicators, rather than estimates, of change in the de facto population over time. Reliable estimates would require measures of bias within these data, which could then be used to weight up these counts. This is a potential avenue for further research.

Figure 1 : Net change in dominant Twitter clusters by month for LAs with the highest proportion of students



[6] AddressBase is the definitive source of address information within Great Britain and is available to public sector organisations under the Public Sector Mapping Agreement.

There is an implied net movement of people into areas with high student populations in May which is consistent with students returning from Easter break to sit final exams. In June, there is a reversal of this pattern which is consistent with students leaving for the summer break.

These results suggest that despite the many issues around representativeness for this data source, it is possible to detect particular patterns of population movement throughout the year. Although the fact that these types of month to month movements exist is not surprising, it is a useful reminder that patterns of population movement are much more complex than those implied by official population estimates. Official estimates are focused on the long-term underlying trends of population change. These are very important measures, but they mask the 'pulsing' effect of population movements, such as month on month changes for student populations which might be better represented in big data sources such as Twitter data.

Future work

The pilot is now focused on finalising analysis for the final report.

References

Backlund H, A. Hedblom, N. Neijman, 2011, Linkopings Universitet, "DBSCAN - A Density-Based Spatial Clustering of Application with Noise" Available at: http://staffwww.itn.liu.se/~aidvi/courses/06/dm/Seminars2011/DBSCAN(4).pdf Accessed on 25-03-2014

Koetsier, J. 2013, "Only 16% of U.S. adults use Twitter, but they are young, smart and rich". Available at: http://venturebeat.com/2013/11/04/only-16-of-u-s-adults-use-twitter-but-theyre-smart-young-and-rich/ Accessed on 18-03-2014

Hawelka, B, I Sitko, Euro Beinat, S Sobolevsky, P Kazakopoulos and C Ratti, 2013 "Geo-located Twitter as the proxy for global mobility patterns" http://arxiv.org/abs/1311.0680 Accessed on 19-03-2014

Tsai C, C. Wu, 2009, "GF-DBSCAN A New Efficient and Effective Data Clustering Technique for Large Databases", Proceedings of the 9th WSEAS International Conference on Multimedia Systems and Signal processing", Aavailable at: http://www.wseas.us/e-library/conferences/2009/hangzhou/MUSP/MUSP38.pdf Accessed on 15-10-2014

5 Smart meter pilot

Background

A smart meter is an electronic device that records and stores consumption information of either electric, gas or water at frequent intervals. These data can be transmitted wirelessly to a central system for monitoring and billing purposes.

The European Commission's Energy Efficiency Directive (EED 2012)[7] is a common framework of measures for the promotion of energy efficiency within the EU. It supports the EU's 2020 headline target on a 20% reduction in energy consumption, and its provision[8] for the roll-out of smart meters requires member states to ensure that at least 80% of consumers have such intelligent electricity metering systems by 2020.

The Department of Energy and Climate Change (DECC) has one of the most ambitious rollout policies within the EU: to put electricity and gas smart meters in every home in England by 2020[9] with roll out starting in 2015.

For electricity, readings will have a minimum specification of 30 minute intervals and will be transmitted at predefined intervals to a body called the Data and Communications Company (DCC). Data access will be permitted for certain specific functions as described in legislation.

Smart meter electricity energy usage data is attractive to statistical organisations as it, subject to data access, potentially allows investigation at low levels of geography and high levels of timeliness. Additionally, within England, these data would represent an almost complete coverage of homes.

It is important to stress that, although energy usage data from trials of smart meter type devices are available for research on individual meters, there are many ethical and legal issues involved with access to these data when the rollout is underway. Research may necessarily involve analysis at individual level, although the ultimate goal will be to develop methods of producing small area estimates or other outputs which preserve the confidentiality rights of consumers.

There is a growing interest in using smart meter data across statistical organisations globally. The applications of most interest for the production of official statistics are:

1. Energy usage and expenditure which is of key interest to policies concerning the management of energy demand/supply in the longer term. For example, the frequency of smart meter data facilitates analysis of energy demand with weather effects such as temperature and rainfall. If relationships can be identified then weather data may provide a useful indicator to identifying energy usage trends at a national/regional level, reducing the need to source smart meter data directly or to collect energy spending through surveys.

2. Occupancy status of homes: low and constant electricity use over a period might indicate that a home is unoccupied. This might have application to a single day or a longer period if wanting to identify long-term vacant properties. Feasibly, small area estimates on the likelihood of homes being occupied on certain days and at certain times might be achieved which could benefit fieldwork processes in national surveys.

3. Household size or structure: it is hypothesised that profiles of energy use during the day might vary by household size or the composition of a household's inhabitants. Small area estimates might be developed.

[7] http://ec.europa.eu/energy/efficiency/eed/eed_en.htm
[8] This provision relates to another EU Directive on smart meter rollout (2009) which required a full cost/benefit analysis be performed prior to commencing rollout
[9] Wales and Northern Ireland have similar policies.

The University of Southampton have been commissioned by ONS to conduct a small research project to investigate the potential of using smart-type meter data to identify household size/structure and the likelihood of occupancy during the day. A final report on this work is being prepared for publication.

The ultimate aim for this research is to develop methods to produce small area estimates for use within either statistical outputs or operational processes such as fieldwork. However, as a first step, it is necessary to work at an individual (yet still anonymous) level to understand patterns of energy usage. Initial research proposals have been discussed with the Government Digital Service Privacy and Consumer Advisory Group and the ONS Beyond 2011 Privacy Advisory Group. If the research is successful and suggest there is real value to be had in developing these small area estimates, the privacy and ethical issues surrounding the use of this data will need increased consideration.

Research objectives

The objectives are to:

- Understand the big data technical/methodological challenges of handling this type of data
- Assess some of the quality aspects of smart meter type data and to form ideas on how to approach further analysis. For example, how to deal with missing values etc.
- Produce higher analysis: to focus on smart meter profiles for identifying occupancy patterns. Less priority to be given to household size/structure or data-led analysis such as a cluster analysis (dependent on data handling restrictions and analyst resource availability)
- Review research studies in academia and other NSIs
- Research the ethical and public perception issues surrounding this type of data
- Identify the cost/benefit to ONS for using smart meter data in specific applications
- Propose future use and further ONS research with this type of data (final report).

Progress

Data collected during consumer trials of smart-type meters have previously been sourced from the Irish Social Science Data Archive. Around 4,000 residential homes are included in this data, and anonymised samples of these were taken to start preliminary analysis to understand the data and to help identify methods of analysing them.

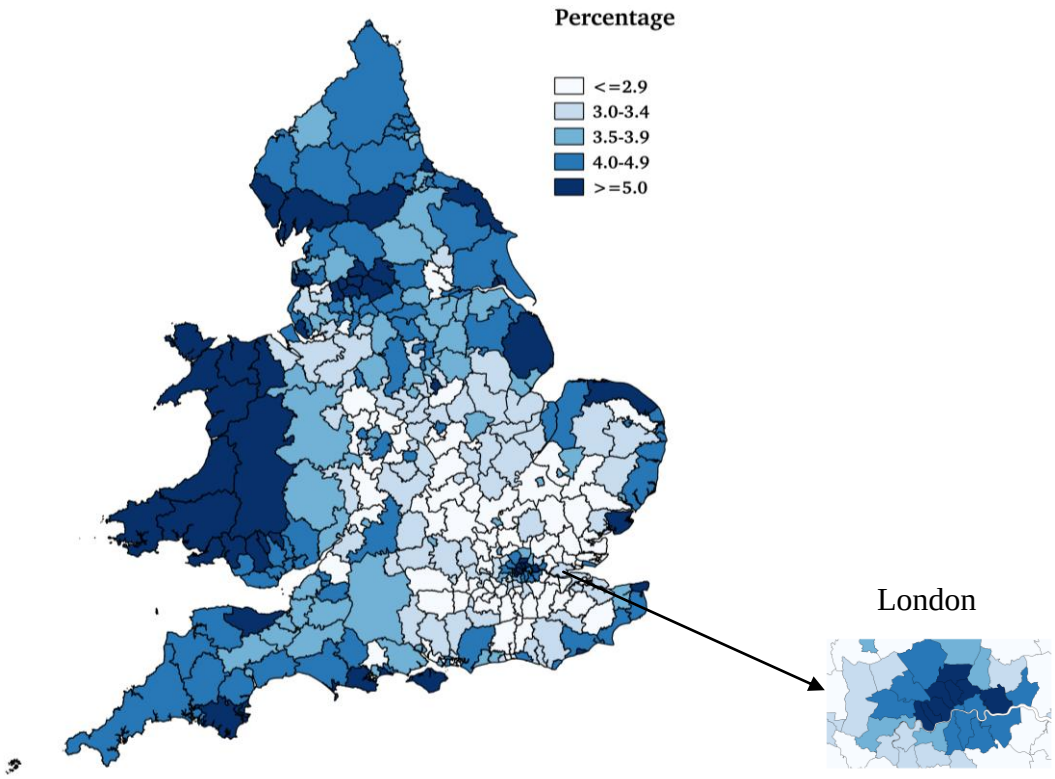
The focus of the research is to assess if an algorithm can be developed to automatically identify whole days when households are unoccupied. The rationale being that a retrospective look at smart-type meter data may highlight, for example, the number of homes unoccupied on census day, which might then be used to validate census results.

During this quarter the research completed so far was written up into a final report for publication. In addition to statistical analysis, the report considers the ethical issues and data access arrangements.

Research has continued into the logical extension of identifying methods which may identify longer term vacant properties, of great use within the production of such estimates, and valuable intelligence in a census or survey operation.

DECC have additionally provided ONS with Local Authority level counts of standard electricity meters in England and Wales, for 2011 and 2012, differentiated by annual usage of electricity. Under the assumption that each meter represents a household, Figure 2 shows that there is a geographic pattern in the percentage of households which used less than 500 kWh of electricity in 2012. This level of electricity usage, by a household over a year, is very low and might be a potential indicator of a long term vacant property. On average 4 per cent of households in England and Wales used less than 500 kWh of electricity in 2012. However higher percentages of such households are observed in coastal holiday areas such as Wales, the Norfolk coast and Devon. Central London also has higher percentages and this is known as an area with a high concentration of second homes which are used for work[10]. Comparison of these counts with 2011 Census data on second homes and vacants is ongoing.

Figure 2: Percentage of households using <500 kWh of electricity in 2012



Future work

Over the next three months the final pilot report will be reviewed and published. A smaller report about identifying long term unoccupied households from smart meter data will also be published.

[10] http://www.ons.gov.uk/ons/rel/census/2011-census/second-address-estimates-for-local-authorities-in-england-and-wales/stb-census-2011-second-addresses-in-e-w.html#tab-Working-second-addresses

Two strands of smart-type meter research will commence in the coming three months:

- Using machine learning methods such as logistic regression to identify households unoccupied for a whole day
- Analysing data from the Energy Demand Research Project. DECC have published this smart-type meter data which contains half-hourly electricity readings for 14,500 households with a smart-type meter in Great Britain between 2008 and 2010

6 Mobile phone pilot

Background

Location data generated through mobile phone usage is of key interest to statistical organisations because it has the potential to inform various important aspects of population behaviour. Current research around the world is focussed on:

- Population densities – at specific times of the day and/or small geographies
- Population flows – for example the number of people who travel from area A to area B
- Tourism statistics[11] – a Eurostat funded feasibility study on the use of mobile positioning data for tourism statistics has generated research within a number of NSIs, most notably Statistics Estonia, Statistics Finland and CSO Ireland.

There are a number of features, specific to these data, that have supported this growing interest including:

- The high coverage of the population who have mobile phones (93% of UK adults[12])
- There are relatively few service providers, so any one provider might have sufficient coverage to produce reasonably representative insights of total population behaviour, reducing the effort required in approaching multiple companies.
- The growth of big data technologies and methods is allowing the service providers to do more and more with their customers' data. Since 2012 the UK's main providers - Telefonica, Everything Everywhere and Vodafone - have all embarked on initiatives to use their customers' data within the development of new data products for sale.

Historically there are many academic research projects demonstrating a use of 'call event' data which contains location information when a customer receives or sends a text/phone call. Of more interest is the use of the geolocational data which is passively generated from mobile phones when they are switched on and either move between cell towers or send out a location reading at intervals (known as 'pinging').

[11] http://www.congress.is/11thtourismstatisticsforum/papers/Rein_Ahas.pdf
[12] Ofcom facts and figures report 2014

It is speculated that this geolocational data might be used to produce travel patterns from an origin to a destination location. ONS has an interest in whether this might be extended to travel patterns for 'workers' as typically produced in a census.

ONS is keen to proceed with this opportunity by approaching mobile service providers to see if they would be willing to produce aggregated counts of such travel patterns for comparison with 2011 Census data, Progress has been delayed due to Government Digital Service (GDS) sensitivity around government departments being seen to be accessing these data, however as at end June 2014, this pilot project has been given internal authorisation to proceed.

Research objectives

Objectives are to:

- Investigate options in sourcing aggregate data on travel patterns of workers from one of the main UK mobile phone providers. There will be an emphasis on understanding the issues involved throughout the stakeholder engagement, negotiation and procurement stages of this 'partnership' opportunity;
- Investigate intelligence around methods used for deriving worker flows using mobile phone data and monitor key considerations;
- Subject to obtaining data: compare the aggregated mobile phone data to 2011 Census data on travel to work flows to assess some of the quality aspects of mobile positioning data and to form ideas on how to approach further analysis;
- Review research studies in academia and other NSIs;
- Research the ethical and public perception issues surrounding this type of data;
- Propose future use within ONS for mobile phone data (final report).

Progress

Stakeholder engagement

Meetings have been held with DfT and other transport bodies in this quarter and intelligence gathered on the use and acquisition of mobile phone data for example:

- Procurement of mobile phone data tends to be through the engagement of a transport consultancy whose responsibility it is to acquire mobile phone data from a mobile phone data provider, possibly via an intermediate data broker. It is unclear what level of modelling is performed by the mobile phone provider, the data broker or the transport consultancy.

- The research objective in public sector transport bodies is to see if mobile phone data can replace or reduce the need for road side surveys which are the traditional means of gathering transport evidence related to transport initiatives. Roadside surveys are expensive and increasingly difficult to field as they require police presence and generate much annoyance and reducing response rates within road users.

Additional internal research

Over the past quarter research using Oyster card data has been prepared as a short paper for publication and is undergoing review.

As detailed in previous update reports, Oyster card data on the counts of journeys from an origin tube station (where an Oyster card first enters the network) to a destination tube station (where the same Oyster card leaves the network) is publicly available. Furthermore, the flows are broken down by time period including journeys starting between the peak travel time of 7am and 10am. ONS used these data to see if the flows of journeys conducted in peak travel time compared well with 2011 Census estimates of travel to work for those travelling mainly by underground metro, light rail or tram. The research shows that the flows correlate reasonably well, although distortions are evident at train interchanges and mainline train stations in particular, where many commuters with Oyster cards enter the tube system. These counts are much larger than the corresponding Census counts of people living in these areas.

Future work

It is proposed that research using mobile phone data for transport flows will be taken forward by the Government Data Science Partnership (GDSP) with ONS leading. The GDSP is a new group set up to provide coordination for data science research across government and includes representatives from the Government Statistical Service, Cabinet Office, GDS and Go-Science.

In parallel to this continuing engagement, a report will be written detailing the current understanding of research activity across public sector organisations together with the key considerations for using mobile phone data in transport flows.

The Oyster card research will be finalised and published as a short paper.

7 Stakeholder engagement

A significant big data project activity is stakeholder engagement and communication. Stakeholder engagement activities seek to achieve the following through communication and other means:

- Engage with data users/the public to understand their concerns around the use of big data within official statistics, and their requirements for new types of outputs
- Engage with external stakeholders to acquire their data/tools/technologies for use in pilot projects
- Engage with external stakeholders to learn from their experience, to develop our knowledge and skills, co-ordinate efforts, to develop partnerships and work collaboratively with them
- Engage with internal stakeholders to co-ordinate efforts, to ensure the project's objectives align with ONS strategic objectives, and to ensure support for the project across the ONS
- Manage stakeholder expectations at various stages of the programme.

The following nine groups of stakeholders have been identified for the project:

- Privacy groups

- International
- Academia
- Private sector
- 'Big Data' companies
- Technology providers
- Government
- ONS
- Data users including the public.

In this fourth quarter of the project key stakeholder groups have been government (developing proposals for and initiating collaborative opportunities) and internal ONS stakeholders (to gain support for future work in this area). In addition engagement has increased with academics to raise awareness of the project, identify common interests and collaborative opportunities.

In the fifth quarter of the project these activities will continue, with continued focus on taking forward future initiatives focused on big data across government.

Key activities in these stakeholder groups are provided below:

- During the first three quarters of the project the ONS Big Data team have contributed to the Cabinet Office Data Science Programme through attendance at the Community of Interest meetings and a cross-profession working group focused on capability and have developed relationships with other key data science advocates across Government, eg Government Digital Service (GDS) and GO-Science. This quarter has seen this relationship formalised with the proposal for a Government Data Science Partnership (GDSP). The key objectives of the GDSP are to deliver high quality data science and build wider government capability through collaborative and coordinated activities across Government. A high level proposal for the GDSP has been developed and signed off and an initial meeting of the key players held to develop a work plan.

- In addition to the activities described above a number of bilateral meetings/conversations/presentations have been held with representatives from different government departments in order to move forward the work of the project, share experiences and investigate collaborative opportunities:

  - A number of conversations/meetings have been held with Department of Transport and other transport bodies to gather intelligence around the use of mobile phone data
  - Discussions with statisticians from Department of Energy and Climate Change around the acquisition of data to support the smart meter pilot
  - Seminar given to Bank of England staff providing an overview of the project

- The Economic and Social Research Council (ESRC) have significant funding to invest in a Big Data Network to help optimise data that is available for research. Discussions have been held with the ESRC to explore the possibility of jointly funding research into public attitudes of the use of big data for research/official statistics.

- Links between the ONS Big Data Project and the RSS will be strengthened during the next quarter of the project. The RSS are hosting a workshop and public meeting on the topic of the future of official statistics in the big data era. Members of the ONS Big Data team will attend and contribute to the workshop and the National Statistician has been invited to be a member of the panel for the public meeting.

- We have also undertaken focused engagement with a number of UK universities offering courses on big data/data science/data analytics, in particular Royal Holloway, Lancaster and Southampton University (the Web and Internet Science group) to understand the courses they offer, the types of skills new graduates studying in this field will have and the research activities that are being undertaken. This engagement has led to the creation of two vacancies for student placements within the ONS Big Data Project, one for a 3 month period and one for a 12 month period, both starting Summer 2015. Recruitment for these vacancies will be pursued in the next quarter.

- In addition to the cross cutting activities described above the ONS Big Data team will liaise with specific academics to provide support to the particular pilot projects:
  - The ONS Big Data Project is represented on the steering group of an ESRC funded Southampton led project[13] which is focused on exploring the feasibility of estimating small area statistics from transactional 'big data' including energy monitoring data and hence building on the ONS funded work on smart meters.
  - Academics from S3RI (Southampton Statistical Sciences Research Institute) will provide support to the analysis of data collected from Twitter
  - The ONS Big Data team will continue to engage with academics from Cardiff University over the Twitter pilot

- Engagement with the private sector has focused on the acquisition of data to be used within the pilot projects.
  - 3 years of daily price quote data has been purchased from MySupermarket.com (Jan 2012 to Dec 2014) for a selection of item categories and supermarkets. Analysis will be undertaken to determine whether this data allows products to be tracked across time, if it can then purchasing these data might be a better long-term option than scraping the data ourselves.
  - In the previous quarter we gained agreement from Twitter that we should stop harvesting Tweets through our application (but that we could analyse the data collected to date) and purchase future data required for analysis through GNIP, a company owned by Twitter offering social media data for purchase. Engagement with GNIP has focused on the specification and negotiation over our data requirements and additional data has been purchased and used with the Twitter pilot analysis.

[13] http://www.energy.soton.ac.uk/category/research/energy-behaviour/census-2022/

- The ONS Big Data Project continue to contribute to the UNECE international collaboration project focused on big data[14] through two of the task teams focused on partnerships and technology.

- A European Statistical System (ESS) taskforce on big data and official statistics has also been established. Members of the ONS Big Data team are contributing to the taskforce which is focused on the Scheveningen Memorandum[15] and its implementation through an action plan and roadmap. The main focus of this is a series of practical pilot projects, the purpose of which is to gain experience of big data within official statistics in the European context.

- Members of the ONS Big Data team will attend the 'New Techniques and Technologies' official statistics conference in Brussels in March. Papers will be presented on the ONS project overall, the Prices pilot and the Smart meter pilot. This will provide an opportunity to expose our work to an international audience, to get quality assurance and review and to make and develop contacts with data scientists from other NSIs.

8 Conclusions

This report has provided an overview of progress on the ONS Big Data Project during the fourth quarter of 2014. Updates on the practical elements of the Big Data project, including the ONS Innovation Labs have been provided. Each pilot project uses a different big data source and has a different set of objectives which, collectively, will help ONS to understand the issues around accessing and handling big data as well as some of their potential applications for official statistics. Alongside the pilot projects a significant Big Data Project activity is stakeholder engagement and communication. This report has also summarised key engagement and communication activities.

[14] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics
[15] http://www.cros-portal.eu/news/scheveningen-memorandum-big-data-and-official-statistics-adopted-essc