Official

ONS Big Data Project – Progress report: Qtr 2 April to June 2014

Jane Naylor, Nigel Swier, Susan Williams ***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. A 12 month Big Data Project (which is to run throughout 2014) has been established to investigate the potential benefits alongside the challenges of using big data and associated technologies within official statistics. The key deliverable from this proof of concept project (due December 2014) will be an ONS strategy for big data. In taking forward this work ONS will uphold all relevant legal and ethical obligations.

Summary

This report provides an overview of progress on the ONS Big Data Project during the second quarter (April – June 2014) and builds on the work that was documented in the first quarter progress report[1]. An update is provided on the practical elements of the Big Data project, the four pilot projects that have been chosen covering both 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 and has a unique 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 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/guide-method/development-programmes/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 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 key component of the project is to include some practical applications of big data to both assess the role they may have within official statistics and help understand the methodological, technical and privacy issues that may arise when handling them.

Four pilot projects have been chosen covering both 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.

Although conducting research on only samples of data, some of these data can be too large and complex to process efficiently using standard ONS computers. There is therefore a requirement to use the ONS innovation labs, a private 'cloud' based environment, for analysis.

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

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

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 since they allow us to handle large and complex data sets and to test new 'big data' technologies.

The 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

[2] http://www.ons.gov.uk/ons/guide-method/development-programmes/the-ons-big-data-project/index.html
[3] http://www.openstack.org/

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. The main general programming and analysis packages currently being used are Python and R. A range of other open source tools are also being explored including Apache Spark, MongoDB and PostgreSQL. However, there are no particular constraints on which open source tools might be used in future.

There are restrictions on the data that can be accessed in the labs, within 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 meter information.

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 on-line. 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 (i.e. more basket items and/or products), and increased frequency.

Supermarket grocery prices have been identified as an initial area for investigation since 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.
- To develop methods for quality assuring scraped data.
- To compare scraped data with data collected using current methods, explore and investigate methodological issues with scraping prices from supermarket websites
- To establish whether price data could be sourced directly from commercial companies and if so, how this compares with data scraped by ONS prototypes.
- To evaluate the costs and benefits of these alternative approaches to collecting price data

Progress

Prototype web-scrapers have now been developed for three on-line supermarket chains (considered sufficient to act as a proof of concept) and are automatically collecting prices for the

[4] http://hadoop.apache.org/

selected basket items each day (approximately 6,500 price quotes). The main focus during this quarter has been on the development of quality assurance processes for the scraped price data.

The scrapers are fully automated and collect prices every morning just after 8:00am. A report is automatically generated providing product counts for each item category. These are then manually inspected to ensure that the target products are being collected. During Quarter 2 there was one instance where a supermarket changed a product category from “Bread rolls and wraps” to “Bread and cakes” resulting in complete failure for this category. The problem was easily fixed and the quality assurance systems now in place will ensure that any future problems can be fixed either the same day, or the next working day. There were other problems that affected the collection of these data such as broadband outages (in some cases resulting in no data being collected for 1 or 2 days). All of these issues have been useful learning experiences since they would need to be considered as part of any longer term strategy to collect web-scraped data.

Most product categories used by supermarkets do not correspond exactly to the item categories used by the CPI/RPI. Therefore, it is not usually possible to define search criteria that return the target products exactly. For example, the category “Red Wine, European” is not the usual way supermarkets categorise or describe a bottle of red wine. The general approach then is to provide broad criteria to ensure all target items are captured (e.g. “Red wine”) and to flag the non-target products as a separate process once the data has been collected. This requires manual intervention and is specific to each website. Classification work has now started and will provide useful data to explore whether inclusion of non-target items would make any difference to current price indices.

Future work

The scrapers will continue to run over the next quarter. Data quality checks will be extended to cover price quotes to check for outliers. Work will also continue on distinguishing in-scope CPI products from others to enable a set of sub-indices to be produced on a comparable basis i.e. to compare with existing methods.

These data will then be compared with data collected from existing methods, and if possible, with corresponding data supplied by PriceStats[5]. The aim is to establish whether price indices derived from these alternative sources are comparable with those collected using existing methods. Some initial investigations will be made into the methodological implications of collecting bulk price data through web-scraping. Collaborative research has been agreed with the University of Huddersfield to undertake a high level review of the potential opportunities and issues with using web scraped data to produce price indices. A meeting has also been arranged in July with MySupermarket.com (an aggregator of supermarket price data) to explore the possibility of collaboration.

[5] PriceStats is a U.S based company that scrapes prices on a global basis and produces daily indices. PriceStats has indicated a willingness to share their data for research purposes.

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 (e.g. 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 (ONS, 2011).

As these populations are more likely to use Twitter (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 publically available, the pilot is very conscious of the ethical issues around how this 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.

Research Objectives

The objectives are:

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

The basic approach is to collect and analyse geo-located tweets. The intention is to collect data continuously from the end of March 2014 through to the end of September 2014. Methodological development will continue during this period and data will be analysed and tracked on a regular basis to get early insights.

Progress

The focus during the early part of this quarter was to deploy a fully tested application that would collect, store and format the required data so that it could be read into an analytical package. A final tested version of the harvesting application was deployed on the 10 April.

A total of 38.9 million geo-located tweets were collected between 10 April and 30 June. The typical number of tweets collected per day was between 500,000 and 600,000. There were a total 12 outages during this period when data was not collected although this has resulted in a fair amount of missing data, it is unlikely this will affect this work as a proof of concept.

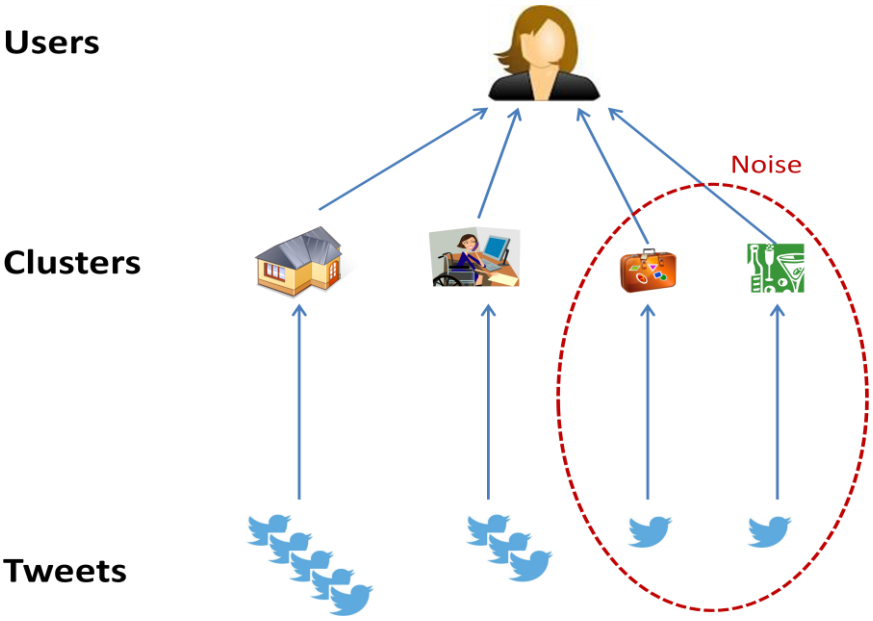
Good progress has been made on methods to create derived variables. This includes the conversion of latitude and longitude data to British National Grid coordinates (i.e. northings & eastings). This is necessary to ensure that clustering methods are accurate. Methods have also been developed for deriving date, day of week and time from the timestamp variable.

One of the main challenges of this pilot is how to make sense of the large amount of data that is being collected. Of greatest interest are clusters of tweets. These would indicate locations of some significance, such as a person's home, their place or work or study, or some other place where they spend a lot of time. Of lesser interest are locations with one-off or infrequent activity. These can be considered 'noise' points that we would want to remove from the analysis.

Good progress has been made in the development of a method for creating clusters for these data, a variation of DBSCAN, or "density-based spatial clustering of applications with noise" (Backlund et al, 2011). Once the clustering of all the data has been completed, the aim is to build a database based on a 3 level hierarchical relational model (Figure 1). The raw tweets and all time related data are stored at the lowest level. All spatial data is associated with the clusters. Clusters will be classified as either valid clusters or noise points, which can easily be filtered out. Clusters will also be classified by type using AddressBase[6] and nearest neighbour analysis. This should help determine whether a tweet was made at a residential or a commercial address, or some other location. It is hoped that this additional information will be useful in getting an understanding of the relationship between tweet locations and residence.

[6] This source contains every address in Great Britain, with coordinates, the address classification and postcode. See http://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-products.html

Figure 1: Twitter Pilot: Planned Analytical Data Model



Future work

The plan is to continue collecting data until at least the end of September as this will cover the period of student moves at the beginning of the academic year. Although the DBSCAN clustering algorithm has been tested on a small sample of data this now needs to be scaled up within the Innovation Labs making use of a more powerful processing environment.

Once the clustering methods have been tested, the next major challenge will be to look at classifying valid clusters using AddressBase. This source contains every address in Great Britain, with coordinates, the address classification and postcode. The coordinates on both datasets provides a mechanism for linking the classification and postcode to the clusters using nearest neighbour methods.

The next major step will be to build the analytical database. This will probably use PostgreSQL, which is an open source relational database, which should suit the planned data model.

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

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 20% energy efficiency and provision[8] is given for the roll-out of smart meters which requires member states to ensure that at least 80% of consumers have intelligent electricity metering systems by 2020.

The Department of Energy and Climate Change (DECC) has one of the most ambitious roll-out policies within the EU: to put electricity and gas smart meters in every home in England by 2020[9] with rollout 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[10].

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

The applications of most interest within the production of official statistics are:

1. Occupancy status of homes; low and constant electricity use over a period might indicate that a home is unoccupied. which could help survey fieldwork planning.

2. 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.

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 anonymous) level to understand patterns of energy usage. If the research is successful and suggests there is real value to be had in developing these small area estimates, the privacy and ethical issues surrounding the use of these data will need much greater consideration.

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

Research Objectives

The objectives are to:

- To understand the Big Data technical/methodological challenges of handling this type of data
- To assess some of the quality aspects of smartmeter 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 smartmeter profiles for determining occupancy status. Lesser 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 smartmeter data in specific applications
- Propose future use and further research within ONS with this type of data (final report)

Progress

This section provides an overview of progress on this pilot both through internal analysis and through a small research project commissioned by ONS and undertaken by Southampton University.

Southampton were commissioned to investigate the potential of using smart meter type data to identify household size/structure and the likelihood of occupancy during the day. A final draft report has been produced and the research shows that:

- Average electricity energy profile by half hour intervals might have some potential to identify households with school age children or households with occupants aged over 65 years although caution is required as the research is based on a small sample (approx 100 households).

- A method has been proposed that might be used to create a probability of a house being occupied at a specific half hour period

- Pre-processing of the data is very time consuming

- Smart meter profiles might be more suited to identifying household size rather than household structure

A separate 18 month ESRC funded project, to complete in 2015 and again conducted by the University of Southampton, will build on the research for ONS. ONS is represented on the project board.

Internally data has been sourced from the Irish Social Science Data Archive[11] and has been converted into a number of different formats for ease of processing. A variety of Big Data software and technologies were tested during the large scale merging/manipulation of these dataSamples of the data have been taken and preliminary analysis has begun to understand the data and to help identify methods of analysing it.

The focus for the internal research is to assess if these data can identify unoccupied households. It is considered that this might have potential to create efficiencies within a census or survey operation. It may also be theoretically possible to identify long term vacants/second homes or holiday homes due to prolonged low electricity use.

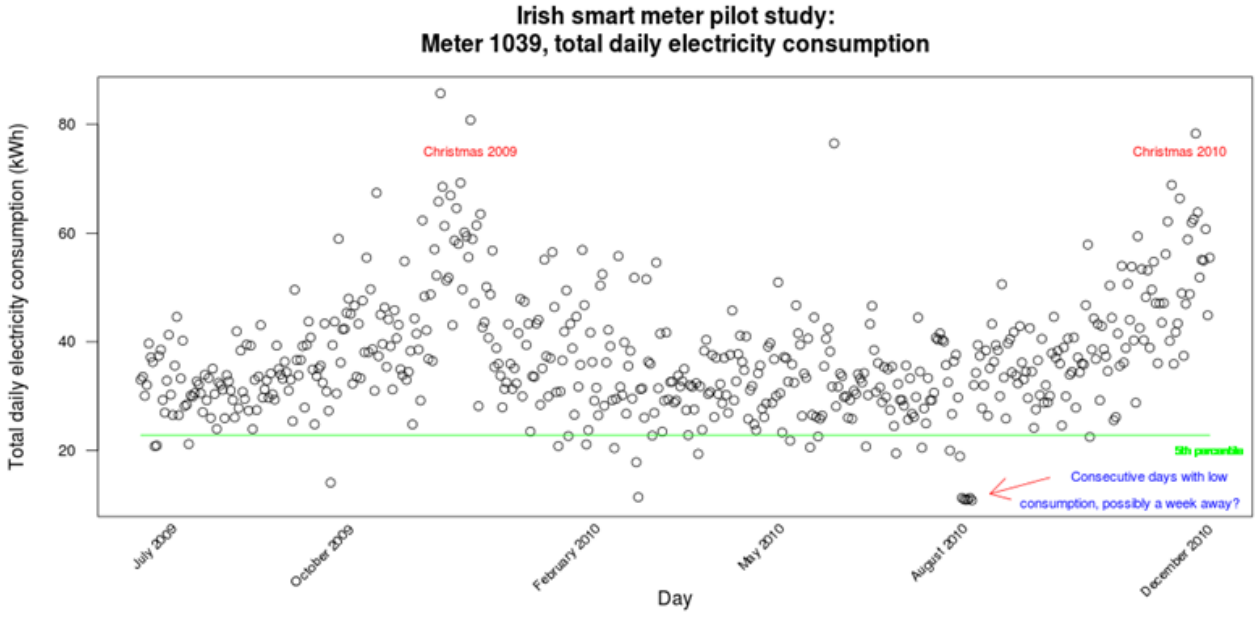


Figure 1: Total daily electricity consumption for meter 1039 between 14th July 2009 and 31st December 2010.

Figure 1 shows the total daily energy used by one household. A seasonal trend can clearly be seen where consumption is high in the winter (peaking at Christmas time) and generally lower during summer months. Low daily values might indicate that a house was unoccupied.

[11] From consumer behaviour trials – these data include 30 minute frequency electricity energy usage data on approximately four thousand homes during 2009-2010. A 6 monthly demographic survey was also conducted so it will be possible to identify some features of the home and the household inhabitants.

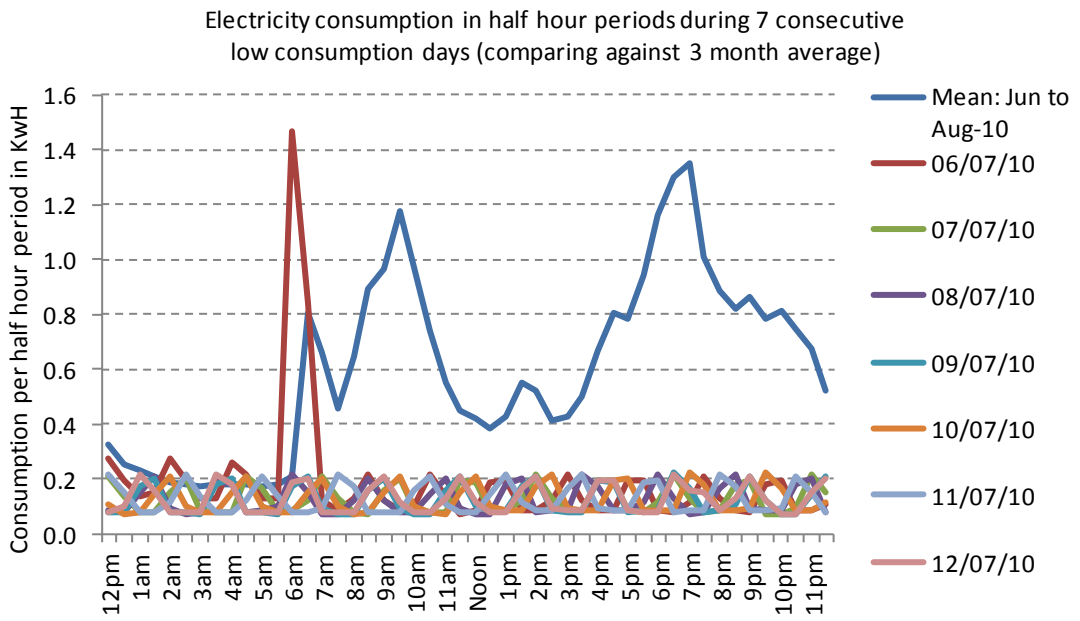


Figure 2 Electricity during 7 consecutive low consumption days for a single meter

Figure 2 shows a run of seven days when a smart meter had low daily energy values. Compared to the average energy profile for this house, it is clearly seen that there appears to be no evidence of occupancy, apart from a single spike in energy use at 6am to 6.30am on the first day. It is necessary to work at this individual (yet anonymous) household level to understand the patterns and develop generic algorithms but the ultimate aim is to develop methods to produce estimates at an aggregate level.

Future work

Over the next 3 months samples of smart meter data will continue to be used to develop ideas for higher level analysis, to be conducted on the full data when a suitable processing environment has been successfully created in the innovation labs. The focus of the research will be around which data can reliably be used to identify unoccupied households

6 Mobile Phone Pilot

Background

Location data generated through mobile phone usage is of key interest to statistical organisations as it has the potential to inform on various key aspects of population behaviour, with current research around the world 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[12] – a Eurostat funded feasibility study on the use of mobile positioning data for tourism statistics has generated research activity in this field 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 (94% of UK adults[13])
- There are relatively few service providers so any one provider might have sufficient coverage to produce reasonably representative insights of total population behaviour, and the effort required in approaching multiple companies around data access is reduced.
- The growth of Big Data technologies and methods are 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/phonecall. Of more interest is the use of "roaming" data which is passively generated from mobile phones when they are switched on and either move between masts or send out a location reading at intervals. It is speculated that roaming 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.

Research objectives

Objectives are to:

- To source aggregate data on travel patterns of workers from a main UK mobile phone provider with an emphasis on understanding the issues involved throughout the stakeholder engagement, negotiation and procurement stages of this "partnership" opportunity.
- To agree a method with the service provider and monitor the issues around the collaboration.
- To 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 this type of data (final report)

Progress

Due to the ethical and privacy concerns around Government departments accessing this data ONS presented the mobile phone pilot proposal to the GDS Privacy and Consumer Advisory Group and

[12] http://www.congress.is/11thtourismstatisticsforum/papers/Rein_Ahas.pdf
[13] Ofcom facts and figures communication report 2013

the ONS Beyond 2011 Privacy Advisory Group. Both groups, although wary of the acquisition of individual level data, were supportive of the use of aggregated data, especially as it is to be aggregated within the mobile phone company.

Background research on academic studies using mobile phone data has revealed a couple of papers with a focus on worker flows. Due to a wide range of working patterns, the method to identify home and work locations is non trivial. In studies, home location is modelled as the area where a mobile phone tends to be found at night time whilst work location is modelled as the location where mobile phones tend to be found mostly during the day (Mondays to Fridays). Challenges include:

- Workers who have more flexible arrangements, such as part-time, nighttime, shifts, work at multiple locations etc. may not be easily identified with such a broad approach.
- All methods currently identified have an inability to distinguish homeworkers,
- As it is understood that the nearest cell tower is the basis for detecting the location of a mobile device, problems will arise in rural areas especially, in that cell towers may have a reach of 2 km or more – thereby unable to detect small distance movement.
- Inability to segment “workers” by key demographics such as age and sex. The prevailing method appears to use information on contracted customers, who typically represent around 50% of all customers, as a proxy for pay as you go customers.
- Weighting up to population estimates is thought to involve the application of marketshares. Careful consideration will need to be given to the way that this is done

Future work

Work will continue to understand and address these challenges when specifying aggregate data that reflects ‘workers’ travel patterns.

ONS will seek meetings with the large mobile network operators to brief them on this research and enquire about their current interest in providing aggregated data for research. If only one company is interested then a single tender will be written; if more than one, then competitive procurement activity will be necessary. ONS will then work with the relevant mobile phone service provider to agree a suitable methodology and will acquire aggregated data for comparison with 2011 Census travel to work flows.

7 Stakeholder Engagement

A significant activity within the Big Data Project is stakeholder engagement and communication. Stakeholder engagement activities seek to achieve the following through communication and other means:

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

The following 9 groups of stakeholders have been identified for the project:

- Privacy groups
- International
- Academia
- Private Sector
- 'Big Data' Companies
- Technology providers
- Government
- ONS
- Users including the public

In the second quarter of the project key stakeholder groups have been international and government in particular to identify and contribute to collaborative opportunities building on engagement from Quarter 1. In addition there has been increased engagement with academics and 'big data' companies (to raise awareness of the project, identify common interests and collaborative opportunities and learn from external expertise), privacy groups (to raise awareness of the project and to get advice and steer) and the private sector (to acquire data).

In the third quarter of the project these activities will continue. In addition more attention will be given to internal engagement within ONS to gain support for emerging recommendations and next steps that will presented in September.

Key activities within these stakeholder groups are provided below:

- During this quarter the ONS Big Data Project has begun to engage with privacy groups. Presentations were made at the GDS Privacy and Consumer Advisory Group and the ONS Beyond 2011 Privacy Advisory Group, to provide an overview of the project, with particular focus on the pilot projects and the privacy challenges they bring. Both groups were positive about this early engagement and provided advice on handling, communications, policy in this area and future directions with the pilots. In particular the groups did not raise any serious concerns around our plans for the mobile phone pilot. The groups were comfortable with our research plans using anonymous smart meter data but felt that strong evidence of the benefits of the data would need to be provided to justify its use operationally. In addition

a concern was raised around the volume of Tweets we are collecting and whether this was within developer terms and conditions, this is currently being investigated with Twitter. During the next quarter the ONS Big Data team will build on the engagement with privacy groups that began this quarter. In particular the ONS Big Data policy will be circulated/presented for comment as well as keeping the groups informed of progress in general with the project and any particular privacy/ethical issues.

- The main activity within international stakeholder engagement has been participation in a 12 month UNECE international collaboration project focused on big data (http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics). A workshop was undertaken in March focused on finding agreement on the key challenges, issues and questions that the use of big data by National Statistical Organisations (NSOs) raises. The sprint was concluded at a 2 day workshop in Rome at the beginning of April with a paper being published as the key output. This paper sets out the key issues with using big data within statistical organisations together with next steps for the project and proposes a number of task teams which are to focus on specific issues related to the use of big data within official statistics.

  Members of the ONS Big Data Project are contributing to two of these task teams (selected due to the overlap of issues that need to be addressed in the UK context as well as internationally) which are due to deliver reports in the Autumn:
  - Task team focused on partnerships. The aim here is to study examples of NSOs working collaboratively with different partners in order to develop guidelines.
  - Sandbox task team. The Irish Centre for High End Computing (ICHEC) in conjunction with the Irish Central Statistics Office have volunteered to assist the project by hosting a big data 'sandbox', a web-accessible environment for the storage and analysis of large-scale datasets which can be used to test and evaluate big data technologies such as Hadoop. The aim of the task team is to identify collaborative projects that can be jointly undertaken between different NSOs.

- A European Statistical System (ESS) taskforce on big data and official statistics has also been established. The taskforce is focused on the Scheveningen Memorandum[14] (http://www.cros-portal.eu/news/scheveningen-memorandum-big-data-and-official-statistics-adopted-essc) and its implementation through an action plan and roadmap. A conference to launch the taskforce was held in Rome at the beginning of April. Members of the ONS team attended the conference and subsequent meetings and are contributing to action plans (that will ultimately feed through to the roadmap) focused on methodology and policy related to big data in the context of official statistics.

- The Economic and Social Research Council (ESRC) have a significant amount of funding to invest in a Big Data Network to help optimise data that is available for research. The first phase of the network focused on administrative data. The second phase supported the establishment of four Data Research Centres with a focus on Business and Local Government Data. The Data Research Centres will make data, routinely collected by

[14] A memorandum on "Big Data and Official Statistics" that was adopted by the European Statistical System Committee on 27 September 2013

business and local government organisations, accessible for academics. One of these centres is being led by University College London (UCL) and is focused on retail businesses. An initial meeting has been held with UCL to identify overlaps and opportunities between their centre and the ONS Big Data Project.

An initial and positive meeting has also 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. This would inform future ONS policy in this area as well as supporting the establishment of the ESRC Data Research Centres.

- An initial meeting has been held with the Royal Statistical Society (RSS) to raise their awareness of the ONS Big Data Project. Links between the ONS Big Data Project and the RSS will be strengthened during the next quarter of the project, in particular considering using RSS channels for communication and the possibility of an RSS meeting on big data.

- There is a need to engage more widely with academic institutions to understand what research activities are being undertaken, what expertise exists, the skills required to work within data science, the types of skills new graduates studying in this field will have and to potentially recruit/attract graduates to the ONS

  To start this process a review of academic institutions offering courses on big data/data science has been undertaken to understand their courses, research and the possibility of student placements/internships. These initial contacts will be followed up in the next quarter.

- Within the ONS Big Data Project there is a requirement to understand the big data expertise that exists within the commercial sector and how it might help achieve the project objectives. The plan is to hold initial meetings with a number of 'big data' companies to start to explore these issues. The approach to this engagement has been opportunistic - engaging with relevant companies that approach us or that contact is made with at conferences/events.

  A presentation on the ONS Big Data Project was given at the ‘Big Data Breakfast’ part of London Technology Week (http://www.wandisco.com/system/files/documentation/BigDataBreakfastReport.pdf). This provided an ideal opportunity to make big data companies aware of the ONS project and to start to make contacts which will be pursued in the next quarter.

- The ONS Big Data team continue to attend and contribute to the Cabinet Office Community of Interest meetings to support their Data Science Programme. This programme aims to understand how new analytical techniques and technologies can improve the way policy and operational services are developed across government. Over this quarter ONS have in particular shared ideas and experiences around the Research Innovation Labs. Specific discussions with the Cabinet Office team have also been held to identify overlaps/synergies between the two projects (where future liaison will be focused):
  - developing policy
  - the public perception/ethical issues

- Cabinet Office open policy approach
- skills and capability

- During the next quarter the following activities will be undertaken to increase awareness of the ONS Big Data Project, to share experiences and to coordinate work in this field across government:
  - Continue to publish articles/links/reports on the Government Statistical Service (GSS) Big Data Group
  - Presentations to be given at the Heads of Analysis Conference and the GSS Methodology Symposium
  - Continue to work closely with those responsible for the GSS Data Strategy over opportunities to share emerging findings from the ONS project with the wider GSS.

8 Conclusions

This report has provided an overview of progress on the ONS Big Data Project during the second quarter of the year. An introduction and update on the practical elements of the Big Data project has been provided including the ONS Innovation Labs. Each pilot project is to use a different big data source and has a unique 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 within official statistics. Alongside the pilot projects a significant activity within the Big Data Project will be stakeholder engagement and communication. This report has also summarised key engagement and communication activities.