Official

ONS Big Data Project – Progress report: Qtr 1 Jan to Mar 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 benefits alongside the challenges of using big data and associated technologies within official statistics. The key deliverable from the project (due December 2014) will be an ONS strategy for big data.

Summary

This report provides an overview of progress to date on the ONS Big Data Project. An introduction and update on the practical elements of the Big Data project is provided. 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 provides an initial identification of 9 groups of key stakeholders and summarises engagement and communication activities to date.

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 also to help understand the methodological and technical 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, smartmeter data and mobile phone positioning data. This report provides an overview of progress to date on the four pilot projects as well as a summary of progress around stakeholder engagement for the project, a key activity for the project.

2 Pilot Projects

The Big Data Project contains a practical component to help ONS understand the issues around accessing and handling big data as well as better identifying some of their potential applications within official statistics.

The approach taken is to conduct four pilot projects, each using a different big data source and having differing objectives which address various key issues around the adoption of big data within ONS.

The four pilots will use:

- Internet price data – to test ways of scraping prices for application within price statistics;
- Twitter messaging – for application of it's geolocation information in identifying movement patterns;
- Electricity smartmeter data – for it's potential for identifying household occupancy, household size or structure; and
- Mobile phone data for its potential in producing travel patterns.

In this first quarter, the technical focus has been to develop tools to scrape prices from the Internet and harvest Twitter messaging for future research. Smartmeter data has been sourced for analysis.

The mobile phone pilot provides an opportunity for ONS to test some of the issues involved with collaborating with a big data provider and wider stakeholder engagement has been held with the Government Digital Service.

The next quarter will see an increase in the level of methodological research, as more data is available for analysis. The technical research will move towards testing different ways of processing smartmeter data, whilst retaining awareness of the processing and storage issues involved with Twitter messaging.

An overview of each of the pilots, progress and future work is provided below.

2.1 Prices Pilot

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 price quotes (key input into these indicators) 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.

The first step is to develop and run prototype web-scrapers. These data will collect daily price quotes for a broad range of products for specified items in the CPI and RPI basket of goods and services. The intention is to run these scrapers for at least three months. Quality assurance processes will be set up to check that the scrapers keep collecting the data as expected.

These data will then be compared with data collected from existing methods, and if possible, with corresponding data supplied by PriceStats[1]. The aim is 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.

Good progress has been made on developing price scraping prototypes. Scrapers have now been developed for two of the four on-line supermarket chains and are automatically collecting prices for the selected basket items each day.

The current focus is on developing a suitable quality assurance process. A ‘dashboard’ is being developed that compares various dimensions of each daily extract with data already collected to help quickly identify unexpected changes.

[1] 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.

Once development is complete the pilot will then enter the main data collection phase. This will look to build up at least a three month time series of daily prices. Data analysis and assessment of method will run in parallel. In theory, web-scraping allows all price data to be collected rather than just a sample. This is a radical change and may require a very different approach for creating price indices. There will be many methodological challenges in particular around the continuity of series and discontinuity impacts that will need to be assessed at this stage.

2.2 Twitter Pilot

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).

In contrast, 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.

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.

Two main pieces of work have been completed during this period.

The first was an exercise to harvest a small set of geo-located tweets using the Twitter API and Python, process them using Apache Hadoop, store them in MongoDB (an open source NoSQL database) and to perform some basic analysis, such as the number of tweets by user.

The second exercise was focused on obtaining a larger volume of data, exploring it in more detail and to start developing appropriate methods. Just over 5.6 million tweets were collected during December 2013 and January 2014. This data is currently being used to develop and test data processing methods.

The current focus is on identifying a suitable clustering method for grouping user tweets in order to identify location of interest, such as home address.

An application for collecting geo-located tweets is now running continuously and it is planned to keep this running until at least the end of September 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

Office for National Statistics (ONS) 2011. "Internal Migration Estimates – Methodology" http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/internal-migration-methodology/index.html, Accessed on 09-06-14

2.3 Smartmeter Pilot

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)[2] 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[3] 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 smartmeters in every home in England by 2020[4] 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[5].

Smartmeter 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.

It is important to emphasise that ONS sets itself high standards on using and safeguarding personal data. Within this research, the data involved has already been made anonymous: all information that might identify the specific location of the household pertaining to each smartmeter has been removed.

So, there is a growing interest in using smartmeter data across statistical organisations globally. The applications of most interest within the production of official statistics are:

[2] http://ec.europa.eu/energy/efficiency/eed/eed_en.htm
[3] 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
[4] Wales and Northern Ireland have similar policies.
[5] Legislation still being devised

1. 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, the likelihood of a home being occupied on certain days and at certain times might be achieved, something of great relevance to 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.

Southampton University have been commissioned by ONS to conduct a small research project to investigate the potential of using smartmeter type data to identify household size/structure and the likelihood of occupancy during the day.

In addition ONS will source anonymised data from various trials of smartmeters which have been made available for research. The priority data are:

- *University of Loughborough's* data from energy usage monitoring trials which is archived by the UK Data Service[6] for future research use. These data link consumption at one minute intervals to a basic household occupancy and appliance ownership survey. This dataset derives from 22 dwellings observed over two years (2008-2009) and will provide an opportunity to test out various big data technologies and methods to understand the benefits and drawbacks of different approaches to processing.

- Data from consumer behaviour trials of smartmeters conducted in Ireland and held in the Irish Social Science Data Archive[7] – 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. These data will be used to perform more statistical based analyses such as occupancy, household size/structure and/or a data led cluster analysis.

- *Energy Demand Research Project*[8] data from trials of smartmeters conducted in Great Britain 2007-2010. DECC hope to publish the English data for research purposes in April 2014. These data represent around 20 thousand homes (with and without a smartmeter installed) but do not have associated demographic survey data thereby limiting its

[6] http://ukdataservice.ac.uk/
[7] http://www.ucd.ie/issda/data/commissionforenergyregulationcer/
[8] https://www.ofgem.gov.uk/gas/retail-market/metering/transition-smart-meters/energy-demand-research-project

usefulness for research. If time allows there may be an opportunity to investigate for similar patterns with the Irish data.

The coming months will see the completion of the Southampton University research and a focus on creating a suitable environment within the innovations labs for processing the smartmeter data acquisitions. Samples of these smartmeter data will be used to develop ideas for higher level analysis to be conducted when the processing environment has been successfully created.

It is anticipated that testing of various big data approaches to handling the smartmeter data as well as higher level methodological analysis will commence in summer 2014. Since these data are very big they offer the opportunity to really test out new big data technologies for processing and analysis.

2.4 Mobile Phone Pilot

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[9] – 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[10])
- 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 complexity of negotiation 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 Vodaphone 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.

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

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.

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. However progress has been delayed due to concerns over the public perception of government departments accessing these data (even though it will be a small sample and aggregated to avoid any disclosure). ONS are currently working with the Cabinet Office’s Government Digital Service (GDS) to manage these issues.

3 Stakeholder Engagement

3.1 Introduction

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 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
- 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
- Manage stakeholder expectations at various stages of the programme

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

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

3.2 Stakeholder engagement key activities: Qtr 1

In this first quarter of the project activities have focused on making contacts with stakeholders to raise awareness and interest in the project, to identify common areas of interest, to acquire data and to work on collaborative projects. Key stakeholder groups have been International and Government. Key activities within these stakeholder groups are provided below:

- The main activity within international stakeholder engagement has been participation in UNECE activities focused on big data. A project proposal was approved and funding provided for a 12 month UNECE international collaboration project on Big Data (http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics) to which the ONS Big Data team are contributing. Participation in the UNECE project has enabled discussions and therefore thinking around strategic issues in relation to big data that are common across NSIs. In addition it has provided an understanding of big data projects/experience and expertise within other NSIs and hence facilitated bilateral discussions on specific topics.

- A European Statistical System (ESS) taskforce on big data and official statistics has also been established and the ONS Big Data Project team are members. The taskforce is focused on the Scheveningen Memorandum[11] and its implementation through an action plan and roadmap.

- The ONS Big Data team have fed into a Cabinet Office led initiative around big data and data science. A virtual team has been formed with members from Cabinet Office, The Government Digital Service (GDS), & GO-Science (Business Innovation and Skills (BIS)), who are taking forward a programme of work designed to demonstrate the applicability of data science across government, and develop strategy for scaling this up in the future. ONS has contributed through attendance at Community of Interest meetings. This has been useful in identifying key contacts in other government departments. In addition some overlap in interests around policy and public perception have been identified that could lead to future collaborative work.

- In addition 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:

3.3 Stakeholder engagement key activities: Qtr 2

In the second quarter of the project the activities described above will continue. In addition more attention will be given to engagement with academia, the private sector and privacy groups as described below.

[11] http://www.cros-portal.eu/news/scheveningen-memorandum-big-data-and-official-statistics-adopted-essc

- During the next quarter there needs to be increased engagement with ESRC and academics in general to increase their interest in the ONS Big Data Project. The activities proposed will also increase their influence.

- A potential avenue for engagement with academics and a source of funding for future collaborative research could be the Eurostat Horizon 2020 scheme (http://www.cros-portal.eu/content/horizon-2020).

- Continued work on the pilot projects will require increased engagement with private sector companies primarily to acquire data. This will increase their interest and influence on the project. These activities may also facilitate thinking and discussions of different commercial models/partnerships with the private sector.

- During the next quarter of the project initial exploratory discussions will be held with the Information Commissioner's Office (ICO) and privacy groups to start to understand public concerns around the use of big data within official statistics, this should be seen as a priority.

4 Conclusions

This report has provided an overview of progress to date on the ONS Big Data Project. An introduction and update on the practical elements of the Big Data project has been provided. 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 provided an initial identification of 9 groups of key stakeholders and has summarised engagement and communication activities to date.