ONS Big Data Project – Progress report: Qtr 1 January to March 2015

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) ONS recognises the importance of understanding the impact that big data may have on our statistical processes and outputs. A Big Data Project has been established to investigate the benefits alongside the challenges of using big data and associated technologies within official statistics. The first phase of this project is now complete and a second phase is being launched to run until the end of March 2016. 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 final quarter of the first phase of the project (Jan – March 2015) and builds on the work that was documented in the previous 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/guide-method/development-programmes/the-ons-big-data-project/index.html

Contents

Background 1
Summary 1
1 Introduction 3
2 Innovation labs 3
3 Prices pilot 4
4 Twitter pilot 5
5 Smartmeter pilot 8
6 Mobile phone pilot 10
7 Stakeholder engagement 12
8 Conclusions 15

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 analysis.

This report briefly introduces the ONS innovation labs, then provides an overview of progress on the four pilot projects in the quarter (Jan – March 2015). In addition a summary of progress around stakeholder engagement is provided, an important activity for the project. This report builds on the work that was documented in the previous progress reports[2].

In 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 these 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. They are used for research, testing and evaluation purposes by authorised ONS employees. They are completely separate from the main ONS network and therefore provide a route for easily accessing open source tools without compromising ONS security. The innovation labs are a key enabler for the ONS Big Data pilot projects.

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

The labs consist of a number of high-specification desktop computers with some additional network storage. The hardware is configured using OpenStack[3] cloud 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 provides 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 meter information. The labs are also used for other ad-hoc projects undertaken by ONS staff. These include software evaluations, mapping open data and further exploration of sentiment analysis. There is continued exploration of technologies such as MongoDB, Hadoop and Spark.

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 (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.

Progress

Classifying web scraped grocery prices using machine learning techniques

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

A key challenge to using web scraped data is item mis-classification. The web scrapers collect all items under a supermarkets' item classification: for example the scraper will navigate to 'Whisky' on the supermarkets' website and then collect all the items under this classification. Unfortunately, the supermarkets frequently put other items under 'Whisky' such as Rum. This issue is common across all the supermarket categories, and requires a systematic and scalable solution to correct for this mis-classification.

Progress has been made in this phase on developing a solution to deal with item mis-categorisation. The techniques being used are from machine learning. Specifically, we are employing supervised classification algorithms such as Naive Bayes, Logistic Regression, Decision Trees, Random Forests and many more. All these algorithms learn from manually classified training data to predict an items' category based on its text description. Early models are classifying prices with a high degree of accuracy (99%) for a small subset of items.

Experimental index creation

Progress has also been made on index creation. Experimental price indices have been published[5] covering a range of different methods of index creation using the web scraped data, and comparing them to indices created using official (Consumer Price Index) data. There are many possible ways of creating indices hence there will be continued research to investigate the optimal approach.

Future work

The Prices pilot will continue within the second phase of the Big Data Project and the priority will be improving the robustness of the web prices collection. This includes moving the scraper to a more robust environment, improving classification/cleaning and improvements to the web scraper itself. The internal web scrapers will be expanded to cover all grocery items. The cost efficiency of a robust in-house web scraper will also be compared to the costs of purchasing web scraped data from companies such as mysupermarket.com.

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. These data can then be used to track mobility patterns.

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

[5] http://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices/experimental-consumer-price-indices-using-web-scraped-data/index.html

people, especially young men, are often slow to re-register when they move. These populations are more likely to use Twitter as opposed to other population groups.

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 are used and will therefore 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 mobility 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 this data 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

Progress during this period has been focused on five main activities:

- Quality assuring and preparing a clean data set prior to running the clustering and classification algorithms.
- Rerunning the algorithms and compiling the final analytical database.
- Migration of the final analytical database to MongoDB
- Re-running of the analyses for the final report
- Preparation of the final written report

Data has been combined from two different sources for different time periods:

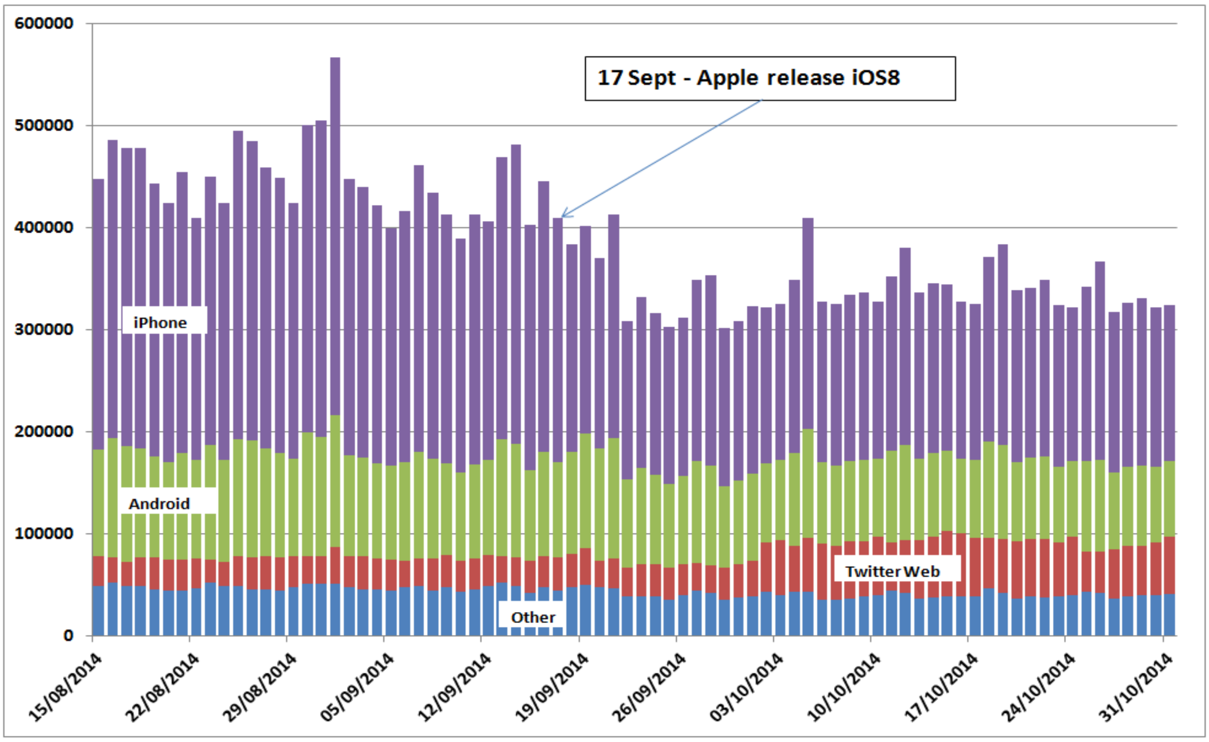
- Data collected directly by the pilot from the Twitter streaming API
- A point-in-time data extract purchased from GNIP/Twitter

As discussed in the previous progress reports, compliance issues with Twitter developer rules required the pilot to move from data collection using the Twitter API to purchasing the data directly. With GNIP’s consent, the pilot combined these data. Some of the differences between each source have taken some time to fully understand and resolve. However, the analytical work is now complete.

There is one new piece of analysis that is particularly worth mentioning as it puts into focus the issue of continuity of big data sources. In the second half of September 2014 there was a general drop in the daily volume of geo-located tweets from about 400,000 per day to about 300,000. Analysis by device type shows that this can almost entirely be explained by a sharp fall in the proportion of tweets from iPhone users (Figure 1). This can be explained by Apple's release of the iOS8 operating system on 17 September 2014, which incorporated more flexible options for managing privacy with respect to location.

This analysis raises some fundamental questions around the use of big data for statistical purposes. ONS is able exert a great detail of control around surveys to minimise these kind of discontinuities. While there is some risk around changes to administrative systems, these can usually be anticipated and managed. Sources like Twitter however are far less constant and their composition is changing against a backdrop of constant social, commercial and technological change. This means that analysis of change can never be taken at face value and that confounding factors should always be considered.

Figure 1: Daily Geo-located Tweet Volumes by Device Type, Great Britain, 15 August 2014 to 31 October 2014



Future work

In addition to completing phase 1 of this pilot, some very good progress has been already been made on the next phase of research. This is looking at the feasibility of deriving socio-demographic characteristics of Twitter users. This research aims to tackle some of the issues identified in the initial research around the representativeness of Twitter data.

5 Smartmeter 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)[6] is a common framework of measures for the promotion of energy efficiency within the EU. It supports the EU's 2020 headline target of a 20 per cent reduction in energy consumption, and its provision[7] for the roll-out of smart meters requires member states to ensure that at least 80 per cent of consumers have such 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[8] 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[9].

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

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.

2. 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.

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.

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. Initial research proposals have been discussed with the GDS Privacy and Consumer Advisory Group and the ONS Beyond 2011 Privacy Advisory Group. If the research is successful and suggests there is real value to be had in developing these small area

[6] http://ec.europa.eu/energy/efficiency/eed/eed_en.htm

[7] This provision relates to another EU Directive on smart meter rollout (2009) which required a full cost/benefit analysis be performed prior to commencing roll-out

[8] Wales and Northern Ireland have similar policies.

[9] Legislation still being devised

estimates, the privacy and ethical issues surrounding the use of these 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 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. 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 smartmeter 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 electricity meters has previously been sourced from the Irish Social Science Data Archive and loaded into the innovation labs. Around 4,000 residential homes are included in these data, and anonymised samples were taken to start preliminary analysis on understanding and handling the data.

Research has focussed on using logistic regression to model the likelihood of a household being unoccupied on a particular day. The pros and cons of using such a model have also been examined. For example, although the model may be more accurate it may also be more computationally intensive than a simpler method.

Research has started into clustering the Irish smart meter trial data. Clustering establishes whether there are groups of households with similar energy profiles. Different groups may be explained by using the survey data provided. Once a suitable clustering method is identified using the Irish trial data, it will be applied to data from a trial of rolling out smart-type meters conducted in Great Britain 2007-2010[10] trial data to understand if similar clusters exist in both datasets, and if not, why this is.

In order to cluster similar energy profiles, variables need to be chosen which in combination can describe the ‘shape’ of the energy profile. This shape can be the average daily energy profile over the trial period, or can take in other variables such as the percentage of weekly energy that is used over a weekend. The variables in the clustering algorithm need to be as independent from each other as possible and capture as much of the variance in the energy profiles as possible.

After an initial literature review two methods[11,12], both using k-means clustering, have been tested on the data. The Knime method performed better and involved calculating 27 different variables,

[10] This data represents around 20 thousand homes (with and without a smart-type meter installed) but do not have associated demographic survey information thereby limiting its usefulness for research.
[11] Cluster Analysis of Smart Metering Data, Research Center for Information Technology, Germany
[12] Big Data, Smart Energy, and Predictive Analytics, Knime

including mean daily consumption, the percentage of weekly consumption used on different days of the week and the percentage of daily consumption used at different times of day.

The results for each meter were then grouped into seven distinct clusters using the k-means clustering algorithm: 65% of the variance in the energy profiles can be explained by the clustering. Repeating the clustering algorithm with varying input or seeding values has shown that the seven clusters are stable.

Future work

Over the next three months the final pilot report will be published.

Due to political sensitivity, long time scales and uncertainty on access to data from smart meters it has been decided to end the research on smart-type meter research. The following research will be published as short papers:

- Using machine learning methods such as logistic regression to identify households unoccupied for a whole day
- Using cluster analysis to see if similar patterns exist in the GB data as in the Irish data. If this is so, then the Irish data's survey information may be useful for continuing analysis of these data.

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[13] – 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 (94 per cent of UK adults[14])

[13] http://www.congress.is/11thtourismstatisticsforum/papers/Rein_Ahas.pdf

[14] Ofcom facts and figures communication report 2013

- 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/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:

- Source aggregate data from a main UK mobile phone provider on travel patterns of workers. The emphasis here is on understanding the issues involved throughout the stakeholder engagement, negotiation and procurement stages of this 'partnership' opportunity.
- Agree a method with the service provider and monitor the issues around the collaboration.
- Compare the aggregated mobile phone data with 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 in ONS for this type of data (final report).

Progress

ONS continues to engage with DfT and other transport bodies in this quarter to gather intelligence on the use and acquisition of mobile phone data. A positioning paper is being prepared, detailing the engagement held with transport bodies and mobile phone network operators, the potential use of data and the issues involved. This paper will form the pilot report on mobile phone data.

The conclusion being drawn is that the acquisition of mobile phone data needs to be better coordinated. This issue will be taken forward by ONS through the Government Data Science Partnership in the next phase of the project.

Additional internal research

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

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 made between the peak travel time of 7am and 10am. ONS used this 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

The mobile phone pilot report or positioning paper will be finalised and published. The Oyster card research will be finalised and published as a short paper.

It is further proposed that research using mobile phone data for transport flows will be taken forward in the second phase of the project through the Government Data Science Partnership (GDSP) with ONS leading the work.

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 final quarter of the first phase of project we held a stakeholder event for the media, privacy groups, academia and other Government departments. In addition during this quarter 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 the second phase of the project these activities will continue, a review of stakeholder engagement and communication activities will be undertaken in order to identify new stakeholders, reprioritise engagement and develop a communications plan.

Key activities for specific stakeholder groups are summarised below:

- During March a briefing was held on the ONS Big Data Project. Attendees (from the media, privacy groups, academia and other Government departments) were provided with an overview of the work that has been undertaken during the first phase of the project. The event also included non-ONS speakers (Will Page (Spotify), Marion Oswald (University of Winchester) and Siobhan Carey (Business Innovation and Skills) who provided different insights on the topic. The objective of the event was to keep those with a vested interest in ONS and its work well connected to developments that could, in future, play a very important part of decision making. This objective was met and feedback was very positive.

- The ONS Big Data team have continued to engage and collaborate with other Government departments around big data/data science. In particular we have begun to work more closely with colleagues in the Cabinet Office, Government Digital Service and Go-Science through the newly formed 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 GDSP work plan is being developed with ONS taking the lead on activities around web scraping and mobile phone data but we will also contribute to cross-cutting work particularly on technologies including sharing experiences of the use of the Innovation Lab.

- We have engaged 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 smartmeter pilot
- Meetings held with Bank of England and Defence Science and Technology Laboratory to discuss big data/data science projects and common areas of interest

- 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. The ONS Big Data team met with the ESRC team and contacts from their Business and Local Government Data research centres in March. A number of collaborative opportunities were identified.

- We have continued to engage 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, the research activities that are being undertaken and to raise the profile of our project. A guest lecture on the ONS Big Data Project was given to MSc students at Royal Holloway. In addition we have recruited a student who will undertake a 12 month placement within the team starting summer 2015.

- Members of the ONS Big Data team are contributing to the European Statistical System (ESS) taskforce on big data and official statistics which is focused on the Scheveningen Memorandum[15] and its implementation through an action plan and roadmap. An ESS Big Data Project has been scoped to implement this roadmap. A key activity within this project will be practical hands-on pilot work focused on specific big data sources and their impact on official statistics. The pilots will be undertaken collaboratively across ESS members. During this quarter the taskforce has developed criteria that will be used to select big data sources that will be the focus of the pilot work.

- Members of the ONS Big Data team attended the ‘New Techniques and Technologies’ official statistics conference in Brussels in March. Papers were presented on the ONS project overall, the Prices pilot and the Smartmeter pilot. This provided 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 National Statistics Organisations.

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

8 Conclusions

This report has provided an overview of progress on the ONS Big Data Project during the final quarter of phase one of the project. 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. This report has also summarised key engagement and communication activities.