Estimating population mobility using big data sources - the benefits and the challenges

Jane Naylor (corresponding author), Nigel Swier, Susan Williams
Digital Services, Technology and Methodology Directorate
Office for National Statistics, Segensworth Road, Fareham, Hampshire, UK, PO15 5RR
Contact: 01329 444976, jane.naylor@ons.gov.uk

Abstract. The digital age is generating vast amounts of data in an expanding variety of formats. This has given rise to the concept of "big data"– large, often unstructured datasets, often available in real time, that require new methods and technologies to handle and analyse. As such it is more important than ever for official statistics to adapt to the changing data landscape. Big data and data science could impact throughout the statistical value chain, potentially delivering significant efficiency gains and quality improvements.

One particular application that has been investigated by a number of National Statistics Organisations is population size and mobility. What potential benefits do big data sources offer over the traditional data sources and approaches (i.e. surveys, census, registers and admin data) in producing estimates of the population and migration?

This paper will illustrate both potential opportunities (improved timeliness, frequency, relevance and reduced costs and respondent burden) but also the challenges (statistical, ethical, commercial) of estimating population mobility using big data sources through three case studies from the UK Office for National Statistics (ONS) - focused on mobile phone data, geolocated social media data and Google trends.

Current and proposed ONS research will also put forward some approaches to overcome the identified challenges and hence realise the benefits of these new data sources for official estimates of the population.

Keywords: big data, population, migration, mobile phone data, Twitter, Google trends

1. Introduction

The digital age is generating vast amounts of data in an expanding variety of formats. This has given rise to the concept of "big data"– large, often unstructured datasets, often available in real time, that require new methods and technologies to handle and analyse, examples being data from the internet, social media, sensors and mobile phones. The potential value of these new data sources for official statistics (traditionally produced using survey, Census or administrative data) has been recognised. In particular The European Statistical System Committee's Scheveningen Memorandum[1] encourages the European Statistical System …. *'to effectively examine the potential of Big Data Sources'*. Many 'big data' research initiatives have been established by National Statistical Organisations (NSOs) across the world. The Office for National Statistics (ONS) in the UK has established a Big Data team[2] to investigate the advantages and the challenges of using big data, and to develop a longer term strategy for using big data and data science in official statistics.

One particular application of big data sources and data science techniques that has been investigated by a number of NSOs is population mobility. At present the UN Global Working Group on Big Data for Official Statistics[3] has a task team focused on mobile phone data, one application being population and migration statistics. In addition the European Statistical System network (ESSnet) Big Data project[4] includes a pilot project to investigate how a combination of big data sources and existing official statistical data can be used to improve current statistics and create new statistics. One of the statistical domains that the pilot is focused on is population statistics.

Traditionally official population and migration statistics are produced using a combination of national surveys, Census data, registers and administrative data. Outputs such as population estimates, estimates of international and internal migration are released by NSOs for different geographical areas on an annual basis. These statistics have a wide range of uses; they are used by central and local government for planning and monitoring policy and service delivery; resource allocation; and managing the economy. Additionally they are used by commercial organisations and academia as well as being of interest to the general public. How could the needs of these users be better met by the use of big data sources over the traditional data sources in producing estimates of the population and migration? Could these new data sources improve the accuracy, timeliness, frequency or relevance of current estimates? Could they allow NSOs to produce entirely new outputs or intelligence around population mobility? Could they reduce the reliance on surveys and Censuses, especially in countries without a population register, thus reducing respondent burden?

This paper will explore these potential benefits but also the challenges of estimating population mobility using big data sources through three case studies from the ONS Big Data team. The next three sections of the paper provide an overview of the cases studies focused on three different data sources; geolocated social media data (in particular Twitter), mobile phone data and Google trends. The final sections of the paper draw together the potential benefits and challenges of using big data sources to estimate population and migration as illustrated through the case studies and make recommendations for addressing the identified challenges in order to realise these benefits.

---

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

[2] http://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theonsbigdataproject

[3] http://unstats.un.org/unsd/bigdata/

[4] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

2. Geolocated Twitter data

Twitter is a popular micro-blogging platform where users post short messages, or “tweets”, with a limit of 140 characters. As of 2016, there are around 313 million regular twitter users globally. Large volumes of these messages and accompanying metadata may contain a range of insights. Users tweeting from a smartphone or other device providing location services may choose to provide a precise GPS location. These are referred to as geolocated tweets. Although less than 2 per cent of tweets are geolocated, the volumes of data are still considerable with hundreds of thousands of such tweets being sent every day within Great Britain.

The potential of geolocated activity traces from Twitter to provide insights into population mobility was the focus of a pilot project undertaken by the ONS Big Data team in 2014/15 [6]. The aim of the pilot was to establish whether it is possible to use geolocated activity traces from Twitter to infer a user’s residence and thus analyse mobility patterns. This would provide new insights into the population and how different groups move around the country and potentially help to better understand and validate official population estimates.

Data were collected on all geolocated tweets sent within Great Britain over a seven month period (1 April to 31 October 2014). This involved collecting data through a combination of real-time collection through the Twitter API and procurement of a bulk point in-time extract. Although individual level data were required for this analysis the ultimate interest and output was solely concerned with aggregate patterns and privacy rights were respected throughout. A density based spatial clustering algorithm with noise called DBSCAN [3] was used for clustering individual geolocated activity traces. AddressBase (the definitive source of address information in Great Britain) was used to find the geographically nearest address point to each cluster and to classify the clusters by type (i.e. residential, commercial, or other). For each individual user the residential cluster with the highest number of tweets (referred to as the dominant residential cluster) was assumed to be the location of usual residence. There were about 340,000 Twitter users for whom there were sufficient data to infer a location of residence for a period of at least one month during the seven month period. The activity traces for each user were broken down into months and then dominant residential clusters were identified for each month. When the dominant cluster from one month to the next was in a different local authority (these are the 404 local government districts in the UK), this was inferred as a mobility flow between local authorities or an internal migration move.

These net flows for each local authority were compared with the proportion of students in the population (based on 2011 Census data) and a distinct signal was identified that follows the cycle of the academic year. For example, in June there is a net outflow from student areas coinciding with the end of studies. Then in September and October, there is a net inflow back into these areas. Figure 1 illustrates this through a month on month comparison of flows for the local authorities with the highest proportion of students. The pattern is strongest for larger regional centres (e.g. Sheffield, Newcastle, Manchester). The pattern for Oxford and Cambridge is weaker despite having a very high proportion of students in the population. The reason for this is unclear, but one possibility is that it relates to international students who have a home address outside of the Great Britain and any geolocated Twitter activity outside of Great Britain would not have been picked up in this study. This pattern of internal migration moves cannot be detected from existing sources and so could be used as a supplementary source of intelligence on the movement of student populations. The pilot demonstrated that it is possible to gain insights into population mobility, in particular internal migration from geolocated Twitter data.

However, the geolocated Twitter data have a number of data quality and access issues that would need to be addressed before this research could be made operational. Firstly the data

are uneven across the user base. Half of all geolocated tweets were made by just 4 per cent of users while 17 per cent of users only sent one tweet. Only 46 per cent of all users had sufficient detail to infer a location of residence. In addition, the median time span between a user's first and last tweet was 47 days. This suggests that many users go through a phase of sending geolocated tweets but do not continue doing so. Thus, Twitter may have limited value for monitoring longitudinal change over periods of more than a few months.

In general social media data are unstable and may be affected by unexpected technological and behavioural changes. A 25 per cent drop in the volume of geolocated tweets during September 2014 was detected during preliminary data analysis and exploration. Investigations into the reason for this decline in volumes identified a link with the release of the iPhone iOS8 operating system. This release included changes to how privacy and location are managed. An analysis of tweets by device type showed that the decline was almost entirely explained by a decline in volumes from iPhone devices (Figure 2). This suggests that many iPhone users took the opportunity to exert greater control over their location settings which subsequently impacted the overall volume of geolocated tweets. This illustrates how the collection of data from social media can be impacted by a combination of technological change (including those of third parties) and the behavioural response of users. This has implications for time series analysis and illustrates why caution is needed when using social media data to estimate patterns that may inform decision-making.

Another quality issue related to the use of geolocated Twitter data relates to bias. Although these analyses can provide new insights into population and mobility, they are based on un-weighted counts and are not estimates. This is an important consideration as Twitter users are not representative of the general population. One possibility for producing estimates could be to infer socio-demographic characteristics of Twitter users and then calibrate to other sources, such as the mid-year population estimates. Another approach might be to use a benchmarking survey to measure rates of Twitter usage across the population. These avenues of research are already being taken forward by the ONS Big Data team.

As well as data quality issues there are also data access challenges when using geolocated Twitter data within official statistics. This pilot started by collecting data through the public Twitter API. Although it is straightforward to collect the target data using this approach, collecting data at this scale falls outside Twitter's API terms and conditions. Thus, if this research were to be made operational, then data access would need to be negotiated with Twitter.

There are also important ethical considerations when using these data, especially when dealing with precise location data. Twitter is designed to be public facing and in addition users must agree to certain conditions about how their data are used. The data used in this project is in the public domain but do Twitter users realise that the content of their tweet and also their location (if they have geolocation enabled) is all in the public domain? The fact that a large number of iPhone users chose to exert greater control over their privacy settings following the iOS8 release raises the question as to whether these users were fully aware of what was happening to their data prior to its release. The question of informed consent related to this project was discussed by the ONS' National Statistician's Data Ethics Advisory Committee[5] in January 2016. The difficulty of this ethical consideration was reflected by the committee not reaching a consensus. The advice was that the research should continue at this feasibility stage but any further or new research using Twitter data should be considered again by the committee in order to balance the ethical risks with the public benefit.

Overall this pilot demonstrated that it is possible to gain insights into population mobility, in particular internal migration from geolocated Twitter data. These insights could be used to

[5] https://www.statisticsauthority.gov.uk/wp-content/uploads/2016/04/NSDEC-270116.pdf

compliment or to quality assure traditional estimates of population mobility and migration. A key benefit is that Twitter data are available in near real-time and more frequently than traditional data sources used to produce these types of official statistics. However, the pilot also illustrated the challenges around data quality, data access, control and ethics that would need to be resolved before using these data within the production of official statistics.

3. Mobile phone data

Statistics collected by the UK communications regulator Ofcom show that at the end of 2014 there were just over 89.9 million mobile phone subscriptions in the UK with 93% of adults owning such a device. When mobile phones are switched on they generate digital information which is transmitted to the mobile network operator (MNO). This information includes details of any calls or texts made or received and also passive location updates. The growth of smart phones and devices capable of connecting to the Internet has further increased the data generated. This coupled with new generation mobile technologies, large increases in computer power and the availability of analytical software means that MNOs are increasingly interested in using their data holdings to create data products of commercial value.

As mobile phones are carried by their users, there is great interest in the use of mobile phone location data to understand population densities and population mobility. The attraction of mobile phone data are their availability in close to real time, for small areas and for a large proportion of the population. As a result the data offer the basis for developing a much richer understanding of population dynamics.

Around the world, NSOs are interested in the potential of mobile phone location data for the production of official statistics on the population. Such data might be used to help quality assure, enhance or even replace official estimates, provide more timely indicators of population change or support the development of new measures of population densities.

Over the past year, the ONS Big Data team has conducted research into the potential of using mobile phone location data to estimate flows of workers from home to work locations. These are currently produced as outputs from the decennial Census. This investigation has involved engagement with the three large UK MNOs[6] and public transport bodies who have a similar interest in using the data to estimate transport flows.

A literature review focused on international research using mobile phone data, with a focus on its relevance to official statistics has been published [8]. The strengths and weaknesses of mobile phone data were identified concluding that the areas of most potential for NSOs are to use mobile phone data for population densities and mobility, i.e. commuting flows. In addition a body of intelligence has been gathered by the Big Data team around the potential benefits but also the challenges of acquiring and using mobile phone data within official statistics.

Two types of geo-location data are generated from mobile telephony: active events which occur when making or taking calls or texting, and passive location updates which take place when a mobile phone connects to different cell tower or is periodically ‘pinged’ by the MNO to ascertain its whereabouts. The MNOs use both forms of location data to produce a variety of data products and services, some of which can be used to estimate population density and mobility.

The ONS produce population estimates on an annual basis. As mobile phones are owned by such a high proportion of the population the number of connections to a single cell tower could

[6] Everything Everywhere (Orange, T-mobile and EE), Vodafone and Telefonica (O2)

theoretically be correlated with the true number of people in the cell area[7] associated with that cell tower. To estimate population totals, weighting adjustments would be needed to correct for the market share held by the MNO and for other factors affecting bias. Research has shown that counts of mobile phones connected to cell towers may provide a credible approach to distributing national population totals to sub regions [1]. Mobile phone data also present an opportunity to produce more timely estimates of populations according to new and flexible definitions. This might include estimates by time of day, day of week or by season. Individual level geolocation data allow analysis of movement patterns for individuals which when aggregated may provide a basis for understanding population movements. ONS is interested in understanding the relationship of such estimates with the home to work flows produced from the Census. Mobile phone data might facilitate more frequent and timely outputs of such data than is currently achieved with the decennial census.

However, there are uncertainties around key methodological and data quality issues associated with using these products and services within official statistics. This arises from a lack of transparency on the methods used by MNO's to develop their data products. Each organisation has developed their own modelling techniques and is reluctant to share their knowledge and expertise because of commercial sensitivities. This is compounded by a number of statistical issues associated with the data, summarised below:

- Definitions - official statistics are based on clear definitions and population statistics tend to refer to the usual resident population, where a usual resident is a person who has lived, or intends to live, in the UK for 12 months or more. Mobile phone location data capture total population flows (for mobile phones, which do not necessarily represent a single person), not just those made by the usual resident population.

- Spatial resolution - there is a limit on the spatial resolution possible in terms of identifying the origins and destinations of journeys with mobile phone data. This is related to the location of cell towers and means that the data are not able to detect journeys that do not move between cell towers. This will be more marked in rural areas where cell towers may be kilometres apart.

- Modelling home and work location - home is generally assumed to be where the mobile is located during the night or when switched on first thing in the morning. Work location is more difficult to model and tends to be set to the location where mobile phones are found during the day (Mondays to Fridays). Methods are based on the detection of regular and repeated journeys made by the mobile. Clearly, workers who have more flexible arrangements, such as those on part-time or zero hours contracts, who work during the night, at weekends, in shifts, or at multiple locations etc., may not be easily identified by such an approach.

- Demographics - population flows are sometimes required to be segmented by key demographics such as age and sex. MNOs use the information held on contracted subscribers directly, although it is known that children's mobiles may be incorrectly categorised as parents are commonly responsible for the contracts. Contracted subscribers typically represent around 60% of the total subscriber base although this share is increasing due to competitive pricing of contracts. MNO do not have demographics for Pay As You Go (PAYG) customers. The prevailing method is to use the age/sex distribution from contracted subscribers as a proxy. This assumption may not hold.

[7] Cell areas vary in size. In urban areas, where cell towers are densely situated, they may have a range of 300 to 400 metres. In rural areas, cell tower density is very sparse and a cell may have a range of 5km or more.

- Bias – mobile phone data are biased, they do not represent the population of interest and the methods used to weight mobile phone data to population totals are not transparent. In the simple case a MNO may purchase regional market share information from data brokers and used it to scale results up to population totals. A more elegant weighting model which takes into account regional variation of market share is used by some MNOs. This method first infers the customer's home location. For a given area, the MNO works out the total number of customers it believes reside there and then uses official population figures to calculate the proportion of the residential population deemed to be customers. The inverse is used as the weight for all customers in that area.

- Missing data - the methodology used by MNOs implicitly assumes that mobile phones are switched on all the time. The scale of missing data (when a mobile is switched off) is unknown and will have implications for the weighting model. The problem of missing data is complicated by the opt-out arrangements for subscribers allowed by some MNOs in order to strengthen their ethical standards. The level of opt-out needs to be known and factored into any weighting.

Another challenge around using mobile phone data within official statistics is that the use of these data may raise privacy concerns around access to, or use of, personal data. These concerns may include informed consent as well as potential disclosure risk and identification of individuals arising from unique patterns of movement. At present ONS are only interested in accessing aggregated commuting flows not the underlying individual data. However, public opinion is still important and people must have confidence in the assurances and safeguards put in place by both ONS and MNOs.

This initial engagement and intelligence gathering exercise has identified some potential benefits but also some challenges with using mobile phone data within official statistics. In order to take forward this research and investigate these challenges further the ONS is planning to obtain access to commuting flows derived from mobile phone data. The aim is to obtain aggregated commuting flows between middle layer super output areas[8] for a sample of local authorities. These will be compared to the flows produced from the 2011 Census to understand the quality issues and to develop recommendations on the use of the data for statistical and research purposes.

A procurement exercise has been launched with the aim of obtaining the commuting flows at marginal cost, i.e. only paying for the service the MNO provides in processing, modelling and aggregating the data. This procurement exercise has been a significant learning exercise for engaging with commercial organisations in this way. This has involved addressing issues associated with payment for data/services, benefits to MNOs in sharing their data and potential opportunities for partnering and terms and conditions of the sharing agreements.

Alongside the work to obtain and analyse commuting flows derived from mobile phone data the ONS will also seek to influence the cross Government approach to using mobile phone data for statistical purposes. Procurement of mobile phone data for use in producing statistical products by public sector organisations has been largely uncoordinated. ONS will be engaging and encouraging collaboration with government bodies including the Government Statistical Service, MNOs and other organisations who may wish to acquire such data.

.

4. Google Trends

[8] A statistical geography used in the UK with an average population of 7,200

Search engines are widely used by people to navigate their way around the Internet. Within the UK, the most popular search engine is Google. By retaining such search queries, Google has built a datasource which allows research into Internet search behaviour over time and within regions of interest. Google has made some of it publically available in an application called Google Trends[9]. This application uses a sample of search queries and produces what is termed a "search volume index (SVI)" reaching back to 2004. Users can restrict research to search queries originating in their desired country or region and time period: a weekly SVI will usually be produced based on any search term of interest. The data made available are extremely timely with a lag of only 2-3 days, a feature which has generated interest in the ability of search queries to predict trends before official figures are produced, described as "nowcasting".

During 2014 the ONS Big Data team undertook some preliminary research to see if search queries were able to inform on the size and dispersal of populations in England who originated from countries that joined the European Union (EU) following its expansion in 2004 and again in 2007[10].

The UK, along with Republic of Ireland and Sweden operated an 'open door' policy in 2004 which contributed to a large number of migrants from the EU8 countries, especially from Poland. In the period October to December 2010 official estimates indicated that 571 thousand[11] Polish nationals resided in England. This compares to an estimate of 59 thousand[12] immediately prior to EU expansion. As the largest migrant population following the 2004 EU expansion in England, research initially focussed on the ability of search queries to identify the population size of Polish nationals. Under the assumption that newly arrived Polish nationals would be conducting their Internet searches primarily in Polish, it was proposed that searches containing the term "polski" might be both sufficiently specific to this population and used frequently enough to generate volumes large enough for Google Trends to report on.

A weekly SVI series was generated for Google searches containing the term "polski". The series started in Jan 2004, which is the earliest time Google Trends can report for, up to the current week. During subsequent analysis, sharp discontinuities were observed within various SVI series, all occurring on 1 January 2011. This coincided with a change in methodology by Google to determine a user's location and it was therefore necessary to restrict the end date of this research to December 2010.

As a comparator for this research, Labour Force Survey (LFS) estimates for Polish nationals in England were used. The LFS is a large continuous survey on the economic circumstances of the UK population. It surveys long term residents in the UK (i.e. individuals who have resided or expect to reside in the UK for 12 months or longer). In addition to its primary purpose of collecting information on the employment status of individuals it also collects information and produced official estimates on nationality and country of birth for the long term residential population. Care is needed in the interpretation of any comparison between LFS estimates and Google's SVI as individuals using Google's search engine include both long term and short term residents.

To enable comparison with LFS data, the weekly SVI data was first aggregated into quarters that corresponded to the reporting periods of the LFS then both series were normalised. Figure

[9] http://www.google.com/trends/explore#cmpt=q

[10] On 1 May 2004 ten new countries joined the EU: the EU8 countries of Estonia, the Czech Republic, Hungary, Latvia, Lithuania, Poland, Slovakia and Slovenia, and the two Mediterranean islands of Cyprus and Malta. Bulgaria and Romania formally joined on 1 January 2007.

[11] Labour Force Survey estimate of Polish nationals residing long term in England, Oct – Dec 2010

[12] Labour Force Survey estimate of Polish nationals residing long term in England, Jan – Mar 2004

3 shows the two transformed data series against each other. Both series have a similar growth pattern from 2004 to December 2010. The correlation between these two series is 0.96. This implies that the growth in the popularity of searches containing the term "polski" could be a very good indicator for the growth in the number of Polish nationals as defined by LFS.

Of note is the observation that the SVI appears to be a leading indicator for the trend in the LFS. This might be expected as the LFS during this time period, only included migrants when they had been resident for 6 months and intended to remain for 12 months or more. Google searches would obviously pick up the search queries from these individuals immediately.

Similar results were obtained for other EU8 populations where population sizes were considered large enough to provide reliable results[13]. Figure 4 shows the comparison for one of the smaller EU8 populations: Estonian nationals. Here the relationship between LFS estimates and SVI does not hold, the correlation is only 0.14. LFS estimates show that there were less than 7 thousand Estonian nationals residing in England in October to December 2010.

In conclusion, this research suggests that search queries might be able to inform on the size and dispersal of some EU8/EU2 populations within England and that search queries could be a leading indicator of such trends. However there are a number of methodological and data quality issues that also need consideration. Firstly the popularity of each search is reported as an index rather than a volume of searches. Although this helps to control for the increasing usage of the Internet over time, the data are not informative of the actual level of interest in the search term and careful interpretation is necessary. In addition a different sample of search queries is used every day. This can lead to volatility in the SVIs produced, especially if the search queries of interest generate a very small share of the overall search activity, although volatility could be controlled by conducting the same analysis on different days and averaging results. During the research period, Google changed the method of determining the location of search queries and applied a retrospective update to all search queries from 1 January 2011. It was observed that the SVI for various analyses had sharp drops or discontinuities at this time, any analysis across a longer time series would require a better understanding of the methodology underpinning location estimation. Another limitation or challenge around using Google Trends to understand population movements relates to the choice of specific search terms. Google trends does not report an SVI if volumes of specific searches are deemed too small so it is crucial to select search terms that should generate larger volumes. For this research it was reasoned that searches made in the native language would be specific to the population of interest. The use of the term for a country's language would also be common as individuals look for translation services and cultural activities relevant to their nationality. It is notable that EU8 populations in England were very small prior to EU expansion. New arrivals from these populations might reasonably be expected to have a preference in using their native language within search queries. Finally, as official statistics have shown, the majority of migrants were young adults who, in the main would be expected to engage heavily with the Internet. All of these factors contribute towards favourable results within this research on search queries. As a comparison, more established populations were investigated such as Indians, Pakistanis and Bangladeshis. Results on these populations were confusing as it was difficult to align the population generating specific search queries using foreign terms with the definition of LFS nationals.

It is important that the selected search term is correlated with the phenomena of interest, in this case immigration. However, care needs to be taken to select a term that does not have an alternative meaning (perhaps a name, place etc) that might confound results. Improvements to this research might involve similar analysis using multiple terms as this might generate greater volumes of searches with which Google Trends can report.

[13] A population size of around 50 thousand was required to generate reasonable results

Notably, a well documented issue around the use of big data sources for inference is the distinction between correlation and causation. The results here show that, between 2004 and 2010, the search volumes for 'polski' are correlated with estimates of Polish nationals but this does not necessarily mean that an increase in search volumes for 'polski' will indicate an increase in Polish nationals. We do not know the underlying theoretical model, we cannot be sure what is influencing the search volumes we are observing and what might impact on them in the future. Since we do not know the underlying model behind the correlation we cannot be sure it will continue over time or for other nationalities (as we have seen for Pakistanis etc.).

If similar results were found for the period since 2010 they would need to be used with caution, perhaps rather than using them as an official estimate they could be used as early indicators of sudden changes to trends and be used alongside and to compliment traditional more robust/reliable estimates. This might include the potential of using the results within the quality assurance of LFS estimates on nationality including the dispersal across England

Further research could also investigate the potential of search queries originating in other countries to inform on imminent immigration patterns. For example, it might be reasoned that potential immigrants would search for information on popular destinations, such as "London" and "accommodation" or "jobs", just prior to migrating. The identification of such search terms could be informed by qualitative research into the internet usage of the particular population groups of interest.

5. Discussion – the benefits and the challenges

Three cases studies from the ONS Big Data team have provided illustrations of both the benefits but also the challenges of using big data sources to estimate the population and population mobility. This section firstly summarises the potential benefits and then discusses the challenges putting forward recommendations and approaches to overcome or at least manage those challenges.

5.1 Benefits

The key advantage of many big data sources over traditional data sources used in official statistics is that they are more timely and are available more frequently and allow for more granularity, particularly on a spatial scale. For example, in England and Wales commuting flow estimates are traditionally produced using Census data every 10 years, although it takes around 2 years to process and produce these outputs. Mobile phone data could be used to produce commuting flows on a more frequent basis with the estimates being much more timely. Google Trends and Twitter data are also potentially available in near real time. These data sources also have the additional benefit that they involve less respondent burden than traditional data collection approaches such as surveys or a Census. It is extremely unlikely that these traditional data sources would be directly replaced by new big data sources (since the ultimate value lies in combining all of these data sources together) but using big data sources may lead to the reduction in the number or complexity of questions in surveys and in the Census and hence reduce respondent burden overall. Although there are often costs associated with some big data sources (such as mobile phone data) others are available for free (Google Trends and some Twitter data) and overall are likely to be cheaper than running a survey or Census and hence offer additional insights without incurring significant direct costs. Also big data sources provide opportunities to gain new intelligence and potentially produce new outputs that meet user requirements, for example mobile phone data could be used to produce estimates for new flexible population bases such as seasonal, daytime/night time population to compliment traditional annual mid-year population estimates. Big data sources could also be used as a leading indicator of a trend (such as Google search queries

for migration patterns) or to quality assure official estimates (such as additional intelligence on internal migration derived from geo-located Twitter data).

5.2 Challenges

There are a number of challenges in realising the benefits identified in using big data sources to estimate population and population mobility and many are associated with data quality. Many big data sources are biased, they do not represent the population of interest, (for example not everyone owns a mobile phone) or we do not even know who is represented in the data, (for example the demographics of Twitter users who chose to enable geo-location). The ONS Big Data team have begun to investigate this issue by using a survey to better understand those individuals who own mobile phones and use Twitter. Methods are being developed to use these survey data as a benchmark to calibrate estimates from the big data source to produce more representative results. Another approach to overcoming the bias within big data sources is to use them as covariates within a model (that could adjust for bias) such as small area estimation approaches. Essentially there is a need to develop an estimation framework for big data, administrative, survey and Census data. This task is far from trivial and will require collaboration across NSOs and academia. The ESSnet Big Data project[14] includes a pilot which has started to look at this issue investigating how to combine big data sources and official statistical data to improve current statistics and create new ones. Another application of big data sources within official statistics (where total population coverage is potentially less critical) is quality assurance. The timeliness of big data sources may provide early indicators of a trend (such as a unexpected increase in the student population of an area based on geolocated Tweets or mobile phone data) that can be used to validate data anomalies seen in official estimates that are not able to pick up these sudden changes.

As has been illustrated by the case studies, big data sources often use different definitions than those adopted within official statistics. For example, mobile phone data does not necessarily align to the definition of usual residence and Google search queries and Twitter data will include long term and short term migrants. The big data sources can be adjusted to align definitions but in addition we should use this as an opportunity to challenge existing or embrace new definitions. Alternative definitions may be appropriate for experimental (rather than official) outputs or definitions driven by big data sources might better align with user needs and hence official outputs should be modified.

Traditionally official statistics have been based on a controlled, measurable data source e.g. a Census or survey designed for the purpose of official statistics. NSOs therefore know all the details of the data collection and processing. NSOs have less control over administrative data and this is a well documented disadvantage of using this type of data for statistical purposes [7]. These issues are further exacerbated with big data sources. With administrative data sources there will usually be plenty of warning of any changes that might affect a statistical output allowing contingency plans to be put in place. However, as is illustrated by the Twitter pilot and the drop in volumes of tweets in September 2014 some big data sources could be affected by changes with little or no warning. Furthermore, it may not always be clear why the source has changed, or even that the source has changed at all. At the very extreme it is possible that a big data source could suddenly be unavailable, Twitter might decide to stop providing access to their data or a mobile phone company might cease trading. Thus, analysts producing outputs and statistics based on this type of data must be extremely alert to these risks and continually evaluate and assess the quality of the data and avoid reliance on one particular data set. Big data products that are made available from commercial companies will have the additional challenge that modelling or processing approaches (such as the algorithms used to identify home or work location from mobile phone data) will be commercially sensitive

[14] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP7_overview

and hence may not be transparent to NSOs using the data. When acquiring data one needs to negotiate for as much metadata and transparency of methods as possible.

Overall new quality frameworks are required for the use of big data sources within official statistics. Understanding and measuring the quality of big data sources is critical to being able to understand and measure the statistics produced using them. ONS has developed methods to measure the uncertainty of population estimates based on survey, Census and administrative data sources [5]. These approaches should be extended to consider big data sources. In general work looking to develop quality frameworks for administrative data needs to be extended to cover big data sources since many of the challenges are the same.

There are also a number of additional challenges around data access when using big data sources for estimating population and population mobility. Many big data sources will need to be acquired from commercial organisations, such as MNOs. In some countries legislation may be in place to enable such data sharing. In the UK work is being undertaken to develop a new legislative framework for better access to data for National and official statistics and statistical research[15]. Even with stronger legislation in place the focus should be around collaboration between NSOs and commercial companies. Eurostat conducted a review of activity across various EU statistical organisations [2] and concluded that access to data requires trust building cooperation between all of the parties involved, to allow projects to grow from small-scale pilot projects to wider collaborations. NSOs must look for ways in which they can add value to these collaborative opportunities in order that both parties benefit. For example the NSO could provide an independent and objective comment on the quality of statistical products.

The challenges surrounding accessing commercial data and establishing partnerships has been the focus of a number of international initiatives. Guidelines for the establishment and use of partnerships in Big Data Projects for Official Statistics have been produced by the UNECE Big Data project[16]. In addition the recently held UN Big Data Global Working Group International Conference on Big Data for Official Statistics[17] dedicated a whole day to 'Access and Partnerships'. These initiatives and discussions between NSOs and commercial organisations must continue to facilitate more collaborative arrangements ultimately leading to greater access to big data sources for official statistics.

Another challenge around accessing and using big data sources to estimate the population and population mobility relates to ethical and privacy issues. The use of data sources such as Twitter and mobile phone data within official statistics will raise privacy concerns around access to, or use of, personal data particularly around the risks of identification. NSOs must commit to protecting the confidentiality of all the information they hold and to balance ethical concerns against the public benefit of making use of particular data sources. The ONS has established its National Statisticians Data Ethics Committee to provide advice on these ethical issues and ensure this balance is met. In addition ONS has also supported research in this area undertaken by IPSOS Mori to understand public views on the use of data science by government (since public perception on issues around privacy can change over time) [4]. The results were used to inform a Data Science Ethical Framework that has been developed[18] to provide guidance to data scientists working across Government.

---

15 https://www.statisticsauthority.gov.uk/publication/delivering-better-statistics-for-better-decisions-data-access-legislation-march-2016/

16 http://www1.unece.org/stat/platform/display/bigdata/Guidelines+for+the+establishment+and+use+of+partnerships+in+Big+Data+Projects+for+Official+Statistics

17 http://unstats.un.org/unsd/bigdata/conferences/2016/

18 https://www.gov.uk/government/publications/data-science-ethical-framework

6. Conclusion

This paper has demonstrated that there are benefits in using big data sources to provide intelligence, enhance or ultimately produce estimates of the population and of population mobility. Data sources such as geo-located Twitter, mobile phone data or Google search queries can potentially be more timely, more frequent and more relevant than traditional data sources used within the production of population statistics. These data sets may also be available at lower costs than traditional collection methods and reduce respondent burden. Big data sources may allow NSOs to produce new outputs (particularly when combined with traditional data sources) to meet user requirements.

The three case studies presented in this paper have illustrated a number of challenges with using big data sources to produce population and mobility estimates but approaches to overcoming and managing these challenges have also been identified. Research is being undertaken to develop methodologies to measure and adjust for bias in big data sources and to integrate these data sources with survey, Census and administrative data. Quality, ethical and commercial frameworks to address these respective issues are being debated and developed and this work is being undertaken collaboratively across NSOs (through UN and Eurostat sponsored initiatives) and with academia and the commercial sector. These collaborations need to continue to be supported if all of the challenges are to be overcome. Big data and data science are still relatively new fields and hence stakeholders need to work together across sectors to share experiences and expertise.

Finally NSOs also need to challenge their own traditional approaches and definitions. Big data sources may not always fit into the established methods and classifications. Using these sources may require the adoption of new approaches or the release of new experimental outputs. This should not be discouraged, even if these outputs may not be of sufficient quality to be an official statistic. Provided these quality issues are understood and communicated the outputs could meet a user requirement, provide some insight or early indicator of a trend. In this way, big data sources could complement or be used alongside an official estimate to improve understanding and measurement of the population and population mobility.

References

[1] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughn, V.D. Blondel, A. J. Tatem, Dynamic Population Mapping using mobile phone data, Proceedings of the National Academy of Sciences of the United States of America, 2014

[2] Eurostat, Feasibility study of the use of Mobile Positioning Data for Tourism Statistics, Consolidated Report, 2014

[3] M. Ester, H. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (1996).

[4] Ipsos MORI, Public Dialogue on the Ethics of Data Science in Government, 2016

[5] ONS, Uncertainty in Local Authority Mid-Year Population Estimates, 2012

[6] N. Swier, B. Komarniczky and B. Clapperton, Using geolocated Twitter traces to infer residence and mobility. GSS Methodology Series No. 41, 2105.

[7] UNECE, Using Administrative and Secondary Sources for Official Statistics – A Handbook of Principles and Practices, 2011

[8] S. Williams, Statistical uses for Mobile Phone data: Literature review. ONS Methodology Working Paper Series No. 8, 2016.

Figure Captions

Figure 1: Net Flows of Geolocated Twitter Users by Month for the 20 Local Authorities in England and Wales with the Highest Proportion of Students, 2014

Figure 2: Daily Volumes of Geolocated Tweets by Device (Great Britain, 15 August 2014 to 31 October 2014)

Figure 3: Google trends search volume index for searches including “polski” term compared to Labour Force Survey estimates of Polish nationals in England,transformed data

Figure 4: Google trends search volume index for searches including “eesti” term within translation category, compared to Labour Force Survey estimates of Estonian nationals in England, transformed data

Figures

Figure 1:

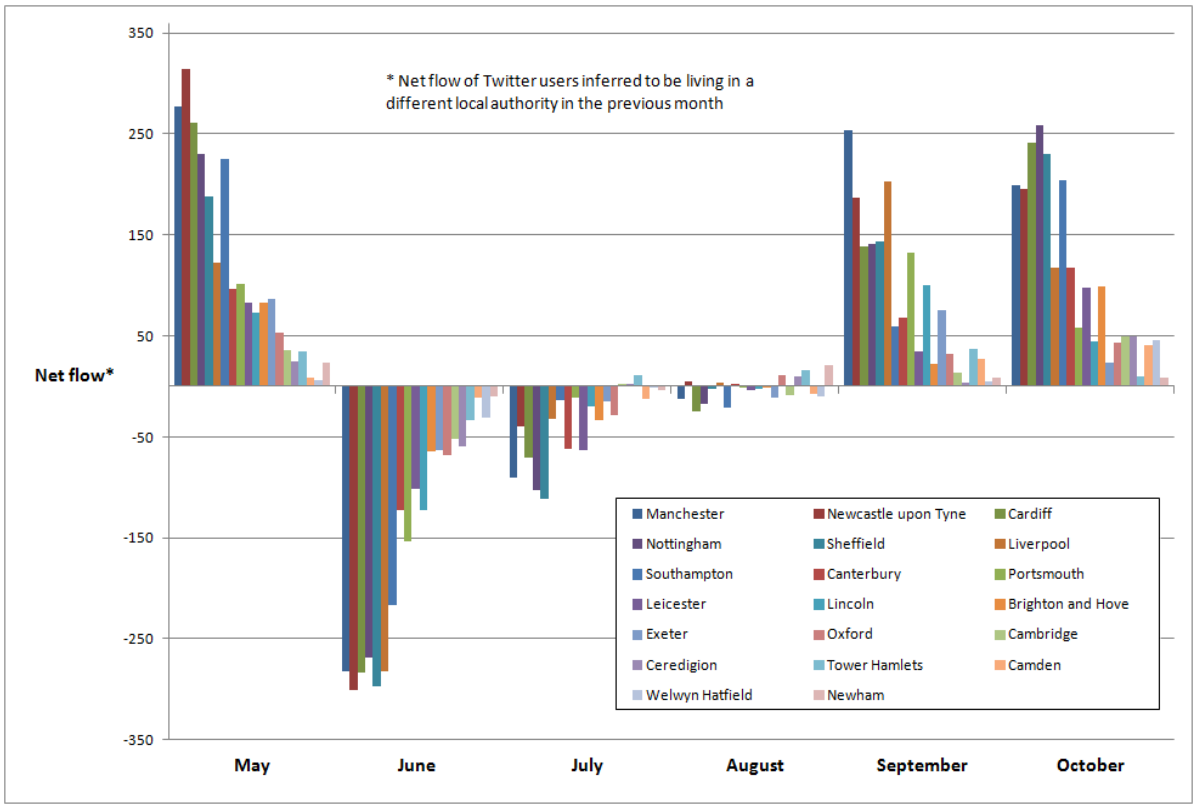


Figure 2:

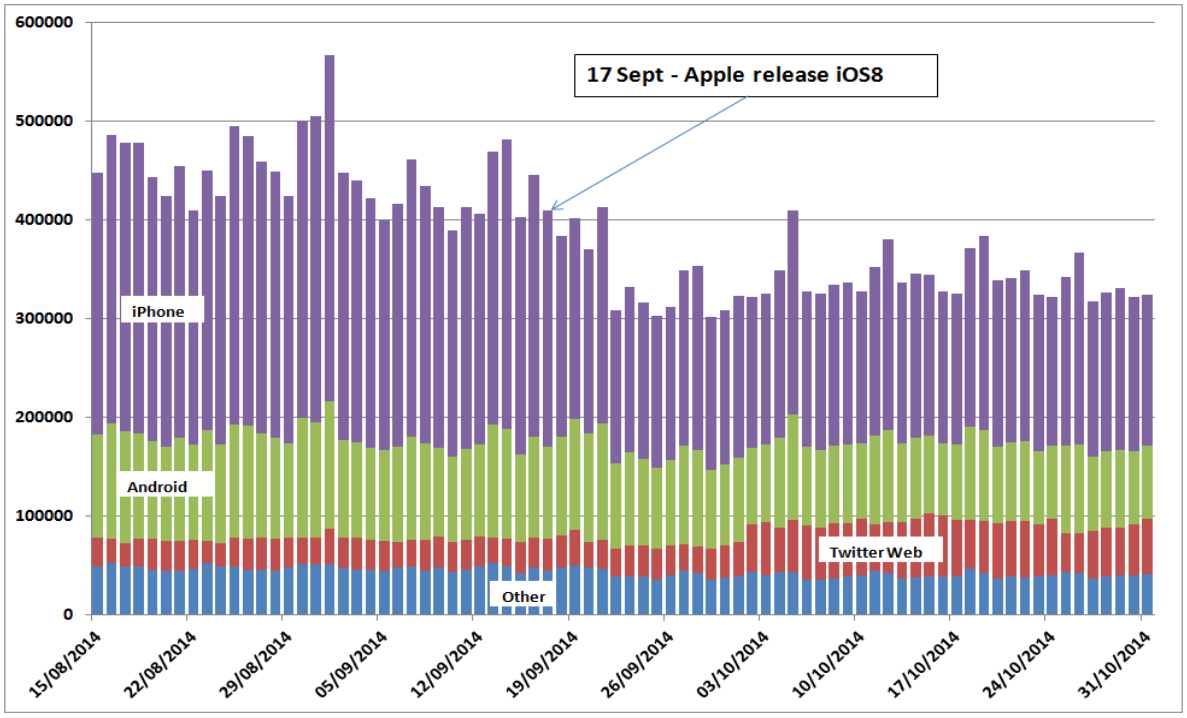


Figure 3:

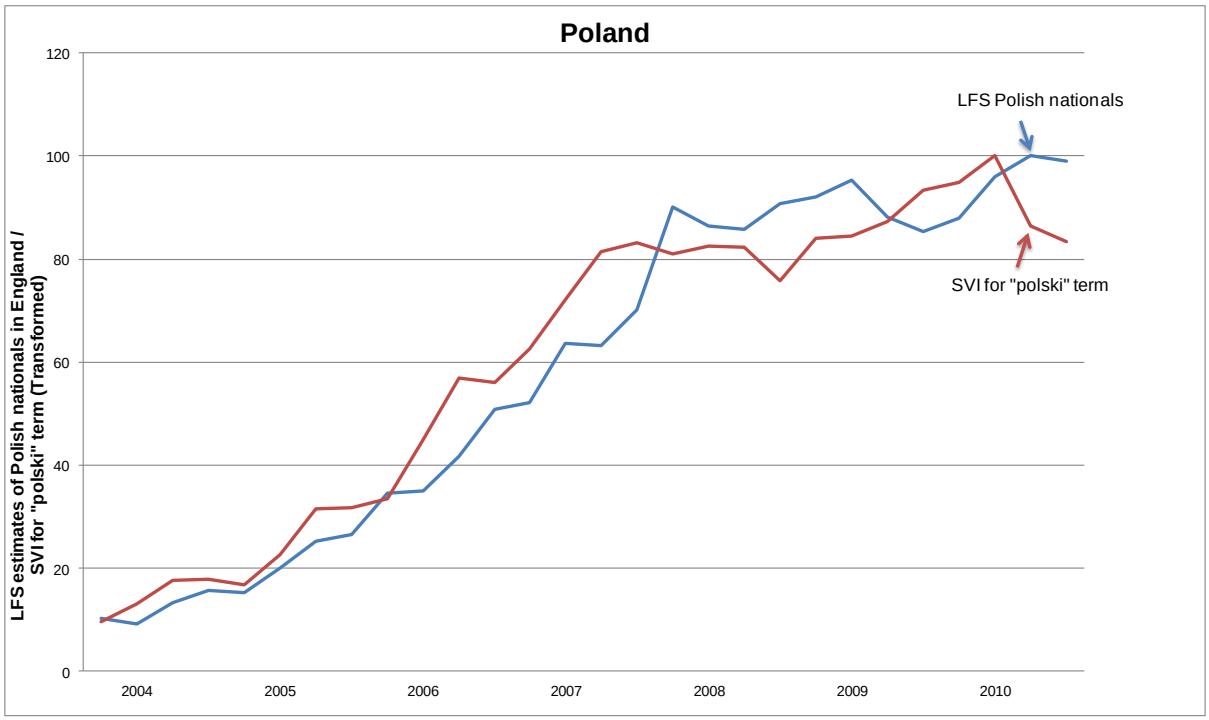
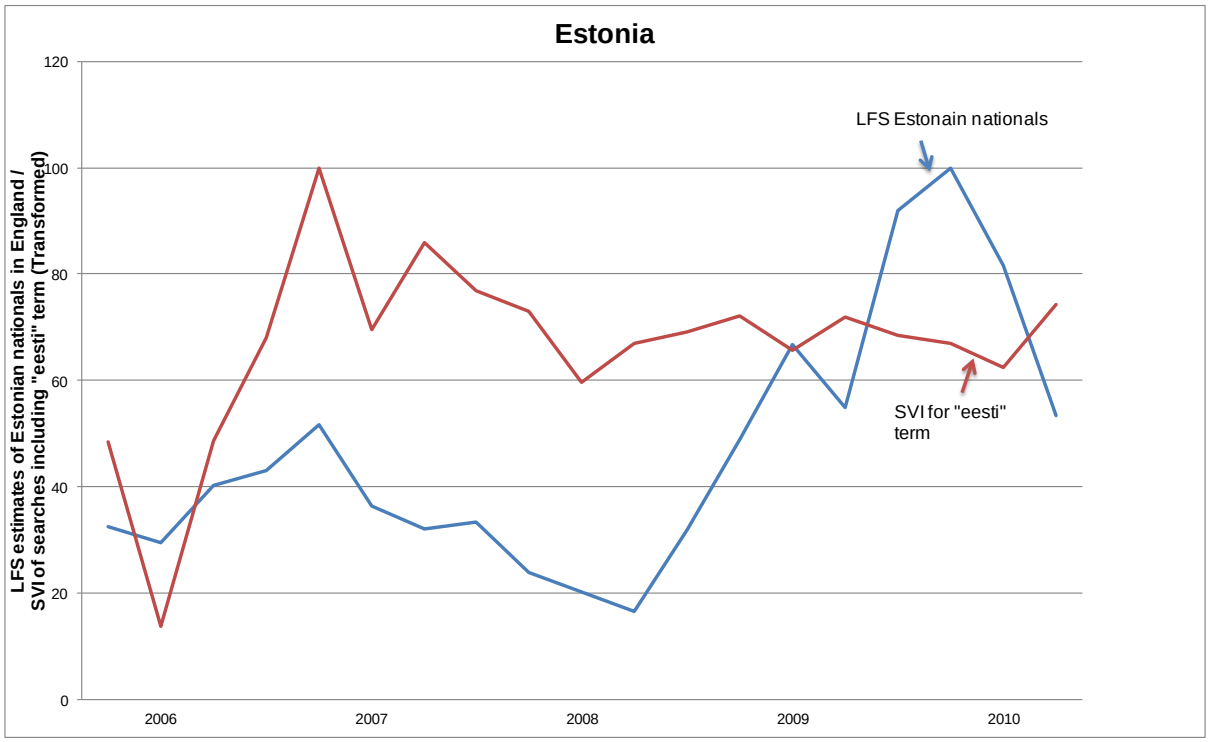

Poland
LFS estimates of Polish nationals in England / SVI for "polski" term (Transformed)
120
100
80
60
40
20
0
2004
2005
2006
2007
2008
2009
2010
LFS Polish nationals
SVI for "polski" term

Figure 4:


Estonia
LFS estimates of Estonian nationals in England / SVI of searches including "eesti" term (Transformed)
120
100
80
60
40
20
0
2006
2007
2008
2009
2010
LFS Estonain nationals
SVI for "eesti" term