Draft Proposal for collaborative research to understand the quality of statistical products derived from mobile phone data

Version 0.1 prepared 8 Feb 2016 by Susan Williams, ONS

The Office for National Statistics (ONS) is the UK's National Statistical Institute and is the largest producer of official statistics on the economy and population. Traditionally, these statistics have been produced from data collected through national statistical surveys and the decennial population census although in recent years there has been more incorporation of the large administrative data collections held in government departments.

The evolution and adoption of new technologies such as the Internet and wireless communications has led to the generation of huge amounts of digital information on the interactions people make. This data may reveal patterns or behaviours in the population that would have relevance for some official statistics.

The geo-coded data generated passively by mobile phones is of key interest to ONS as it has the potential to inform on population densities and mobility, both of relevance to ONS outputs. The primary focus to date has been on the potential of using mobile phone data to model commuting flows as customarily produced in Census.

Although still in an early phase, public transport bodies have been commissioning pilot studies on the use of mobile phone data to model transport flows including those from commuters. There is currently no standard for the validation of the outputs, with Mobile Network Operators, analytical organisations, transport consultancies and the public transport bodies themselves all performing QA as best able with the data available to them.

The Department for Transport has further commissioned the Transport Catapult to conduct a series of workshops to gather intelligence across the range of organisations involved with the production of transport flows. One key observation from these workshops is that there is a desire to have guidance on the validation process, potentially with an independent body providing oversight.

ONS is independent of government and has a role in providing objective comment on the quality or fitness for use of data. ONS is keen to form collaborative relationships with MNOs to assess the quality of the statistical products derived using mobile phone data and to help improve these outputs if possible.

Benefits of collaboration would include:

- Independent validation for methodology/outputs
- Suggestions for improving methodology
- Increased assurance on fitness for purpose for customers of the modelled data
- Greater adoption of using the modelled data
- Identify the need for new data collection through surveys (e.g. working patterns, demographic differences between contracted and pay as you go customers, land use etc.)
- Improving public perception through demonstration of the national benefit of using this data
- Enhanced brand reputation through demonstrating corporate responsibility

Draft specification for research to compare mobile phone data commuting flows with Census Origin-Destination travel to work

Objective of research

To compare commuting flows modelled using mobile phone locational data with MSOA level 2011 Census data. Any differences to be examined and learning made to help understand both data.

Data required from provider

Table 1 Analysis data sets required

Analysis Set	Data	Priority	Analysis
1	Fully weighted MSOA level estimates of Origin-Destination (home-work) matrices, for an area spanning at least 3 Local Authorities. i.e. to identify all commuting flows into and out of the sample area. For clarity: if an individual is inferred as having a commuting journey from MSOA A to MSOA B and does this commute 3 times per week (on average), then they will form a flow of 1 from MSOA A to MSOA B. These flows are to represent numbers of commuters by main mode of transport. (For commuters with more than one transport mode, main mode is defined as the mode used to cover the greatest distance of their overall commuting journey). Main mode of transport to include road vehicle, train or other mode.	Essential	Compare with MSOA level 2011 Census data on commuting flows
2	Equivalent MSOA flows, identifying **average weekday / weekend** numbers of people by main commute. This data may also be differentiated by average flow per hour.	Desirable	To compare with 2011 Census data, in order to understand the difference between average weekday flow and main commuting flow (as defined in Census).
3	For each MSOA in the sample area: calculate the average number of times per week that the home-work journey is made by each commuter resident there. For each MSOA: the distribution of these averages (i.e. count of 'commuters' who commute, on average, less than once a week,	Desirable	To help understand the difference between average weekday commuting count and Census data.

	less than twice a week, less than 3 times, less than 4 times, less than or equal to 5 times).		

Data Principles:

- Modelling should use mobile phone data in an area where there is at least a market share of 15% of all mobile phone subscribers.
- Counts should be of unique users (i.e. not total mobile counts as some users have more than one mobile).
- Averages and inference of commuting journeys should be calculated using at least four full and consecutive weeks of location data – in months of Feb/Mar if possible (i.e. not including Easter). A year closest to the Census year of 2011 is also desirable.
- Main work location desired - i.e. some people may have more than one job.
- Apply disclosure rules as necessary (e.g. threshold values for minimum flow count).

Research Proposal

The research will start with a comparison between mobile phone data commuting flows and 2011 Census travel to work flows at MSOA level using Analysis dataset 1 referenced in Table 1. Mobile Phone data flows at this level should be reliable and not be greatly affected by disclosure controls.

It is proposed that comparisons will involve validation measures including, but not restricted to:

- Correlation/Scatterplots
- Relative mean squared error (RMSE)[1]

Secondary investigations:

Contingent on what is observed in the initial comparison the research will investigate the following features of an area:

Investigation	Rationale
Industry characteristics in area – agricultural, manufacturing, retail, hospital, education etc.	Investigate if different industries have working behaviour where it is more or less likely to reliably identify 'workplace' in mobile phone data model. (e.g. shift worker / parttime patterns might distort)
Population changes	Investigate population growth for impact on mobile phone data flows. E.g. new housing estates, general population growth from Census

[1] Fij will denote the mobile phone data flow from the i-th Origin MSOA to the j-th Destination MSOA

Rij will denote the Census flow from i-th Origin MSOA to j-th Destination MSOA.

The formula for this measure is RMSE = 1/(NxM)*sum [i=1..N][j=1..M] [(Fij-Rij)/Rij]^2

	date
Employment changes	Investigate employment rate changes since Census
Demography/Topography of areas	Urban/rural, Index of Multiple Deprivation, Census area classification, Land use

**NB This research is to understand comparison with Census data – and ancillary data available within ONS.

If LSOA level mobile phone data is provided as in Analysis dataset 2, then the same research will be conducted at this level.

It is understood that transport bodies are receiving average weekday commuting flows derived from mobile phone data. This data is conceptually different to Census commuting flows and it would be desirable research to investigate how these average weekday commuting flows should compare so that guidance may be developed for comparisons with Census commuting data. It is proposed that this research is only to be conducted at MSOA level. Data specified in Analysis dataset 3 and ideally 4.

A workshop will be arranged to go over the research findings and discuss additional research ideas as well as, subject to agreement, proposals for publishing.

Comment [w1]: Does this section need more specific T&Cs? Such as ONS will sign the MNOs NDA but will want to publish generic findings of value to the whole industry etc.??

Is there a standard text for this – has ONS done research with commercial data (access given for free) and stipulated Terms as well?

Location of data

If it is agreed in the research proposal and contract that LSOA level will ultimately be required for research, it is proposed that all data required will be loaded into the new Virtual Micro Lab (VML) as a project to be accessible by named researchers only. However, if the research is only to be conducted at MSOA level then this research may be conducted on an ONS secure server, restricted to named persons conducting the analysis.

Activities and Responsibilities

id	Activity	Owner	Date due (TBA?)
1	Research proposal to be agreed	All	Mar/Apr 2016
2	Appropriate contracts to be arranged	All	Apr 2016
3	Arrange for non-disclosive/aggregated mobile phone data derived commuting flows to be created (as per research proposal).	Data supplier	Apr/May 2016
5	Encrypted mobile phone data to be provided to ONS	Data supplier	May 2016
6	Mobile phone data to be loaded into VML or Secure Server as per contract arrangement	ONS	May 2016
7	Initial research on MSOA level data	ONS	June 2016
9	Workshop to go over results/learning of research	All	July 2016
10	Working paper to be prepared for publishing	All	Autumn 2016