ESSnet Big Data

Specific Grant Agreement No 1 (SGA-1)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/.....

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2015.007-2016.085**

Work Package 5

Mobile Phone Data

Minutes of internal meeting in Madrid 7-8 June 2017

Version 2017-06-12

Prepared by: David Salgado (INE, Spain)

Elisa Esteban (INE, Spain)
Soledad Saldaña (INE, Spain)
Luis Sanguiao (INE, Spain)
Maria Novas (INE, Spain)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

DAY 1: Wed, 7 June 2017

Participants:

Ciprian Alexandru	**INSSE, Romania**
Marc Debusschere	**Statistics Belgium, Belgium**
Elisa Esteban	**INE, Spain**
Baldur Kubo	**Cybernetica, Estonia**
Suelí Lorenzo	**INE, Spain**
Miguel Ángel Matínez	**INE, Spain**
Pasi Piela	**Tilastokeskus, Finland**
Roberta Radini	**ISTAT, Italy**
Benjamin Sakarowith	**INSEE, France**
Soledad Saldaña	**INE, Spain**
David Salgado	**INE, Spain**
Luis Sanguiao	**INE, Spain**
Martijn Tennekes	**CBS, Netherlands**
Margus Tiru	**Positium, Estonia**
Susan Williams	**ONS, UK**
Markus Zwick	**DESTATIS, Germany**

Planned Agenda:

09:00-09:30	Welcome and agenda adoption	
09:30-11:00	Description of compiled mobile phone data sets	FR, UK, DE, IT, NL, BE, FI
11:00-11:30	Coffee break	
11:30-13:00	Description of target statistical outputs	FR, UK, DE, IT, NL, BE, FI
13:00-14:00	Luch	
14:00-15:00	Methodological proposals I: general framework, ecological sampling and geostatistics	ES
15:00-16:00	Methodological proposals II: admin data methodology and Heckman correction technique	RO
16:00-16:30	Coffee break	
16:30-18:00	Debate on input data sets, outputs and methodology	All

Session Contents:

1. **Welcome and agenda adoption.**
   Miguel Angel Martínez, director of the dept. Methodology and Statistical Production in Statistics Spain (INE) opened the meeting and welcomed participants.
   The WP coordinator pointed out the main issues of the day's agenda.

2. **Description of compiled mobile phone data sets.**
   In turns, each participant presented the description of their mobile phone data sets.

   [FR]

   - Three-year long agreement between Orange, INESEE and Eurostat to exploit a data sets of five months of CDRs (mid May to mid October of 2007) to experimentation inside Orange infrastructures.
   - Recent access to a customer data set with geographical information (postal code).
   - Complex access to CDRs even inside Orange and difficulties to update tools and configure a user community.
   - No guarantee of further access when the convention expires.

   [UK]

   - Engagement with MNOs and public Transport Bodies to access fully modelled commuting flows derived from mobile phone data together with high level overview of the methodology used.
   - Two data sets from data collected from February to April of 2016: commuter flows originating or ending in three London Local Authorities and number of commuting journeys made during the weekdays or weekends.

   [DE]

   - Aggregates hourly data on total number of events/SIM cards.
   - Total counts potentially broken down by nationality, or by age group.
   - Probably access to data in summer.

   [IT]

   - CDRs refers to calls and text messages in the province of Pisa from 1^{st} January to 12^{th} February.
   - No information about antenna location.
   - Geographical attributes coarse-grained to municipality level.

   [NL]

   - Aggregates CDRs data for eight months in 2013 and 2014 from Vodafone via Mezuro per hour, municipality, and municipality of residence.

[BE]

- Two data sets received and waiting for two additional sets, adapted to specific statistical products.
- Data are counts of devices per grid cell.
- Situation with MNO Proximus is unclear for the moment. Data provision has been temporarily suspended.

[FI]

- Partial aggregates (count of devices) from two of three Finnish MNOs. Negotiations with the third one will be resumed shortly.

More details can be found in meeting presentations.

3. **Description of target statistical outputs**
In turns, each participant presented the description of their target statistical outputs.

[FR]

- Already no clear outputs.
- Home detection and daytime population, mobility.

[UK]

- The main objective is to compare commuting flows derived from mobile phone data with Census Travel To Work.

[DE]

- Population density and mobility.
- Inbound tourism.

[IT]

- Measuring urban population.
- Mobility.

[NL]

- Daytime population.
- Tourism.
- Mobility.

[BE]

- Census population by living place and workplace.
- Matrix living place-workplace.
- Dynamic population mapping.
- Other: mobility, tourism, time user, circular migration, etc.

[FI]

- Inbound and outbound tourism.

More details can be found in meeting presentations.

After the expositions of data sets and statistical outputs, there was a debate on these issues. The main conclusions were:

- The situation regarding data sets is very different in each country and data access is not yet a closed issue.
- It is necessary to establish a standard description of data sets to allow comparability and methodological treatment. The standard description should be available at two levels at least:
  - microdata (CDRs and signaling information): at least the standard description should include a clear definition of the identification of the mobile phone and of the temporal and spatial attributes.
  - partial aggregates (count of devices, etc.): in this case, variables depends very sensitively on the target statistics. The biggest problem arises because the MNOs are the ones that transform their raw records into aggregated data sets.
    It would be an useful output of the project to agree on the description of different kinds of data sets according to the statistics considered.
- The main target outputs amount to counting populations, which will probably entail the estimation of other local parameters, such as local market shares of MNOs.

4. **Methodological proposal I: general framework, ecological sampling and geostatistics.**
   Statistics Spain (INE) presented a methodological framework to produce concrete outputs using the mobile phone data sets.
   As the statistical outputs of interest refer to counting individuals in the domains of tourism, mobility and/or daytime population, the suggested model sets the problem as a problem of estimation of population size (either population of people, tourists, commuters, etc.).
   One of the richest field in methods to estimate population size is Ecology. Several methods to estimate a population size can be found in this discipline. In addition, some Geostatistical techniques seem to be applicable to our problem.

5. **Methodological proposal II: admin data methodology and Heckman correction technique.**
   INSSE (Romania) gave a revision of the different ways in which administrative data can be used to improve official statistics quality and how to link them with mobile phone data.
   He also presented a brief introduction to the Heckman correction techniques. This collection of methods offers a means of correcting for selection bias for non-randomly selected samples so it could be applied to our project since our estimations are only for those people with mobile phone registered to the MNOs under consideration.

6. **Debate on input data sets, outputs and methodology.**
   Statistics Spain (INE) summed up the main conclusions in the day:

   - Documentation.
     All presentations will be uploaded to the open Wiki, except data access slides from France, Italy and the Netherlands that will be placed on the restricted Wiki. UK has to check if data access presentation may be on the open Wiki too.
     Marc encourage everybody to upload as much information as possible to the open Wiki. The restrictions arise because of the particular agreements with data providers.
   - Data sets.
     Proposal of standard descriptions of data sets in Excel format. It will be updated with the new information available of each country and additional variables if needs.
     There will be two file versions, one for data at the mobile device level and another one for partial aggregates. They will be available in the restricted Wiki.
   - Target outputs.
     The main target outputs are counts of different units for
     - Daytime of present population.
     - Tourism (inbound/outbound).
     - Mobility.
   - Methodological framework.
     - No concrete proposal from any existing official statistical technique that can be applied in our study case, so we look for methods in the different fields of statistics that could be used. Some proposals:
       * Ecology techniques.
       * Geostatistical methods.
       * Heckman correction.
     - An interest in working with simulated data sets is expressed to assess how well the proposed models work.
     - Include negative results in the deliverable to show all failed attempts and avoid spending efforts in methods that have been proved not to work.
     - The role of Official Data in the estimating procedures is debated:
       * Benchmark to calibrate mobile phone data.
       * Improvement of mobile phone data estimations with official data, but not calibrating.
     - Multiplicity of devices per person problem.
       There is no knowledge about how MNOs solve this problem.
       In Belgium, Proximus elaborated a small sample to estimate them and extrapolate results to the entire population.

6

Marc reported that Eurostat is going to introduce some questions about mobile phone usage in the Tourism Survey. If the questionnaire is not yet closed, it may be possible for us to propose some new question to include in it.
Susan will provide an article about statistics on mobile phone usage elaborated in the UK.

DAY 2: Thu, 8 June 2017

Participants:

Planned Agenda:

09:00-11:00	Description of national ongoing/intended data procesing	FR, UK, DE, IT, NL, BE, FI
11:00-11:30	Coffee break	
11:30-13:00	Internal technical reports: overview and questions Positium Sharemind HI - encrypted mobile Big Data secure processing	Positium
13:00-14:00	Luch	
14:00-15:00	Agreements and plan of action	All

Session Contents:

1. **Description of national ongoing/intended data processing.**
   In turns, each participant presented the description of their ongoing/intended data processing.

   [FR]
   - Daytime population:
     - quality estimation of data, determining how to evaluate the performance of an home detection algorithm
     - comparison of mobile phone data with official statistics
     - market share estimation
     - defining the best granularity
   - Mobility:
     - link the mobility behaviour to the social level of the residence
     - network of users

   [UK]
   - Identification of use cases for statistics and considering different ways of working.
   - Preparing Datathon in autumn 2017.

   [DE]
   - Good expectation with Deustche Telekom.

  - Data is supposed to be available in August.

[IT]

  - Collaboration between ISTAT, NRC and the University of Pisa to develop a methodology to classify the user from the mobile phone data using the Sociometer tool.
  - The intention is to integrate those data with administrative data relating to the resident and commuting population and to request new data sets with wider and more detailed information (current data set do not have antenna location).

[NL]

  - Their objective is to collaborate directly with MNOs, without the intervention of Mezuro.
  - Testing Bayesian location algorithm to approximate location of events and different ways of combining the mobile phone data with administrative and statistical sources to obtain daytime population.

[BE]

  - No new methodological proposals because no new data sets are available. Previous documents can be found in the Wiki.
  - Administrative and statistical source can improve considerably the utility of mobile phone data.

[FI]

  - The main objective is to measure seasonal population in Finland.
  - This entails especially the measurement of inbound tourism.

2. **Internal technical reports: overview and questions**
   Margus Tiru from Positium presented the main results in the second report that will be available at the end of June.

3. **Positium Sharemind HI - encrypted mobile Big Data secure processing.**
   Baldur Kubo from Cybernetica presented the Sharemind secure computing platform. It is presented as an optional tool to process data in an encrypted and secure way that could overcome the lack of access to mobile phone data due to privacy and confidentiality restrictions. The software is in compliance with late regulations on personal data protection. Cybernetica and Positium are conducting projects to deploy this tool in statistical production and would be willing to cooperate with NSIs from the ESS.

4. **Agreements and plan of action.**

   - Task distribution:
     - Upload the standard description of data sets/target outputs to the open Wiki. Deadline 23^{th} June: All.

  - Models building (including estimation of selection probabilities and other auxiliary information): Spain.
  - Heckman correction techniques to correct from selection bias in the suggested model: Romania.
  - Analysis of the misalignment problem and the relevance of the diverse degrees of geographical breakdown in the analysis: UK, Belgium.
  - Compilation/development/adaptation of computer tools to implement the models: Spain.
  - Application of methodological proposal to simulated data: Spain.
  - Application of methodological proposal to real data: Netherlands, France, Belgium, Finland.
- Next meeting is planned in February 2018 and we could take advantage of the dissemination meeting in Sofia to hold them back-to-back. Further action with Peter and Martin is agreed: Spain.