

## ESSnet Big Data

### Specific Grant Agreement No 2 (SGA-2)

<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 **11104.2016.010-2016.756**

### Work Package 5

### Mobile Phone Data

## Minutes of internal meeting in Madrid 20-21 March 2017

Version 2018-04-12

**Prepared by: Soledad Saldaña (INE, Spain)**

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)

[p.struijs@cbs.nl](mailto:p.struijs@cbs.nl)

telephone : +31 45 570 7441

mobile phone : +31 6 5248 7775



**DAY 1: Tue, 20 March 2018**

Participants:

<b>Sandra Hadam</b>	<b>DESTATIS, Germany</b>
<b>Fabio Ricciato Albrecht Wirthmann</b>	<b>Eurostat</b>
<b>Benjamin Sakarowith</b>	<b>INSEE, France</b>
<b>Roberta Radini</b>	<b>ISTAT, Italy</b>
<b>Susan Williams</b>	<b>ONS, UK</b>
<b>Marc Debusschere</b>	<b>Statistics Belgium, Belgium</b>
<b>Pasi Piela</b>	<b>Statistics Finland, Finland</b>
<b>Soledad Saldaña David Salgado Luis Sanguiao</b>	<b>Statistics Spain (INE), Spain</b>

Planned Agenda:

13:00-13:30	Welcome and agenda adoption	
13:30-14:30	Towards a reference methodological framework for analysis of MNO signalling data	Eurostat
14:00-15:00	Methodology I: the two-phase life cycle model and mobile phone data	ES
15:00-16:00	Methodology II: the core data model and location of events	FR, IT
16:00-16:30	Coffee break	
16:30-17:00	Methodology III: a hierarchical model to estimate populations counts	ES
17:00-18:00	Methodology IV: discussion and recap	All

## Session Contents:

**1. Welcome and agenda adoption.**

Miguel Ángel Martínez-Vidal, director of the Dept. Methodology and Development of Statistical Production at Statistics Spain (INE) opened the meeting and welcomed participants.

The WP coordinator briefly expressed the structure of the agenda commenting the main items. No change was suggested and the agenda was adopted.

**2. Towards a reference methodological framework for analysis of MNO signalling data.**

Eurostat presented a proposal for a reference methodological framework for processing mobile phone data for the production of official statistics, based on the principles of layering and a so-called “hourglass” model adopted in Computer Network Architectures.

The basis of the framework amounts to defining three main layers: the MNO data layer (D-layer), the statistics layer (S-layer) and a convergence layer (C-layer) with a few common definitions connecting the two former specialised layers.

He remarked the importance to develop such a framework that could be evolved and adapted to different settings. It was pointed out too the need to shift from CDRs to the much richer information contained in signalling data.

**3. Methodology I: the two-phase life cycle model and mobile phone data.**

Statistics Spain (INE) presented how the two-phase life-cycle model used for admin data can be flexible enough to be applied in the processing of mobile phone data for statistical purposes, to identify errors and to evaluate their quality.

The need of including a third phase for aggregate data was remarked.

**4. Methodology II: the core data model and location of events.**

ISTAT and INSEE exposed the core data model and their experiences with spatial interpolation.

- ISTAT presented their results using best service areas information.
- INSEE presented the core data model to process mobile phone data included in the deliverable on methodology (originated in Positium’s second internal report). The goal of this model is to describe the position of every individual at every moment in a standardised way.  
This model could set the guidelines for a common methodology for different NSIs

in order to obtain different aggregation and indicators from mobile phone data. The main drawback is that it will require an access to data at an individual level more extended than CDRs.

INSEE also showed some results related with the estimation of the cost of spatial interpolation by using the area proportion method. Finally, the importance of spatial interpolation and the need for validation indicators for these procedures were stressed.

#### 5. **Methodology III: a hierarchical model to estimate populations counts.**

Statistics Spain (INE) exposed the hierarchical model developed for population estimations using aggregated mobile phone data.

The developed methodology adapts the species abundance problem, where the focus is on counting species individuals, to aggregated mobile phone data, where the aim is to estimate population counts of diverse target populations (outbound tourists, resident tourists, commuters, ...).

The proposal is not a definitive solution, but a first piece of a methodological framework that can be increasingly improving.

More details of the model are exposed in the deliverable on methodology.

#### 6. **Methodology IV: discussion and recap.**

Statistics Spain (INE) summed up the main topics in the day and asked the participants for the next two issues:

- Comments and improvements for the Methodological and IT deliverables.
- Feedback regarding the elements for the final deliverable on quality.

Summarised outputs are:

- Statistics Belgium remarked the need for a trade-off between feasibility and availability, and advised to try only to enrich both deliverables (methodology and IT) but not to rewrite them.  
Regarding the deliverable on quality, it was suggested to include a chapter with a list of research topics that can be explored.
- INSEE pointed out the utility of a common framework and a common language to deal with data and the importance of simulated data set to prove methodological proposals. This was strongly agreed by all participants.

It was stressed the importance of include in the quality deliverable what NSIs can bring out from data to offer to MNOs so that they were interested in collaborating with NSIs.

- ISTAT offered to include the IT architecture used in their development in the IT deliverable.
- Statistics Spain (INE) suggested as further complementary ideas for the inference stage of the process to consider state space models as a good framework. The main advantage of this option is its easy scalability and the existence of much work in seasonal adjustment field already developed. It was remarked too there exists a lot of possibilities to be explored both in IT and in statistical methodology for our problem.

**DAY 2: Wed, 21 March 2018**

## Participants:

<b>Sandra Hadam</b>	<b>DESTATIS, Germany</b>
<b>Fabio Ricciato Albrecht Wirthmann</b>	<b>Eurostat</b>
<b>Tracy Li</b>	<b>Flowminder</b>
<b>Benjamin Sakarowith</b>	<b>INSEE, France</b>
<b>Ciprian Alexandru</b>	<b>INSSE, Romania</b>
<b>Roberta Radini</b>	<b>ISTAT, Italy</b>
<b>Susan Williams</b>	<b>ONS, UK</b>
<b>Marc Debusschere</b>	<b>Statistics Belgium, Belgium</b>
<b>Pasi Piela</b>	<b>Statistics Finland, Finland</b>
<b>Soledad Saldaña David Salgado Luis Sanguiao</b>	<b>Statistics Spain (INE), Spain</b>

## Planned Agenda:

09:00-10:00	Flowminder/WorldPop CDR analysis - applications and methods	Flowminder
10:00-10:45	IT I: R package to estimate population counts	RO
10:45-11:15	Coffee break	
11:15-13:15	Quality I: Perspectives to apply this framework on real data sets	FR, UK, DE, IT, BE, FI
13:15-13:45	Quality II: Deliverable on quality: structure, agreements and plan of action	All

## Session Contents:

**1. Flowminder/WorldPop CDR analysis - applications and methods.**

Tracy Li, from Flowminder, presented an overview of the activities conducted by Flowminder.

Flowminder is a non-profit foundation that operationalise applications to support vulnerable population and sustainable development. They collect, aggregate, integrate and analyze anonymous mobile phone data, satellite and household survey data in order to map the distributions and characteristics of vulnerable populations in low- and middle-income countries. They make extensive use of high-resolution population maps from WorldPop project.

WorldPop is a research program providing an open access archive of spatial demographics datasets, focused on low- and middle-income countries to support development and humanitarian applications.

The way they work with CDRs data and some of the issues to take into consideration when working with kind of data were exposed.

The presentation ended with some of the Flowminder project's along the years.

**2. IT I: R package to estimate population counts.**

INSSE presented the R packages *pestim* and *mobloc*. These two packages are explained with detail in the deliverable on IT.

R package *pestim* provides tools for population estimation implementing the hierarchical model developed in the deliverable on methodology.

R package *mobloc* provides mobile location algorithms and tools presented in the same deliverable.

**3. Quality I: Perspectives to apply this framework on real data sets.**

Participants with available mobile phone data were asked about their view on how to finish this work package and the possibility of assessing the proposals presented in the meeting with their real data sets.

- Statistics Finland presented the pilots they are carried out covering inbound and outbound tourism with data from two of the three Finish mobile network operators.

Their approach is fully experimental to gain insight into the data. They expressed their intention to test the methodology and software developed in the current project for the last deliverable.

- ONS, through an open call, could work on fully modelled data on commuting flows in the metropolitan area of London. The analysis allowed expert to conclude that the usually required quality standars for Official Statistics were not met.
- DESTATIS also expressed their interest in testing the hierarchical model with their data.
- INSEE is focused on location of events and the local interpolation process. They remarked the need of robust validation indicators to asses that process. Some inspi-ration can be found in geostatistical indicators. They will also test the methodology and software tools developed in this project using a partial set of their CDRs.
- ISTAT also expressed their intention to test the hierarchical model with their avail-able data.

As objectives of this application to real data participants pointed out the following: (i) a first proof of concept, (ii) the limit of scalability and performance of the software, (iii) the adequacy of data from MNOs as compiled during SGA-1 as inputs for the software tools.

#### 4. **Quality II: Deliverable on quality: structure, agreements and plan of action.**

Statistics Spain (INE) proposed a structure for the final deliverable, with 5 possible chapters:

- a) Identification of source of errors according to the two-phase life-cycle model.
- b) Quality assessment (accuracy, especially) of the inference stage (hierarchical model).
- c) Statistical standards (prototype of standard at least for aggregated mobile phone data sets).
- d) Application of the full model to the real mobile phone data compiled during SGA-1. Efforts to combine real data set with the methodology and IT tools developed.
- e) Conclusions and future perspectives about the work package on Mobile Phone Data.

INSEE remarked the need to include the evaluation of location interpolation, at least as an importante immediate future prospect. Eurostat suggested to add in the future perspective chapter a section of lessons learnt during the project.