WP5 SGA2 Internal Document No. 1:

Proposed Course of Action for March-June 2017

March 10, 2017

WP5 on Mobile Phone Data

ESSNET ON BIG DATA, 2017

This document contains a proposed course of action for the period March-June 2017 ending in the first physical meeting at Madrid in June 7-8. Shortly, we propose to focus on three first relevant aspects to produce a methodological embryo to cover our first goal of SGA-2 (statistical methodology) and to provide an input to WP8 on Big Data methodology. These three aspects are (details below):

1. To provide a detailed description (both from the computer and statistical point of views) of the data we have to conduct our work.

2. To provide a detailed description of the intended statistical output we want to reach with the data in each country. This description must be an explicit mathematical formulation of the statistical problem (e.g. as the problem of estimation in a finite population).

3. To provide a description of the planned (or even already executed in some cases) data processing in each case. This description must be used to extract this first methodological embryo, which in turn must produce some more methodological proposals to be tested with the data (as far as possible – sometimes the processing may be limited by the concrete agreements with the MNOs).

1. Data description

We need to share the full details of the data that our corresponding NSIs are allowed to process within the reached agreement either in-situ or via transmission to our offices (probably in some aggregated or coarse-grained form). To be concrete, we need:

1. To specify both the computer and statistical details of the (anonymised) ID variables of the data sets. From the computer standpoint, what kind of variable(s) are these? Alphanumeric variables with a specific length and/or structure? More importantly, ***how are these variables generated from the different ID variables in a mobile network (IMSI, MSISDN, subscribersÍDs, . . .)?***

2. To specify the time attribute variables contained in the data sets. Again, both from the computer and statistical point of view. Are they timestamps? Are they standard dates? With time zones? Are they identification of time intervals within each day? Any concrete codification? More importantly, ***how are these variables generated from the raw data in a mobile network?*** And what accuracy do they have?

3. To specify the spatial attribute variables contained in the data sets. Again, both from the computer and statistical point of view. Are they coordinates? In which system? Are they identification of coarse-grained localisations (municipalities, city districts, electoral constituencies, grid cells, . . .)? Any concrete codification? More importantly, ***how are these variables generated from the raw data in a mobile network?*** Are they antenna positions? More elaborate localisations (triangulation of signals, . . .)? And what accuracy do they have?

4. To specify any other variable:
    a) Sociodemographic variables (if any) of the subscribers (e.g. nationality, province, ...).
    b) Roamer/non-roamer: this is a fundamental variable for some statistics (tourism, especially).

For the matter of concreteness, I'd like to propose you to create a fake data set (just 5-10 registers; not more) in each case with exactly the same structure as the real data. This may be a very helpful visual aid.

2. Output description

We need to explicitly and very clearly specify what statistical output(s) we are to produce out of these data in each case. In particular, we need to clearly spell out what population aggregate(s) we are to estimate (population total, population density, total of national tourists, total of commuters, ...) and their levels of disaggregation, both in time and geographically (each week, each day, each hour, ... and in each city districts, each municipality, ...). This includes the definition of the involved statistical variables (tourist, commuter, ...) – of course, if this is a standard definition established in a European regulation, please just specify.

From the methodological point of view, this is a key step because in our opinion we need to formulate whether we are solving a problem of estimation in a finite population or a classical inference problem.

3. Data processing description and methodology

For those already processing and following some methodological proposals, we need to know these details to begin producing a mobile phone data methodology. Our proposal is that from these descriptions of each country with access to data, those not having access will analyse them and combine them to produce a unified approach. The ideal situation would be to produce a two-way communication from the concrete case to the generic methodological feeding the former with new proposals and feeding the latter with concrete difficulties detected in the practice.

Please in this sense we need to share reports, documents, articles and any written work you have (we do not intend to write everything anew).

4. DEADLINES

We need to fix some deadlines for the work to advance. Our proposal is:

Data description (point 1)	31 March
Output description (point 2)	21 April
Processing description (point 3)	Physical meeting in June

Apart from Webex meetings we can fix along the way, I propose to fix two Webex meetings to revise jointly the first two points in the week 27 March - 1 April and in the week 1-5 May (for the third point we can use the physical meeting itself).