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

Deliverable 5.5

Some Quality Aspects and Future Prospects for the Production of Official Statistics with Mobile Phone Data

Version 2018-05-31

Prepared by: David Salgado (INE, Spain)

Marc Debusschere (Statistics Belgium, Belgium)
Ossi Nurmi, Pasi Piela (Tilastokeskus, Finland)
Elise Coudin, Benjamin Sakarovitch (INSEE, France)
Sandra Hadam, Markus Zwick (DESTATIS, Germany)
Roberta Radini, Tiziana Tuoto (ISTAT, Italy)
Martijn Tennekes (CBS, Netherlands)
Ciprian Alexandru, Bogdan Oancea (INSSE, Romania)
Elisa Esteban, Soledad Saldaña, Luis Sanguiao (INE, Spain)
Susan Williams (ONS, UK)

ESSnet co-ordinator:

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

Contents

[1]Prepared in collaboration with Fabrizio de Fausti and Luca Valentino.

1

Introduction

This document proposes the first steps in the quality assessment for the use of mobile phone data in the production of official statistics. It corresponds to the final fifth deliverable of the work package on mobile phone data of the ESSnet on Big Data (ESSnetBD, 2017).

In deliverables WP5.1 (2016) and (WP5.2, 2017) we focused on the access to mobile phone data. The original goal was multiple, namely to take stock of the access to this data source in the ESS, to compile enough data sets to conduct a complete research from methodological, IT, and quality points of view in a hands-on bottom-up approach, and to pave the way for the integration of these data in the routinely production of official statistics in the ESS. Having access to data, the next planned step was to investigate the statistical methodology, the IT environment, and the quality issues necessary for their usage.

In deliverable WP5.3 (2018) we proposed some elements for the construction of a methodological framework. Basically, after noticing that mobile phone data, as administrative data, are generated without statistical metadata, we propose to adapt the two-phase life-cycle model for statistical microdata by Zhang (2012) to describe the analogous process with mobile phone data. Although this needs further research with more advanced work in this direction (Reid et al., 2017), this brings mobile phone data to a common statistical language in use with administrative registers. One of the key elements of this model is the identification of errors along the process of data generation from network events (objects) to individuals (statistical units). Data quality is embedded in the approach from the very beginning.

In the deliverable WP5.3 (2018) we also identified and proposed methodological elements for important steps in the entire process, namely, (i) geolocation of network events, (ii) space-time interpolation of events, and (iii) a hierarchy model providing the

inference from aggregated mobile phone data to target populations. These are separate elements to be integrated in an entire end-to-end process.

In deliverable WP5.4 (2018) on IT, we offered three main outputs. Firstly, a description of a computer platform to access and process statistical microdata in an MNO's premise was included. This comes from the actual access conditions INSEE enjoys to access and process data in Orange Labs' premises. This is the only example of a platform with such nature which this work package's members have been able to work on given the limited access to data. Secondly, a first R package called `mobloc` for the geolocation of network events has been developed. The code is open and freely available at our GitHub page (WP5GitHub, 2018). Thirdly, a second R package called `pestim` for the implementation of the aforementioned hierarchical model has also been developed. The code is open and freely available at our GitHub page (WP5GitHub, 2018), too.

In the present deliverable, some quality issues are gathered from the identification, proposals, and work done upon the different elements of the process. Firstly, to provide a general context we very briefly revise the European Statistics Code of Practice (ESS, 2011) commenting on the most relevant aspects which mobile phone data will probably impinge on (see section 2.2). Since we lack an end-to-end process a fully-fledged detailed analysis based on quality indicators cannot be undertaken, but some conclusions from the separate elements already identified can be supplied. The closest example to this end-to-end process carried out within this project is the outsourcing exercise practised by the ONS. This exercise provides important conclusions regarding quality (see section 2.3).

As an illustration of novel elements in the quality assessment entailed by the new methodology, in chapter 3 we have exclusively focused on the accuracy dimension for the estimates obtained using the aforementioned hierarchical model. Now, having statistical models entering into play, do not only accuracy measures like confidence intervals or coefficients of variation need to be taken into account, but also model checking and model assessment issues must be proposed and agreed upon to fully evaluate and report about the quality of the methods. We make some concrete proposals.

The last two chapters 4 and 5 articulate a structured set of conclusions for the whole project as well as relevant prospects for the future and the continuation of the development of a production framework. It is our belief that the current project has provided the opportunity for Official Statistics to move from a purely exploratory stage in which the knowledge about mobile phone data was superficial to a more mature stage where key elements for a future end-to-end process have been identified and first concrete methodological proposals have been made including the development of IT tools already with considerations on data quality incorporated in the analysis.

Some of the ongoing developments will still be active after the end of the project. A number of strategic recommendations for the future close the present document and this work package.

2

General vision of quality

Executive summary

This chapter presents an overview of general quality issues regarding the European Statistics Code of Practice ESS (2011) – CoP, hereafter– and an important conclusion regarding outsourcing the whole production of statistical analysis based on mobile phone data. We comment one by one on the main aspects on which mobile phone data will affect each of the 15 principles of the CoP. The main sources of impact arise from (i) the need to include MNOs as an early part of the statistical process during the data generation and preprocessing steps and (ii) the novel statistical methods to introduce as a consequence of the unattainability of traditional probability sampling techniques with this new data source. Finally, using ONS' experience of outsourcing the whole analysis of commuting patterns in some zones of the UK, we can provide an empirical evidence showing why this choice of incorporating mobile phone data fails to meet common data quality standards in the production of official statistics. This view is complemented with potentialities shown by Istat's experience using a limited set of Call Detail Records.

2.1. Overview

Beyond doubt, quality must be an ultimate goal in the production of official statistics. Highly relevant policy-making and decision-taking are firmly rooted upon the information provided by national and international statistical offices. Furthermore, even the increasing number of private statistical producers using both mobile phone data and more Big Data sources use this information as an important part of their production processes to calibrate their estimates. Thus, if all these new data sources are to be included in the production of official statistics, quality must be assured before their dissemination and consideration of being a final official product.

In this spirit, the ESS developed and profusely uses a so-called Quality Assurance Framework (ESSQAF, 2012) fully based upon the CoP, which provides a set of principles to be fulfilled for a statistics to meet the necessary quality standards. Regretfully, to apply

even at an experimental level this framework and to compute some of its indicators, a more or less complete production process should be in place, which is not the case with mobile phone data. The limitation in the access to data and the complexity of the end-to-end process have restricted our results to just identify, make initial proposals regarding some key elements, and develop some first prototypes. Furthermore, the adaptation itself to mobile phone data of the two-phase life-cycle model by Zhang (2012) introduced in deliverable 5.3 (see WP5.3 (2018)) already contains the identification of error sources in the production chain. A complete exercise on quality would be to analyse each of these errors. Lacking an end-to-end process with real data renders this impossible.

Nonetheless, we can already provide some initial considerations regarding the principles of the CoP (section 2.2). Moreover, as stated in the introduction, a potential final outsourced product based on mobile phone data has been analysed. We can provide some strong conclusions regarding this choice of incorporating mobile phone data in Official Statistics (section 2.3).

2.2. Mobile phone data and the European Statistics Code of Practice

The CoP provides a set of 15 principles supplying the quality standards to be met by an official statistics within the ESS. Here we comment on each principle in relation with the use of mobile phone data according to the experience gained in the present project.

2.2.1. Principle 1. Professional independence

The principle states that *professional independence of statistical authorities from other policy, regulatory or administrative departments and bodies, as well as from private sector operators, ensures the credibility of European Statistics*.

Independence now must be reinforced from two new flanks. On the one hand, the use of statistical methods not based on probability sampling and introducing the construction of statistical models and/or the choice of algorithms and their parametrisation brings the need the choose some a priori hypotheses (in the form of models, of priors, of algorithms, ...). These choices must be conducted according to this principle and with as much transparency as possible clearly making them explicit. This is especially critical in those methods used to provide information of a target population from the data at disposal (inference).

On the other hand, the inclusion of MNOs at an early stage of the production chain during data generation and data preprocessing arise as a novel element. According

to our analysis of the production process with mobile phone data in terms of the two-phase life-cycle model initiated in WP5.3 (2018), MNOs will ineludibly be part of the end-to-end process with this new data source in the data generation and preprocessing stage. Thus, both data access/use conditions and preprocessing protocols must be again designed, executed, and monitored to fully endorse this principle. Decisions by MNOs (for diverse reasons: business, operational, ...) cannot have an unsupervised influence on the production of official statistics with mobile phone data. Since data access/use is still in the air, in the future all potential agreements must take this principle into consideration.

2.2.2. Principle 2. Mandate for data collection

The principle states that *statistical authorities have a clear legal mandate to collect information for European statistical purposes. Administrations, enterprises and households, and the public at large may be compelled by law to allow access to or deliver data for European statistical purposes at the request of statistical authorities*.

The key question behind this principle in relation with mobile phone data is whether the current legislation across the ESS endorses NSIs to compel data holders by law to allow access/use for European statistical purposes. As described in deliverable WP5.2 (2017), legal issues immediately arise as one of the major obstacles for NSIs to access these data even on a research footing.

Although each national legislation shows some particularities, by and large the National Statistical Acts across the ESS provide legal support for NSIs to request access to mobile phone data (ESSTFBD, 2017). Indeed, this support is also provided by European Statistical Regulations (European Union, 2009). However, potential conflicts with telecommunication and data protection regulations need further clarifications in some cases. The experience of this project strongly suggests that the participation of national DPAs in any agreement between NSIs and MNOs can provide a satisfactory solution for all parts (NSIs, MNOS, and subscribers) so that private and public interests are met under a guaranteed respect of privacy and confidentiality.

This principle must be explicitly taken into account in any future negotiation and communication strategy clearly exhibiting (i) NSIs' rights to request and use these data for statistical public interest and (ii) our position before data holders (MNOs) not as customers but as necessary partners in the production of official statistics with this source.

2.2.3. Principle 3. Adequacy of resources

The principle states that *the resources available to statistical authorities are sufficient to meet European Statistics requirements*.

Mobile phone data, as well as most Big Data sources, entail a challenge for NSIs in terms of resources. These must be available (in terms of staff, finance, and computation) and commensurate with needs. But according to our experience novel skills, more computation capabilities, and a new relational framework with data holders will be needed to use this new data source.

These new needs require a management effort not only to fulfil them but also to cope with already existing ones. For example, we will need professional profiles of official statisticians becoming closer to data scientists thus blurring the distinction between classical statisticians and computer scientists. Certainly, training and/or recruitment programmes, depending on each NSI's framework, will have to be put into place. The role of the ESS with programmes as the ESTP (ESSESTP, 2018) and the network of masters EMOS (ESSEmos, 2018) will be progressively more important.

2.2.4. Principle 4. Commitment to quality

The principle states that *statistical authorities are committed to quality. They systematically and regularly identify strengths and weaknesses to continuously improve process and product quality*.

In producing official statistics with mobile phone data, this commitment must be clearly reinforced not only in the output statistics but also from the very beginning in the complex pre-processing and processing of data as they are generated in the telecommunication network. We remind the reader that mobile phone data are generated without statistical metadata, thus their adequation for statistical process needs this clear commitment to quality so that any processed variable satisfies a minimum level of quality for its use in the whole statistical production process. An appropriate quality policy must be defined and according to this principle made available to the public.

2.2.5. Principle 5. Statistical confidentiality

The principle states that *the privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes are absolutely guaranteed*.

Contrarily to the usual privacy and confidentiality issues raised by the MNOs to grant access to their data, this principle should be used as a clear statement that offi-

cial statistical producers ***already access, use, and process even identified personal data to produce openly disseminated statistical outputs which meets stringent privacy and confidentiality requirements***.

A novel challenge appears when sharing part of the production process with MNOs. As outlined in WP5.3 (2018), some data integration steps with official data should be carried out at diverse stages of the analysis. Guarantees must be put in place and provided so that this integration is undertaken under this principle so that no data is unduely undisclosed.

2.2.6. Principle 6. Impartiality and objectivity

The principle states that *statistical authorities develop, produce and disseminate European Statistics respecting scientific independence and in an objective, professional and transparent manner in which all users are treated equitably*.

This principle impinges on two aspects of the process with mobile phone data. On the one hand, regarding the stages of the process shared with MNOs (mainly in the access and preprocessing stages), there arises a potential conflict between NSIs' obligation to openly inform in a transparent manner about the whole process and MNOs' right to guard their industrial secrecy in their internal processes, especially, when monetising their data. Further empirical research with concrete statistical outputs must be pursued to reach an optimal agreement. MNOs' claims to protect their economic activity are legitimate and must be made compatible with public interests.

On the other hand, given the need to introduce novel methodologies where now some a priori technical decisions must be taken (choosing e.g. priors, algorithms, parameterisations, etc. in contrast to classical a priori hypothesis-free sampling designs), the exercise on transparency and scientific independence must be reinforced.

2.2.7. Principle 7. Sound methodology

The principle states that *sound methodology underpins quality statistics. This requires adequate tools, procedures and expertise*.

A difficult challenge is now posed regarding the methodology. On the one hand, generally speaking, official statistics provide information of a given target population (of households, enterprises, establishments, etc.) using information from a sample (understood in a very general way as a subset of that population of interest). With mobile phone data, the target population remains the focus of interest and not the population of mobile devices or network events per se. In this sense, we still have a sample of the

target population, a fairly large in some cases, but still a sample (from a very rigorous point of view, under classical inference models (Casella and Berger, 2002), even with data from the full population, the concept of sampling is still in the core of the analysis).

On the other hand, novel problems (thus novel outputs) can be potentially posed when using this wealth of data. Sometimes these are referred to as data-driven problems or even signal extraction problems, in which valuable information may be detected, analysed, and disseminated. However, these require a full access to statistical microdata. The reader may consult the programmes of the NetMob series of conferences (NetMob, 2017) to gain some insight on the variety of problems potentially to be tackled with mobile phone data.

In any case, traditional sampling design methodology cannot be used. In this traditional approach target variables do not need to be conceived as realizations of random variables (except for the treatment of non-sampling errors), so that analysts are free of a priori hypotheses. Novel methods (either algorithmic or model-based) do require in some form or another these assumptions. All in all, the inference framework connecting the sampled data with the target population is clearly affected and must be investigated to place this connection on firm roots.

2.2.8. Principle 8. Appropriate statistical procedures

The principle states that *appropriate statistical procedures, implemented from data collection to data validation, underpin quality statistics*.

This principle explicitly compels NSIs to use appropriate procedures along the classical survey-based production process (survey design, questionnaire design, data collection, data editing, ...).

New and/or existing procedures must be proposed or reinforced in the process with mobile phone data. For example, data integration together with data transformation from event-based objects into statistical units stands up as a fundamental procedure. The cooperation with MNOs is essential in assuring data quality and setting up an adequate statistical process.

2.2.9. Principle 9. Non-excessive burden on respondents

The principle states that *the reporting burden is proportionate to the needs of the users and is not excessive for respondents. The statistical authorities monitor the response burden and set*

targets for its reduction over time.

As a proposed element in our definition of Big Data (WP5.2, 2017; WP5.3, 2018) we underlined the fact that these new data sources share the feature that data refer to third people and not to data holders themselves. That is to say, the response burden is beared by MNOs, not by subscribers. Since we have not succeeded in having access to data (just limited for research and none for standard production), it is extremely difficult to assess how much burden is excessive and how much is not. A goal for future projects should be to research on providing a measure of this burden on MNOs analysing diverse factors such as hardware/software costs, staff, number of queries to MNOs' operational databases, ...

Also, it is important to remind that, as a public service, access to mobile phone data must be carried out evenly among all MNOs. The diversity among MNOs in the monetisation of their data poses an extra challenge in achieving the optimal conditions to access/use their data.

2.2.10. Principle 10. Cost effectiveness

The principle states that *resources are used effectively*.

This principle explicitly recognises the need to conduct "proactive efforts [...] to improve the statistical potential of administrative data [and other sources, we must say] and to limit recourse to direct surveys". Mobile phone data clearly points in this direction.

Also, in the quest for standardised solutions increasing effectiveness and efficiency, the ESS is now facing the opportunity to provide this kind of solutions when developing a completely new production framework for the use of mobile phone data in official statistical production. From our experience in this project, in our view, the development of these tools requires intensive research efforts which are not attainable for any NSI operating in its own. The ESS as a whole has the opportunity to provide this supporting framework for its members.

2.2.11. Principle 11. Relevance

The principle states that *European Statistics meet the needs of users*.

When considering traditional statistics carried out through classical survey methods, some limitations are usually pointed out by users, especially regarding both time and space breakdowns of the statistical outputs. Regarding the former, this is considered

in the principle 13 on timeliness and punctuality. Regarding the latter, it is widely known how sample sizes pose a limitation on the degree of territorial breakdown (under standard accuracy conditions) of the statistical outputs. Indeed, the whole discipline of small area estimation (Rao and Molina, 2015) is an illustration of this fact.

Mobile phone data stand up as a potential solution to reach unprecedented geographical scales in providing information. However, two challenges appear. Firstly, statistical disclosure control and the risk of identifiability in low-density zone is present if highly disseminated information is to be provided. Secondly, a full identification of concrete statistical outputs of interest must be conducted in collaboration with stakeholders. Tourism in its many forms –outbound, inbound, resident– mobility (commuting patterns), and many others are rather clear, but more statistical domains remain to be explored.

2.2.12. Principle 12. Accuracy and reliability

The principle states that *European Statistics accurately and reliably portray reality*.

This principle underlines the importance of producing accurate and reliable outputs with an emphasis in sampling errors (also non-sampling). Now novel methods are in place and the analysis of errors along the process must keep its central role. The adaptation of the two-phase life-cycle model introduced in WP5.3 (2018) does not only provide a way to represent the process in statistical abstract terms but also incorporates the identification of errors along the process so that quality is at the very core of the model.

Furthermore, again motivated by novel methods making use of statistical modes, model checking and model assessment must be an integral part of the accuracy and reliability assessment of the process. In the forthcoming chapter 3, we provide some indicators in this direction.

2.2.13. Principle 13. Timeliness and punctuality

The principle states that *European Statistics are released in a timely and punctual manner*.

This is strongly connected with our comment on principle 11 above. In some traditional structural statistics, data for a year of reference are collected and processed during the next natural year so that the dissemination of results takes place up to two years after the phenomenon has occurred. Even after the efforts by NSIs to shorten this period with traditional data sources, this is too long for some policy-making and decision-making in

many scenarios.

Mobile phone data, since they are digital data instantly processed by MNOs, provide an opportunity to reach even real-time dissemination. This has not been researched in the current project since a strong agreement is needed with MNOs. The trade-off between processing burden and users' needs are to be explored to reach a balanced agreement.

2.2.14. Principle 14. Coherence and comparability

The principle states that *European Statistics are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different sources*.

This principle points in the direction of the standardisation of official statistical production across the ESS. With mobile phone data, where an end-to-end process is still lacking, this is indeed an opportunity. In the current project several key elements for a production framework (including methodology, IT tools, and quality issues) have been identified and some initial proposals have been formulated. The need for standard definitions and statistical metadata for mobile phone data is already apparent.

2.2.15. Principle 15. Accessibility and clarity

The principle states that *European Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance*.

The dissemination stage is the clear object of this principle. Since no concrete statistical output has been obtained, dissemination has not been tackled in the project. However, some comments can already be made. Depending on the degree of both time and space breakdown of outputs, new dissemination techniques will need to be put in place (interactive visualization with movies, interactive maps, etc.).

Custom-designed analysis can potentially enter into conflict with the private interests of MNOs to sell their own products. Probably a joint collaboration in this line would be beneficial for both parts. Access to microdata for research should be evaluated in terms of privacy and confidentiality issues, especially regarding the risk of identifiability of subscribers. Probably, a minimal degree of aggregation might not be surpassed.

Finally, maximal transparency regarding the policy of access and use of data as well as the methodology must be also followed.

2.3. Dealing with fully processed mobile phone data

The closest example to an end-to-end process with mobile phone data supplying statistics for potential dissemination and use by stakeholders has been carried out separately by the ONS by outsourcing the whole analysis of commuting patterns in some zones of the UK and by Istat using a limited set of Call Detail Records duly aggregated according to the methodology presented in WP5.3 (2018). Important conclusions can be drawn from both experiences in terms of the quality of these potential statistics.

2.3.1. ONS experience of outsourcing

The ONS sourced, via public tender, two small samples of aggregated commuter flows. Each sample was fully modelled by a mobile network operator (MNO) using four weeks of geo-location traces of mobiles subscribed to that MNO's network. The mobile-phone-data-flows (MPD-flows) were specified to be equivalent to data as is typically produced in the decennial Census (known as Travel to Work (TTW) data). The objective for ONS was to compare these MPD-flows directly with TTW.

In common with administrative datasets, mobile data is generated for mobile telephony operations and not for statistical use. There were difficulties in using MPD within commuter flows including alignment to the various TTW concepts such as population base, definitions for a worker, usual residence and usual workplace etc. Although some metadata was provided, the detail was not sufficient to be able to make a considered evaluation of the quality of the modelled MPD-flows.

As an input data-source, minimal information on MPD quality was provided to ONS although the two data suppliers have developed basic checks such as confirming that there were no problems with the network for the days and study areas used in the analysis. Mobile traces that do not have enough geolocation data over the period to determine patterns of movement are also excluded, however the number of mobiles removed, what types of subscriber they may be and how results may be affected was not provided. The modelled commuter flows were subject to quality assurance checks by both the data suppliers and one supplier provided ONS with their checks and results.

The algorithms to produce commuter flows are more complex than those required to generate population densities. Commuter flows, as modelled by the three main MNOs in the UK, require observation of repeat daily return journeys from an inferred residence location to another location. Additional criteria are introduced, such as requiring the

"stay" periods at the second location to be of a sufficient duration to viably infer it as a workplace. A period of 4 to 6 weeks is chosen so that repeat journeys might be required to be observed 2 or 3 times per week on average.

Clearly, there are multiple parameters and assumptions within the modelling of commuter flows and it would be necessary for ONS to understand and test them. Sourcing fully modelled and aggregated estimates from a data supplier presents problems with the required level of transparency. Although data suppliers are willing to give high level overviews for their methodologies, they have made large investments into their research and are naturally reluctant to give out commercially sensitive information that might be made public. This made it difficult to understand the various stages of the modelling and the necessary quality measures needed to verify each stage.

The ONS comparison with TTW data revealed quality issues for MPD-flows into rural areas and for short distance commutes. These issues are most likely due to the use of cell towers as the measure of geo-location. Besides complex algorithms to first determine the approximate location of a mobile (usually a cell area or best service area), there is a second mapping algorithm of this approximate location into the standard official geography required for reporting: defined by ONS as the Middle Layer Super Output Area, a statistical geography containing a resident population of around 7000. It was not possible to investigate the mapping algorithms as the MPD-flows were too heavily modelled.

It is difficult to specify how a less aggregated set of data, suitable for more adaption within ONS, might have been sourced. The MNOs analyse the geo-location traces at the cell area level and it might appear sensible to ask for commuter flows based on this level of geography. However, the use of a disclosure threshold, below which no MPD-flow is released, was already seen to remove many of the small flows at MSOA to MSOA level. At cell area to cell area level the impact of this threshold would be more destructive.

In future research partnerships with mobile networks it might be more practical to operate a call off arrangement for ad-hoc analysis of MPD on a case by case basis so that different methods might be tested. For ONS, this would require the submission of a public tender with clear requirements that no payment for data were involved. Closer collaboration with the data suppliers would be useful to identify the quality measures required for inputs and outputs, as this information might be made available.

2.3.2. The Istat experience[1]

The Italian National Institute of Statistics, Istat, accessed to anonymised MPD at disaggregated level, thanks to a strong and fruitful partnership with an Italian MNO. Istat accessed to CDRs for a province (the province of Pisa, composed by 37 municipality=LAU) for 6 weeks, from the first of January 2017 to the 12nd of February 2017. This gives Istat the opportunity to apply some analysis to measure the quality of input data, as well as the quality of the expected outputs, i.e. population estimates at LAU level and mobility pattern within the province.

2.3.2.1. Investigating the quality of the input

As far as the investigation of the input quality, the analysis was firstly concentrated on the pattern of calls and text messages in the reference period. The period between two public holidays (New Year Eve and Epiphany) is considered as festive in Italy and the trend is irregular compared to the following weeks. During week-days the trend is regular with a deep drop over in the weekend. This trend is also documented in the telephone traffic of other countries (de Jonge et al., 2012; Douglass et al., 2015; Furletti et al., 2017).

This analysis discovered the regularity of patterns for calls and text messages, as regards weekdays and weekend days, as well as the daytime and nighttime. Reassuringly, these results have been already observed by the literature in several countries.

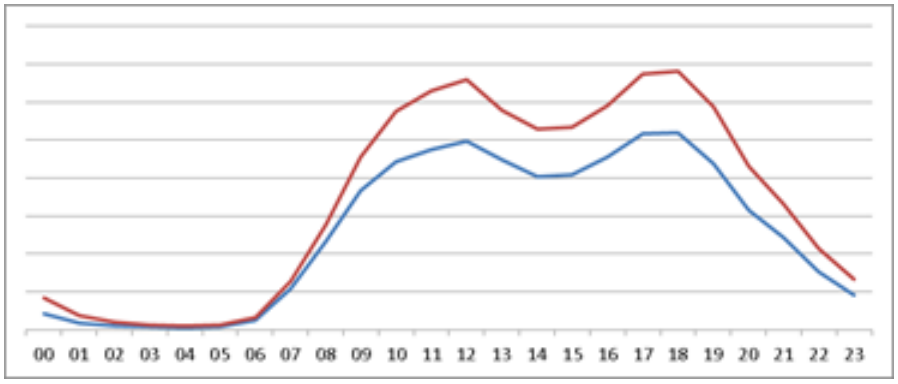

Figure 2.1 Number of speech activity in the working days (in red) and at the weekend (in blue).

Moreover, in this way we identified an anomalous data supply in a single day (the 31 of January) for couple of municipalities, which caused the anomalous peak

[1]Prepared in collaboration with Fabrizio de Fausti and Luca Valentino.

highlighted in the following figure 2.2. In this case we noted that the number of calling SIM was within the expectation of other periods whilst the number of CDRs was anomalously high. After investigating that there was not the occurrence of special events (e.g. Christmas night, popstar concert, massive flash mob...) which affect the PD users behavior, further analyses showed that some data had been sent erroneously twice.

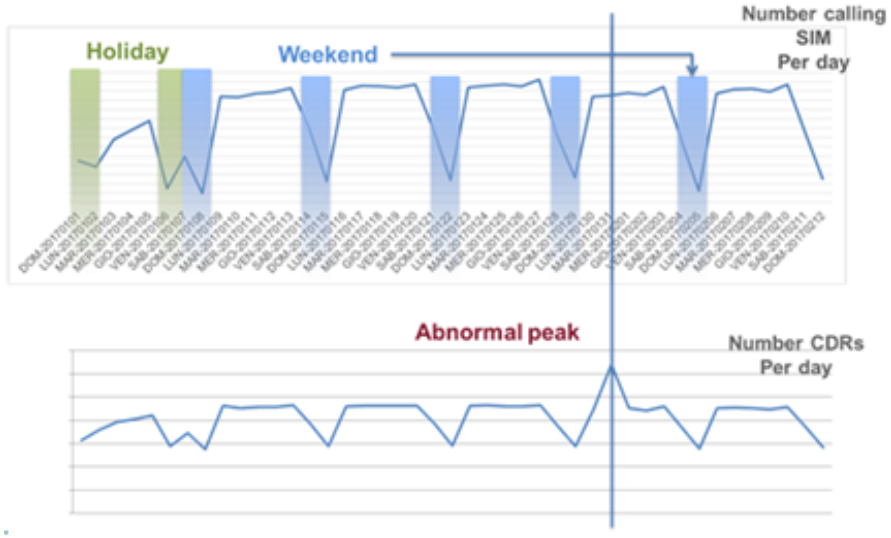

Figure 2.2 Trend graph comparison for number calling SIM per day and number CDRs (Calls and SMS) per day.

Furthermore, the analysis of the graph suggested excluding from the subsequent calculation the data for the first week, as they regard to a holiday time in Italy, and this will not be valuable information for the population and mobility pattern estimates.

At individual level, some SIMs had made a single call in the observed period. These CDRs were excluded from the subsequent analysis, because they represent possible so called "passing-by" (Furletti et al., 2017), e.g. a person who calls while driving on the highway or when arriving at the airport or train station and he/she is departing to another place. Furthermore these analyses showed that many SIMs call few times and few SIMs call a lot of times.

2.3.2.2. Investigating the quality of the output

After the check of the supplied set of data as a whole, we moved to the analysis of MPD on the territory, at the level of municipality. Firstly, we analysed the MPD coverage of the 37 municipalities in the province of Pisa. We observed that the territories of 5 municipalities don't have antennas, they are very small in terms of population counts; moreover, we did not receive data for 4 municipalities, even though having antennas. These 4 municipalities were deleted from the subsequent analysis.

In order to evaluate the output quality, we made use of auxiliary information regarding the data of: resident population, land use, the presence of POIs (Point of Interest: University, Industrial Poles or site, Monuments or touristic place) and population density concentration in urban localities (LOC). It is worth noting that these information are owned by the NSIs (some of them are publically available).

To investigate the correlation of MPD with official population figures, we firstly concentrated the analyses on the nighttime population. The approximation of residential population with nighttime mobile phone users has been stated in several works (Kang et al., 2012; Deville et al., 2014; Douglass et al., 2015). In this work, we identify mobile phone users with SIMs.

The nighttime population per municipality has been calculated with different techniques for call location:

- Best Service Area (BSA[2]):
  This type of localization allocates the calls proportionally to the BSA area within the area of the municipality;
- The Voronoi tessellation[3];
- The joint use of BSA and UAM (Urban Area of the Municipality):
  This type of localization allocates the calls proportionally to the BSA area within only to the urban area of the municipality;
- The antennas position:
  In this case, the localization is given by the position of the tower antenna.

Figure 2.3 shows a scatter plot of the counts of nighttime active SIMs versus the January 2017 residential population estimates for the province of Pisa at municipality level; the counts are obtained by the abovementioned different types of localization.

[2]See section 3.2.2 of deliverable WP5.3 (2018).
[3]See section 3 of deliverable WP5.3 (2018).

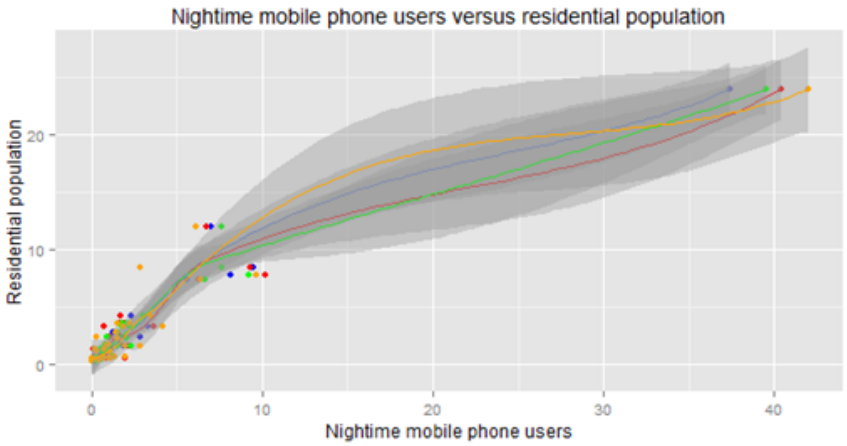

Figure 2.3 Scatter plot of nighttime mobile phone users versus residential population (municipalities in Pisa province) with Pisa city.

It shows that there is a reasonable good relationship that was approximately linear as depicted by the LOESS regression interpolation. The different colors represent the techniques of localization, in particular:

- blue represents the BSA;
- red represents the Voronoi;
- green represents the BSA and UAM;
- orange represents the antennas position.

Figure 2.4 shows the same scatter plot excluding from the analysis the municipality of Pisa.

The R-squared of the linear regression interpolation are provided in table 2.1, proving the overall goodness of the model in predicting residential population via the nighttime mobile phone users, obtained according to the different localization techniques. The results are reported both excluding from the analysis the extreme value represented by the city of Pisa and including it.

The high correlation between phone users and residential population is also proved when the analysis is focused on the SIMs active during the day. We used the phone users' population identified by these SIMs as predictor of the residential population, in

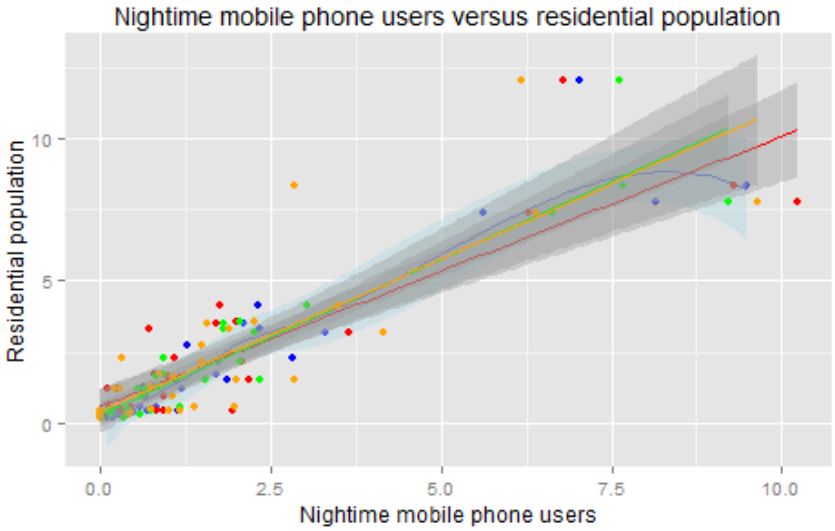

Figure 2.4 - Scatter plot of nighttime mobile phone users versus residential population (municipalities in Pisa province) without Pisa city.

this case the SIMs are assigned to the municipality resulting from the nighttime position.

Type of localization	R^2 with Pisa	R^2 without Pisa
BSA (Blue)	0.8967	0.8564
Voronoi (Red)	0.8816	0.7767
BSA and UAM (Green)	0.8951	0.8761
Antennas position (Orange)	0.8305	0.6674

Table 2.1 Regression parameters analysis.

These results are considered satisfactory and encourage us in using CDR to estimate something that cannot be actually observed with sample survey and administrative data, but which could be a new output within official statistics. In fact, the MPD allow us to identify areas that might be problematic for census counts, for instance, areas at risk of over or under coverage can be identified by comparing population estimates coming from MPD with the counts of people enrolled in registers. The risk of over/under-coverage can be defined at a very small scale and this information can be used both at a sample stage, when designing the coverage sample survey, and at the estimation stage,

when small area population estimates have to be provided.

3

Accuracy and model checking

Executive summary

This chapter addresses the accuracy dimension of data quality. This is a fully methodological chapter complementing the hierarchical model proposed in deliverable WP5.3 (2018) to make inferences from the aggregated mobile phone data. Here, the focus is now placed on quality measures and, in particular, on measures of accuracy for the population count estimates obtained with this model. As a novel ingredient in traditional quality assessment, since we make use of statistical models, measures for model checking must be introduced in connection with principles of the CoP commented in the preceding chapter. We make some concrete proposals, which are indeed already implemented in our prototyping R package `pestim` to provide a proof-of-concept.

This chapter is devoted to the accuracy dimension of data quality. In particular, we shall focus on the accuracy of the estimates of the population counts using aggregated mobile phone data. Since we proposed novel elements for a methodological framework beyond the design-based inference model, an effort must be made to clearly explain how the accuracy is now dealt with.

By and large, we shall firstly concentrate on how the concepts behind traditional coefficients of variation and confidence intervals should now be approached. Secondly, given that a statistical model is at the core of our proposal it is compulsory to deal with the goodness of fit. As briefly argued in WP5.3 (2018), model-based inference can provide more accurate estimates *if the model is correct*. Thus, model assessment must be ineludibly tackled when producing estimates according to this methodology. We provide some first elements to assess the hierarchical model proposed to estimate population counts.

3.1. Credible intervals and posterior coefficients of variation

3.1.1. Credible intervals

In probability sampling the construction of confidence intervals is a usual technique attached to accuracy assessment of point estimations. Given a design-based estimator $\widehat{Y}_{U_d}$ of a population total $Y_{U_d} = \sum_{k \in U_d} y_k$ in a population domain $U_d \subset U$, under the usual assumed normality conditions, the confidence interval for Y_{U_d} of confidence level $1 - \alpha$ is given by (Särndal et al., 1992)

$$I_\alpha \left(\widehat{Y}_{U_d} \right) = \left[\widehat{Y}_{U_d} - z_{1-\alpha/2} \sqrt{\mathbb{V} \left[\widehat{Y}_{U_d} \right]}, \widehat{Y}_{U_d} + z_{1-\alpha/2} \sqrt{\mathbb{V} \left[\widehat{Y}_{U_d} \right]} \right].$$

As usual, the variance must be estimated so that we report the following approximate confidence interval

$$\hat{I}_{1-\alpha} \left(\widehat{Y}_{U_d} \right) = \left[\widehat{Y}_{U_d} - z_{1-\alpha/2} \sqrt{\widehat{\mathbb{V}} \left[\widehat{Y}_{U_d} \right]}, \widehat{Y}_{U_d} + z_{1-\alpha/2} \sqrt{\widehat{\mathbb{V}} \left[\widehat{Y}_{U_d} \right]} \right]. \tag{3.1}$$

Since we adopted the Bayesian approach in our proposal (WP5.3, 2018), we can provide a natural counterpart by constructing the so-called *credible intervals* (Gelman et al., 2013). Notice that, contrarily to usual confidence intervals, a credible interval provides a range of values such that the true parameter under estimation (the population count N_i in cell i in our case) belongs to this range with the desired probability. Despite this important subtlety in the meaning, we believe it provides an accuracy assessment as appropriate as traditional confidence intervals for official statistics.

There exist at least three possible ways to construct these credible intervals:

- We can construct an interval for which the mean (assumed to exist, as always in our case) lies in the centre of the interval.
- We can also construct the so-called *equal-tailed intervals*, in which the probability of lying outside the interval is equal both below and above the interval. These intervals will certainly contain the median.
- We can construct the so-called *highest posterior density intervals*, which are the narrowest possible intervals. For unimodal distributions, these intervals will contain the mode.

We illustrate the construction of these intervals with an example for a single cell (the construction would be identical in the rest of cells). Taking $N^{\text{MNO}} = 19$, $N^{\text{Reg}} = 97$, and the following priors for u, v, and λ:

1. $u \simeq \text{Unif}(\text{xMin} = 0.95 \cdot N^{\text{MNO}}/N^{\text{Reg}}, \text{xMax} = 1.05 \cdot N^{\text{MNO}}/N^{\text{Reg}})$;
2. $v \simeq \text{Triang}(\text{xMin} = 0.95 \cdot N^{\text{Reg}}, \text{xMax} = 1.05 \cdot N^{\text{Reg}}, \text{xMode} = N^{\text{Reg}})$;
3. $\lambda \simeq \text{Gamma}(\text{shape} = 1 + \alpha, \text{scale} = N^{\text{Reg}}/\alpha)$, with $\alpha = 10$;

we arrive at figure 3.1.

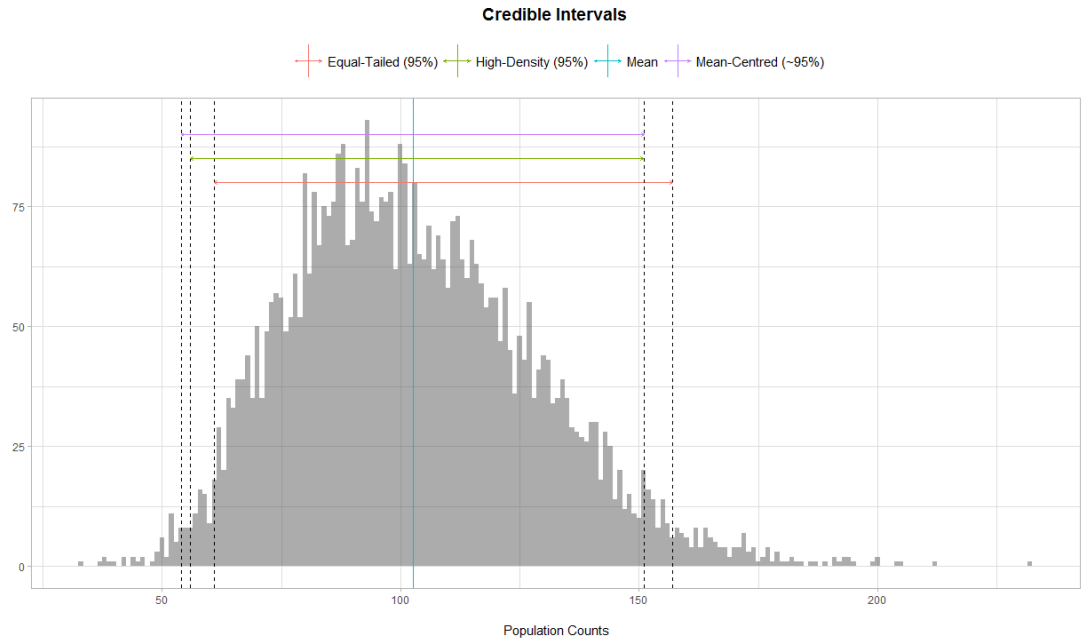

Figure 3.1 Credible intervals for $N^{\text{MNO}} = 19$, $N^{\text{Reg}} = 97$ and weakly informative priors.

3.1.2. Coefficients of variation

In probability sampling the accuracy of estimates is also traditionally assessed through coefficients of variation. Given a design-based estimator $\widehat{Y}_{U_d}$ of a population total $Y_{U_d} = \sum_{k \in U_d} y_k$ in a population domain $U_d \subset U$, the coefficient of variation of $\widehat{Y}_{U_d}$ is defined as

$$\text{cv}\left(\widehat{Y}_{U_d}\right) = \frac{\sqrt{\mathbb{V}\left[\widehat{Y}_{U_d}\right]}}{\mathbb{E}\left[\widehat{Y}_{U_d}\right]},$$

where the expectation and variance are taken over the sampling design $p(\cdot)$. The coefficient of variation is estimated using the fact that $\widehat{Y}_{U_d}$ is (asymptotically unbiased):

$$\widehat{\text{cv}}\left(\widehat{Y}_{U_d}\right) = \frac{\sqrt{\widehat{\mathbb{V}}\left[\widehat{Y}_{U_d}\right]}}{\widehat{Y}_{U_d}}, \tag{3.2}$$

where $\widehat{\mathbb{V}}\left[\widehat{Y}_{U_d}\right]$ is also (asymptotically) unbiased.

In our proposal to estimate population counts we can naturally produce a similar figure of merit to assess the accuracy of a point estimation. All we need is a point estimator together with a measure of dispersion.

In this sense, let us briefly remind that, once we have the posterior probability function for the number of individuals $\mathbb{P}\left(N_i \middle| N_i^{\text{MNO}}; N_i^{\text{Reg}}\right)$, we can produce different point estimations. In WP5.3 (2018) we have proposed to use the posterior mean, the posterior median, and the posterior mode as potential choices[1].

Now we must associate some measures of dispersion to these point estimators. Regarding the posterior mean, along with the original definition in design-based inference, we can focus on the posterior variance, so that a natural extension of the definition for the coefficient of variance would be a similar expression to (3.2) although using posterior versions of these moments:

$$\text{cv}^{(1)}\left(\widehat{N}_i^{\text{mean}}\right) = \frac{\sqrt{\mathbb{V}\left[N_i\right]}}{\mathbb{E}\left[N_i\right]}, \tag{3.3}$$

where the moments are computed with respect to the posterior distribution.

For the posterior median, a natural dispersion measure can be the interquartile range IQR of the posterior distribution. Thus, we can compute

$$\text{cv}^{(2)}\left(\widehat{N}_i^{\text{median}}\right) = \frac{\text{IQR}\left[N_i\right]}{\text{Med}\left[N_i\right]}. \tag{3.4}$$

Regarding the posterior mode, as explained in the foregoing section, for our unimodal posterior distribution the high-density credible interval always contains the mode. Thus, it seems to us natural to associate the length of this interval as a measure of dispersion to construct a mode-based coefficient of variation. We propose to compute:

$$\text{cv}^{(3)}\left(\widehat{N}_i^{\text{mode}}\right) = \frac{u^{\text{hdi}}\left[N_i\right] - l^{\text{hdi}}\left[N_i\right]}{\text{Mode}\left[N_i\right]}, \tag{3.5}$$

[1] All these three choices are included in the `pestim` package (WP5.4, 2018).

where u^{hdi} and l^{hdi} stand for the upper and lower linits of the high-density credible interval.

We can illustrate the computation of these coefficients of variation with the same example as above. To get a numerical reference of the final computed coefficients of variation in each case (mean, median, mode) let us firstly compute the coefficients of variation of the prior distributions according to the same expressions:

Prior	Mean	Median	Mode
f_u	2.9	5.0	5.0
f_v	2.0	2.9	2.9
f_λ	30.2	41.1^*	43.1^*
N	25.2	35.0	31.3

Table 3.1 Coefficients of variation of population count estimates in a cell with $N^{\text{MNO}} = 19$, $N^{\text{Reg}} = 97$ and weakly informative priors.

* computed empirically.

So far we have dropped both the cell and time dependence for the population count estimates $N_i(t)$. Since the model assumes the estimates to be independent in each cell, the inclusion of the cell dependence is trivial. Regarding the time dependence, we notice that the computation is carried out using the random values generated by the posterior distribution $\mathbb{P}(N_i(t_k)\big|N_i^{\text{MNO}}; N^{\text{Reg}})$, for $k = 0, 1, 2, \ldots$. Thus it is also trivial to incorporate the time dependence. The R package `pestim` already incorporates these routines to return these coefficients of variation.

Finally, these considerations on accuracy can be further illustrated by the application of these computations to the whole simulated population included in the R package `pestim` (WP5.4, 2018). The results are represented in figure 3.2.

This figure has been computed with the following weakly informative priors:

- In each cell i, f_{u_i} is chosen to be a uniform distribution with interval center at $N^{\text{MNO}_i}/N^{\text{Reg}_i}$ and interval radius $\pm 0.15 \times N^{\text{MNO}_i}/N^{\text{Reg}_i}$.
- In each cell i, f_{v_i} is chosen to be a uniform distribution with interval center at N^{Reg_i} and interval radius $\pm 0.15 \times N^{\text{Reg}_i}$.
- In each cell i, f_{λ_i} is chosen to be a gamma distribution with shape parameter $1 + \alpha$ and scale parameter N^{Reg}/α, where $\alpha = 1/(cv^2 - 1)$ (cv stands for the coefficient of variation).

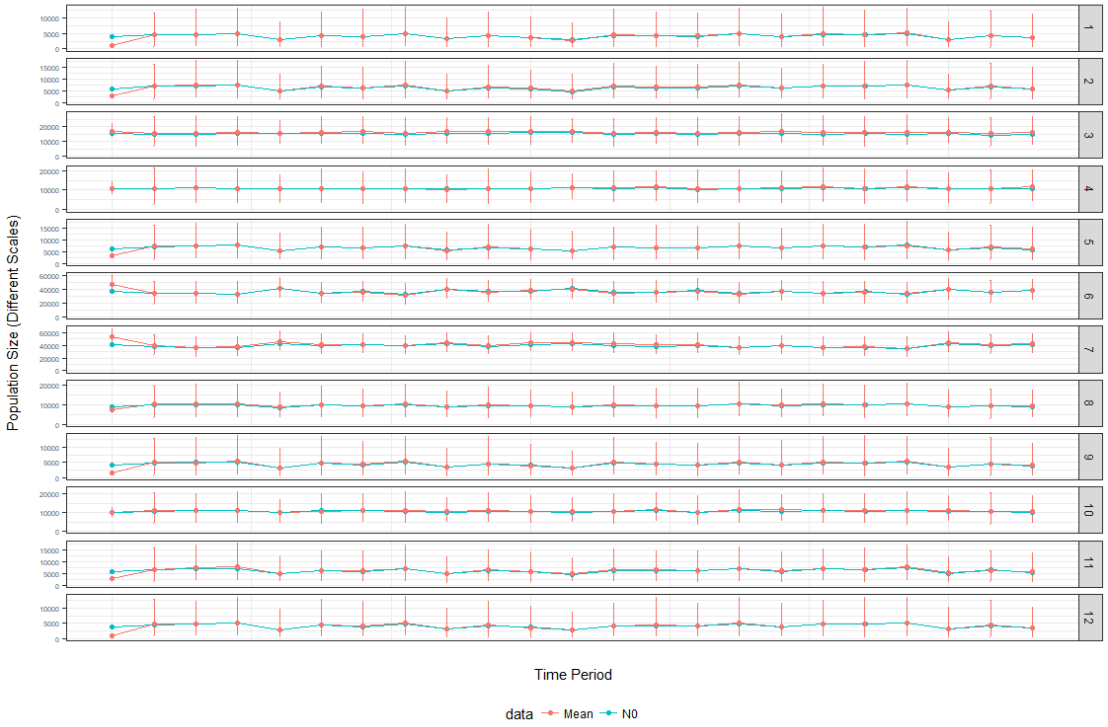

Figure 3.2 Estimates of population counts with their 95% equal-tailed credible intervals in the simulated population included in the `pestim` package. A selection of time periods is shown for ease of visualisation.

- For the transition probability matrices, we also choose uniform distributions with coefficients of variation of 10% for each cell.

The length of the credible intervals is justified (apart from the 95% degree) by the prior uncertainty in the choice of the foregoing distributions, especially for the transition probability matrices.

3.2. Model checking

In our deliverable on methodology (WP5.3, 2018) we argued why the traditional probability sampling setting was not attainable with mobile phone data and novel methods must be used to produce estimates. We set up a hierarchical model with this purpose. When using statistical models we must clearly assess the goodness of fit, especially when producing official statistics which, in consonance with the European Statistics Code of

Practice (ESS, 2011), should be as independent as possible of a priori hypotheses of any kind.

In the foregoing section some measures on the accuracy of the output estimates were proposed. Here we concentrate on the goodness of fit of the model to the data. Following Gelman et al. (2013) we do not aim at finding out whether our model is true or false (*all models are essentially wrong*, Box dixit), but at assessing how much *the model's deficiencies have a noticeable effect on the [...] inferences* (Gelman et al., 2013).

There exist plenty of possibilities to assess the fit of the model. However, following the line of the preceding section, we shall propose simple indicators providing compact information about the goodness of fit. The central element to produce these indicators will be the posterior predictive distribution so that we draw simulated values from this distribution and compare these samples with the observed data. The bottom line is that, if the model fits the data, there should be little differences between the replicated and input data.

We formalise this line of reasoning. Let us concentrate for the time being on the inference exercise at the initial time t_0. Let us denote by $N^{\text{MNO, rep}}$ the replicated number of individuals detected by the network. We need to compute the following posterior predictive distribution

$$\mathbb{P}\left(N^{\text{MNO, rep}}\middle|N^{\text{MNO}}\right) = \int_{\Omega_{u,v,\lambda}} \mathbb{P}\left(N^{\text{MNO, rep}}\middle|u,v,\lambda\right)\mathbb{P}\left(u,v,\lambda\middle|N^{\text{MNO}}\right) dudvd\lambda, \quad (3.6)$$

where we have dropped out both the subscripts denoting the cell (since they are independent) and the time dependence and where N^{MNO} stands for the actual input data in the hierarchical model.

We propose to use the following set of indicators for the model checking:

- Bias indicator:

$$\mathsf{b} = \mathbb{E}\left(N^{\text{MNO, rep}} - N^{\text{MNO}}\middle|N^{\text{MNO}}\right), \quad (3.7a)$$

  Notice that this measure is intended to quantify the difference on average between the replicated data and the actual input data. An associated measure is its relative counterpart:

$$\tilde{\mathsf{b}} = \mathbb{E}\left(\frac{N^{\text{MNO, rep}} - N^{\text{MNO}}}{N^{\text{MNO}}}\middle|N^{\text{MNO}}\right). \quad (3.7b)$$

- Variance indicator:

$$\mathsf{v} = \mathbb{V}\left(N^{\text{MNO, rep}} - N^{\text{MNO}} \middle| N^{\text{MNO}}\right), \tag{3.7c}$$

Notice that this measure is intended to quantify the variability of the replicated data around its average value. An associated measure is its relative counterpart:

$$\tilde{\mathsf{v}} = \mathbb{V}\left(\frac{N^{\text{MNO, rep}} - N^{\text{MNO}}}{N^{\text{MNO}}} \middle| N^{\text{MNO}}\right). \tag{3.7d}$$

- Mean square error estimator:

$$\mathsf{mse} = \mathbb{E}\left[\left(N^{\text{MNO, rep}} - N^{\text{MNO}}\right)^2 \middle| N^{\text{MNO}}\right], \tag{3.7e}$$

Notice that this measure is intended to quantify the variability of the replicated data around its true value. An associated measure is its relative counterpart:

$$\widetilde{\mathsf{mse}} = \mathbb{E}\left[\left(\frac{N^{\text{MNO, rep}} - N^{\text{MNO}}}{N^{\text{MNO}}}\right)^2 \middle| N^{\text{MNO}}\right], \tag{3.7f}$$

In all cases, the moments are taken with respect to the posterior predictive distribution (3.6). The computation is carried out generating random values according to this distribution (see appendix A).

An application to the same example used in the preceding sample for a single cell with $N^{\text{MNO}} = 19$, $N^{\text{Reg}} = 97$ and weakly informative priors on page 25 yields the results in table 3.2. As the reader can observe from this table, the model reproduces the input data with an expected bias of ± 1 individuals and a (rounded) deviation of ± 8 individuals.

	b	v	mse
absolute	1.1	58.3	59.4
relative (%)	5.6	16.1	16.5

Table 3.2 Model-checking indicators for $N^{\text{MNO}} = 19$, $N^{\text{Reg}} = 97$ and weakly informative priors.

The computation of the indicators (3.7a) to (3.7f) is undertaken with the same Monte Carlo techniques used for the computation of estimates (WP5.3, 2018). This is explained

in detail in the appendix A.

The inclusion of the cell dependence is again trivial, since they are independent. In table 3.3 we report these (relative) indicators for each cell $i = 1, \ldots, 12$ in the simulated population included in the R package `pestim`.

Cell	$\tilde{\mathsf{b}}$	$\tilde{\mathsf{v}}$	$\widetilde{\mathsf{mse}}$
1	4.9	10.2	10.4
2	7.7	25.6	26.1
3	2.5	6.0	6.0
4	5.7	13.4	13.7
5	7.5	31.2	31.7
6	7.3	21.1	21.7
7	5.1	12.0	12.3
8	4.4	11.8	12.0
9	3.1	6.3	6.3
10	3.7	7.2	7.4
11	6.1	29.7	30.1
12	3.5	12.3	12.4

Table 3.3 Model-checking relative indicators (in percentage) for the simulated population included in the R package `pestim`.

Notice that having a more or less low bias, the model generates data with a noticeable variability in some cases[2].

Introducing the time dimension to assess the model checking is slightly subtler. Notice that to apply the former approach we need to replicate the input data according to the model. When considering the time dimension, the input data in the model are the transition matrices $\pi^{\text{MNO}}(t_0, t_n) = \left[\frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}\right]_{1\leq i,j\leq I}$, where $N_{i\cdot}^{\text{MNO}}(t_0) = \sum_{j=1}^{I} N_{ij}^{\text{MNO}}(t_0, t_n)$. These are used to estimate a probability transition matrix with rows $\mathbf{p}_i(t_0, t_n)$ which are then used together with the initial time estimates $\mathbf{N}(t_0)$ to produce the population count estimates in all cells $\mathbf{N}(t_n) = \mathbf{N}(t_0) \cdot p(t_0, t_n)$ (up to the closest-integer function – see WP5.3 (2018)).

This means that to follow our approach we need to assess the model fit to these transition probabilities $\pi^{\text{MNO}}(t_0, t_n)$. We will then focus on assessing how close the

[2]We must state that there some numerical issues behind these computations which are still under analysis. These results should be taken with care.

replicated samples of the matrix $p(t_0, t_n)$ are to the input data $\pi^{\text{MNO}}(t_0, t_n)$. Thus, working with rows, we write as above

$$\mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n)\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right) = \int_{\Omega_{\boldsymbol{\alpha}_i}} \mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n)\big|\boldsymbol{\alpha}_i(t_0,t_n)\right)\mathbb{P}\left(\boldsymbol{\alpha}_i(t_0,t_n)\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right) d\boldsymbol{\alpha}_i(t_0,t_n) \tag{3.8}$$

The set of indicators for each cell i runs parallel to those proposed above (eqs.(3.7a) to (3.7f)). They are provided in the appendix A.

As an illustrative example, let us use the simulated population included in the package `pestim`. As priors for the transition matrices, we set uniform distributions with coefficients of variation equal to 0.10 for the 12 cells. The relative indicators for a subset of the whole span of time periods are represented in figures 3.3, A.1 (in appendix for readibility's sake), A.2 (in appendix for readibility's sake).

As the reader can immediately observe, the generation of input data is fairly good enough, but there exist outlying cell-to-cell transition probabilities in all time periods. The main reason behind these outliers lies on those small probabilities (very close to zero) for which relative measures are highly sensitive to small variations.

Having assessed the initial time estimation part of the model, on the one hand, and the transition matrices estimation part, on the other hand, we can now also assess both elements in conjunction. To do this, let us further assume the relation

$$\mathbf{N}^{\text{MNO}}(t_n) = \mathbf{N}^{\text{MNO}}(t_0) \cdot p(t_0, t_n). \tag{3.9}$$

Then we can also generate replicated samples $\mathbf{N}^{\text{MNO, rep}}(t_n)$ and compare them with the actual input data $\mathbf{N}^{\text{MNO}}(t_n)$.

In figures 3.4 the reader can observe how the number of individuals according to the network are generated according both to the model and the input data.

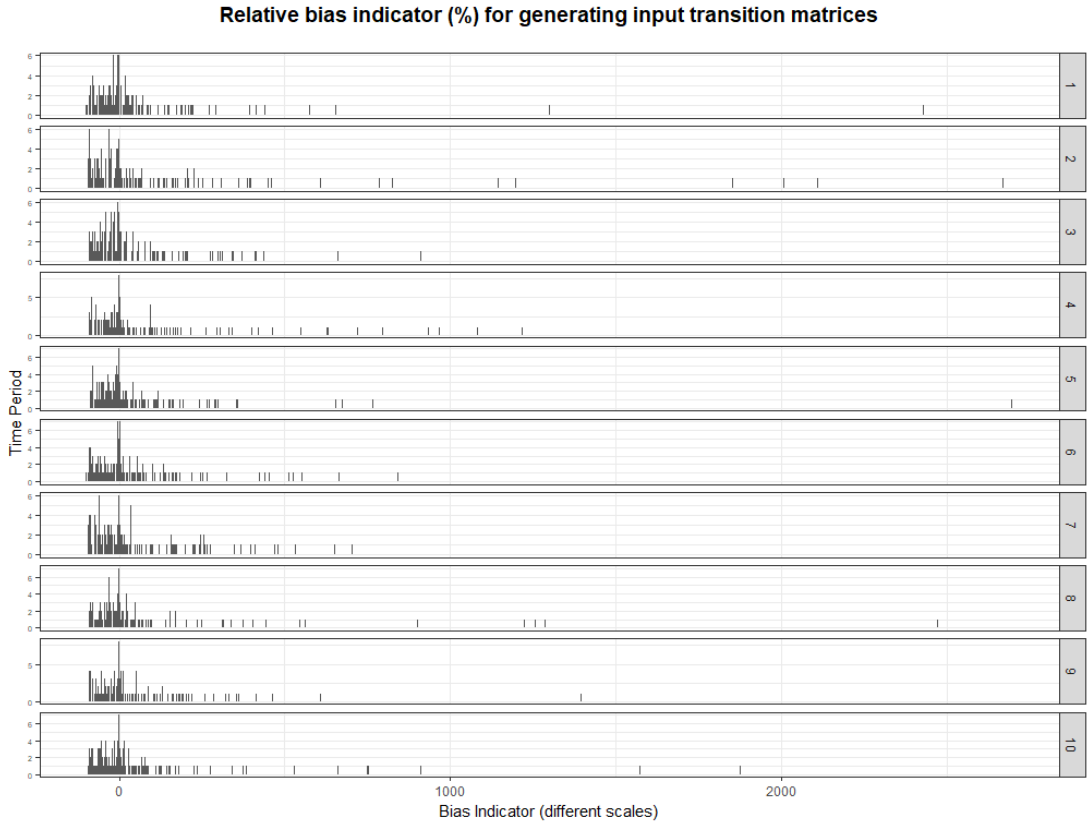

Figure 3.3 Relative bias indicators (in percentage) for uniform priors with coefficients of variation equal to 0.10 for all cells.

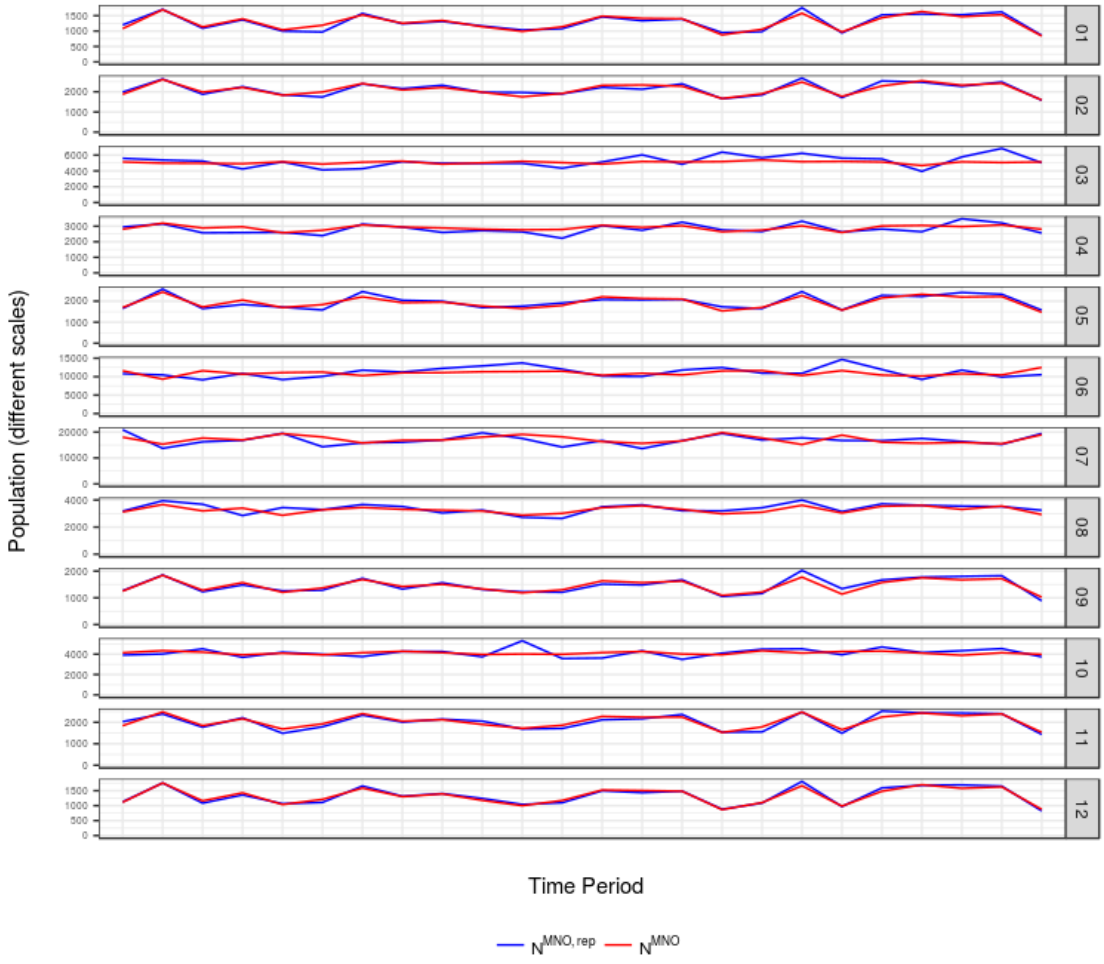

Figure 3.4 Comparison of estimated replicated and actual data for the number of individuals according to the network.

4

Conclusions

Executive summary

In an executive fashion we underline the main conclusions as a bullet list:

- Access
  - Raw telecommunication data.- No access whatsoever has been granted. Highly disruptive and requiring completely novel skills for official statisticians. A close collaboration between NSIs and MNOs needed.
  - Statistical microdata.- Access not yet achieved (just for some limited research and never for production). This must be a strategic goal for the ESS. Statistical microdata not leaving MNOs' premises stand as the most promising solution to access/use.
  - Aggregated data.- More control on aggregation procedures needed.
  - Final product.- Outsourced analyses do not meet quality requirements to be considered official statistical products.
  - DPAs.- Intervention of DPAs will be fundamental.
  - Obstacle.- Risk perceptions by MNOs in different directions stand as the main obstacle.
- Methodology
  - Preprocessing.- The three main variables are the identification variables, the spatial attributes and the time attributes. Geolocation of network events arises as a key step needing further research. Voronoi tessellations already ruled out.

  - Aggregation.- Currently aggregation procedures are out of the control of NSIs. Further research regarding techniques and parameterisations even if in the final solution MNOs themselves carried out the aggregation.
  - Inference.- In inferring from aggregated data to the target population traditional probability sampling does not suffice. Proposals based on hierarchical models like in ecological sampling solutions have been made. Hierarchies allow us to deal with the integration of mobile phone data with other sources and to account for the uncertainty in these sources.

- IT
  - Since access is highly limited, no concrete empirical insight into Big Data infrastructures has been gained. Prototyping software to provide proof of concepts has been developed for different stages of the whole statistical process.

- Quality
  - Many CoP principles will be affected mainly by the fact that part of the process needs to be conducted by MNOs and by the use of novel methods needing a priori hypotheses.
  - In using statistical models, the analysis of the accuracy dimension needs to be completed with model checking and model assessment measures.

- Strategy
  - A reformulation of the research strategy is needed so that access, on the one hand, and methodology-IT-quality issues, on the other hand, are developed in parallel (with a very close communication between both activities).
  - In this new strategy, (semi-)simulated data (e.g. based on agent-based simulations both at the mobile device and aggregated levels) must play a central role.

As already pointed out in this and preceding deliverables (WP5.1, 2016; WP5.2, 2017; WP5.3, 2018; WP5.4, 2018), the limited access to mobile phone data for the present research has posed important constraints in the analysis and development of hands-on bottom-up results for the use of this new data source in the production of official statistics. However, having adapted an exploratory mind we have been able to reach very relevant conclusions for the incorporation of mobile phone data into the standard

production process at national and international statistical offices. These conclusions do not only range from the access, methodological, IT, and quality aspects of the original structure of the project envisaged in the public tender but also in important strategic and management issues regarding the future of the research on this topic within the ESS.

In the following sections we tackle each of these aspects trying to provide an exhaustive executive summary for official statistical producers without the need to dive into all the preceding material (of course, up to many details). Our starting point for all these considerations is a revision of the definition of Big Data for the production of official statistics. Instead of the classic n Vs (velocity, volume, variety, ...) we have found it more appropriate to define these new data sources in terms of the following characteristics:

1. Data do not contain information of the data holder but of third people.
2. Data play a central role in the business of the data holder.
3. Data are not generated with a specific metadata structure for statistical purposes.

Then we can add more characteristics in terms of the more technical aspects regarding velocity, volume, variety, ...

Notice that survey data clearly fails to satisfy condition 3, which makes them different to administrative data and Big Data. Indeed, these two new data sources are highly similar from this perspective with the only distinction of the public/private ownership[1]. It has taken some time for the ESS to integrate administrative data in the standard production process. Some aspects have been successfully resolved (e.g. the access to administrative registers is clearly supported by a European Regulation (European Union, 2009)). Some others are still under intense development already with important results (see, e.g., ESS.VIP Admin (2018)). For mobile phone data, the current situation seems to point in the same direction with the aforementioned distinction of these data being under private hands.

This suggested analogy is not only relevant to provide a common view of our data sources, but especially to promote the development of production frameworks integrating methodology, IT, and quality management with any sort of data. The work of the current work package on mobile phone data must be read in this sense of providing first steps towards the construction of such a production framework. Despite not having full access to data, we have been able to identify important elements and first conclusions

[1] Here *ownership* is used in a very general sense, not tackling legal details about who legally owns the data.

for the development of this framework. Combining these elements into a statistical process remains for the future.

4.1. The statistical process with mobile phone data

The end-to-end production process using mobile phone data can be represented by figure 4.1 starting from the raw telecommunication data to the final disseminated official statistical product. By and large, there exist four types of data along the process, namely (i) the raw telecommunication data generated by the electromagnetic interaction between mobile devices and antennae, (ii) the preprocessed microdata for their statistical exploitation, (iii) the aggregated data for each (possibly overlapping) cell in which the territory is divided, and (iv) the final data regarding the target population to be disseminated.

Although we still lack a standard description of these types of data sets, they can be easily identified. Raw telecommunication data are not prepared for their statistical processing. They generate from the operating activities in an MNO. They are highly technical and need heavily computational preprocessing before they can be exploited in statistical terms. Statistical microdata sets are the result of this preprocessing. Notice that these provide information at the mobile device level. Further transformation in terms of individuals (not devices) needs to be practised. Then, these data are aggregated in terms of the territorial division used by the analyst. They basically provide information on the number of mobile devices/individuals per cell (possibly also along time). Finally, these data are to be used to provide information about the target population thus constituting the final statistical product.

The evolution of the data entails the division of the process into production stages. Every stage requires the development and standardisation of different issues (access, methodology, IT, quality). Furthermore, to achieve high-quality standards, error sources must be clearly identified in each stage for their appropriate treatment. There already exist tools in the use of administrative data for official statistics purposes with these goals. Under our proposed revised definition of Big Data for Official Statistics, at least for mobile phone data, we claim that these tools can be pushed forward to provide a unified framework to deal with errors and quality (see WP5.3 (2018), Zhang (2012), Reid et al. (2017)).

We now supply more detailed conclusions about each stage.

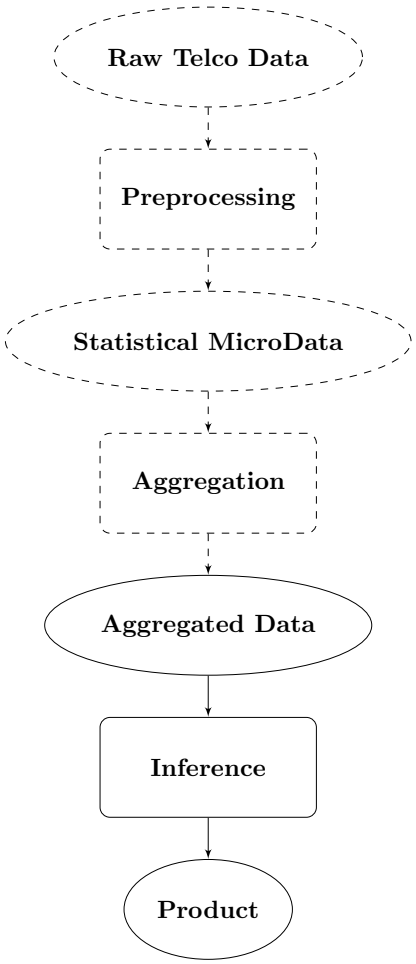

Figure 4.1 Sequence of large processing steps of mobile phone data.

4.2. Access issues

We have gathered different experiences regarding the access to the foregoing types of data sets:

- Raw telecommunication data.- No access whatsoever has been granted to this type of data by any MNO to any partner of the project. Indeed it seems clear that even in a near future this access will be not granted. This would be highly disruptive in the operations of the MNOs and would also require completely novel and different skills for official statisticians to process these data. Indeed, a close collaboration

between NSIs and MNOs in the future will be needed at this point (see below).

- Statistical microdata.- Direct access to this form of data for CDRs and indirect access for signalling data have been achieved only in very limited conditions and only for research purposes (see WP5.2 (2017)). As a clear conclusion from our experience in this project, NSIs must clearly access/use these data as a strategic goal in order to achieve high-quality statistical outputs.
- Aggregated data.- Direct access to this form of data has been also achieved in the project. However the aggregation cannot be conducted by NSIs themselves thus leaving some details out of our control. The core methodological inference proposal in the project has been developed using these data as input data.
- Final product.- Also final products have been agreed to be received by an NSI in the project (see section 2.3). This final product was scrutinised by experts at the office to assess its potential validity as an official statistical product. A clear conclusion was reached: this sort of final product elaborated outside the NSIs does not reach the quality conditions to be considered an official statistical product.

Regarding the conditions upon which access has not been successfully achieved, the situation is remarkably complex. Legal issues are immediately alluded to by MNOs as a first obstacle. There exist different legal situations in different countries and global statements cannot be made. Nonetheless, by and large, we do not perceive the legal issues as the main obstacle since national DPAs usually provide legal support to access these data under strict confidential and privacy-respecting conditions (at least as traditional data sources, which are indeed identified personal data). Generally speaking, current legislations seems to provide enough legal support to request the use of these data for the production of official statistics across the ESS.

The main obstacle, in our view, is the risk perceived by MNOs regarding granting data access/use to the NSIs. This risk perception takes form in different ways. Sometimes, they are not willing to share a source of promising revenues perceiving official statistical producers as dangerous competitors. In other cases, in their highly volatile markets, clients are perceived to potentially finish their subscriptions if their data are shared. Risks are also perceived regarding legal issues and the potential sanctioning action of the DPAs not to mention their image in the public opinion. Also, costs related to data extraction/access/use operations usually stand up as a relevant element arising from the operational complexity to access these data.

As we shall state below in the section on future prospects, collaboration with MNOs under a clear legal framework with an appropriate communication strategy seems to be the most appropriate solution.

4.3. Methodological issues

Methodological conclusions can be also offered for the different sub-processes depicted in figure 4.1:

- Preprocessing.- The three main attributes for a mobile device/individual in the statistical microdata set are the identification variables, the spatial attributes and the time attributes. Both time and identification variables require less degree of complexity than spatial attributes for their computation. The geolocation of network events in the preprocessing stage has been clearly identified as a key step in the production process. One of the main conclusions in this respect has been to rule out the territorial division in Voronoi cells as an efficient and accurate technique to assign spatial attributes to each mobile device. Concrete proposals have been supplied in WP5.3 (2018) taking into account the directional orientation of cells and the signal strength between mobile devices and antennae.

  The assignment of spatial attributes must be complemented with time attributes so that the joint analysis of both attributes (trajectories) can allow us to assign and classify so-called anchor points for each mobile device/individual.

  Complementarily, a core data model to standardise statistical microdata sets and the data architecture has also been proposed, although more experience with full access to these data across the ESS is needed. The computation of these spatial attributes and space-time interpolations is central.

- Aggregation.- The aggregation of microdata information into the different territorial cells has been so far carried out by MNOs themselves for those cases in which access has only been agreed at this form of data sets. This obliges NSIs to provide an aggregation scheme under a given parametrisation. For the time being, this is a somewhat blind operation and further research regarding variations and robustness of this parametrisation is needed. Standard definitions and structural metadata, in general, are needed.

- Inference.- To conduct the inference from aggregated data to the target population of interest we have recognised that traditional probability sampling is not enough. We have proposed to use hierarchical models in a similar way to how ecologists approach the species abundance problem across a given geographic territory. Furthermore, the use of a hierarchy in the modelling exercise allows us to deal with the integration of mobile phone data with other sources in a natural way and, at the same time, to account for the uncertainty in these sources.

4.4. IT issues

Big Data are clearly identified as IT resource-demanding from diverse points of view (access, storage, processing, visualization, ...). Our limited access to data has prevented us to face these common challenges. Regarding microdata, access and use have been undertaken indirectly without data coming out of the MNOs' premises and in almost all cases by their own staff. When this form of data has been received at statistical offices, the volume has been remarkably low and even with a limited set of variables (e.g. no antenna position provided in some cases, widely coarse-grained spatial attributes, ...). We have provided in WP5.4 (2018) an overview of a computer platform to deal with these data at the MNOs' own premises based upon our experience. Details about the IT infrastructure will come closely hand-to-hand with the agreements to access and use the data. In any case, everything suggests right now that microdata will not (and should not) leave the MNOs' premises. Regarding aggregated data, their needs for Big Data IT infrastructures for storage and access are less demanding, although processing needs appear regarding the inference stage.

The prototyping software developed in this project has been proved to yield meaningful results with (semi-)simulated data at a small scale (mainly at the PC scale with more or less long computation times). Now the challenge to set up the scalability frontiers is settled by using real data.

4.5. Quality issues

As stated above, the similarity with the use of administrative data in the production of official statistics strongly suggests that a close revision of both methodological and quality assessment methods must be revisited for their adequate reuse with mobile phone data (even Big Data in general, we dare to say). Even the Total Survey Error framework potentially appears as a useful tool with this new data source. The identification of error sources is still crucial for a high–quality final output.

However, novel techniques, and especially those requiring a modelling exercise, require a close analysis of quality indicators so that estimates can be considered robust against model misspecifications and a priori hypotheses. A first set of quality indicators regarding accuracy has also been provided in terms of posterior biases, variances, and mean square errors.

4.6. Strategic issues

This work package, along with the whole ESSnet on Big Data, has followed a hands-on bottom-up strategy dealing successively with data access, methodology, IT issues,

and quality assessment. This entails that failing to access data puts later stages under a high risk. We conclude that a reformulation of this strategy is needed so that access, on the one hand, and methodology-IT-quality issues, on the other hand, are developed in parallel (with a very close communication between both activities). In this new strategy, (semi-)simulated data (e.g. based on agent-based simulations both at the mobile device and aggregated levels) must play a central role.

5

Future prospects

Executive summary

In an executive fashion we underline the main future prospects as a bullet list:

- Strategy
  - A two-branch research strategy is needed. Access must be the goal of first track. Methodology, IT and quality issues must be the goal of the second track.
  - In this strategy, (semi-)simulated data based on agent-based simulations both at the mobile device and aggregated levels must play a central role in connecting both tracks.
- Access
  - Reinforce and clarify the legal support for NSIs to access and use mobile phone data to produce official statistics in the ESS.
  - Search for a satisfactory technological solution to use data in MNOs' premises possibly integrating official data in a completely anonymised way.
  - Assess in detail the data extraction costs and related infrastructures for their use by NSIs.
  - Develop a communication strategy shared by MNOs and NSIs clearly explaining to the citizenship and the public opinion the circumstances under which their mobile phone data will be used to produce official statistics.

- Methodology
  - Develop methodology for the generation of (semi-)simulated data.
  - Fully adapt the two-phase life-cycle model and related works to mobile phone data identifying needs for microintegration and event-unit transformations, error sources and their treatment.
  - Develop methodology for (i) the assignment of spatial attributes (geolocation of network events), (ii) space-time interpolation, and (iii) anchor-point identifications.
  - Further develop the hierarchical model for inference under a diversity of input data (including different priors, geostatistical considerations, ...).
- IT
  - Develop software for the generation of (semi-)simulated populations.
  - Along with the business perspective, find and agree on the technical details of a technological solution to use data in MNOs' premises possibly integrating official data in a completely anonymised way.
  - Investigate the scalability of the software implementations (both the packages `mobloc` and `pestim` and the software for generation of synthetic data) in order to implement them in a distributed computing platform.
  - Revise and optimise some computation algorithms in the implementation of the hierarchical model (including e.g. the investigation of quadratures instead of Monte Carlo techniques to compute certain integrals).
  - Include as many visualisation facilities in the software implementations and possibly consider the integration with GIS software.
  - Promote prototypes to efficient production tools.
- Quality
  - Identify, analyse, and assess all potential errors and their sources.
  - Provide standards for different aspects such as the data architecture, structure and content of aggregated or output data files, ...Standards must rule the combination of different process steps.
  - Provide precision or accuracy measures for all variable assignments, estimates and/or model checking and assessment. Revise the related quality indicators. If spatial correlations are also considered, investigate about the adequate indicators to evaluate the final outputs.

- Provide comparison indicators for final estimates from the mobile phone data process and official figures (correlations, cosine similarities, spatial autocorrelation measures, etc.)
- Develop efficient software implementations for all the preceding measures.

- Management
  - Assignment roles associated to the diverse elements of the research proposal.

A schematic representation of elements for research are given in the picture 5.1.

In this section we provide a detailed list of different elements and issues for the development of the production framework initiated in the current work package for the use of mobile phone data in the production of official statistics. We also include some reflections and/or recommendations regarding both the strategic and management aspects of this research in the immediate future.

5.1. Strategic issues

As commented above, the whole current ESSnet has been oriented in the axis going from accessing concrete data sets over developing the necessary methodology and IT tools for producing concrete statistical outputs to the quality issues needed for these outputs to be an official statistical product. For Official Statistics this bottom-up approach is undeniably fundamental because we must prioritise the production of concrete results for policy making and decision taking in those organisations using our statistical outputs.

However, making the development of methodology, IT tools, and quality assessment depend on having concrete real data sets is too risky. We identify as a key strategic element in the future research with mobile phone data to carry out a double-track approach. On the one hand, efforts must concentrate on the highly entangled issues surrounding access and use of these data not only for research but especially for standard production. On the other hand, the development of statistical methodology, IT infrastructure, and quality management tools must be undertaken independently to having real data.

The key intersecting point of these two tracks is to produce (semi-)simulated data taking advantage of the available knowledge on these data. By (semi-)simulated data we mean synthetic data both at the microdata (mobile device) and aggregated levels with a

structure as similar to real data as possible. At the aggregated level we can proceed as follows. Starting from official population data N_i^{Reg} at a given territorial partition into cells $i \in \mathcal{I} = \{1, 2, 3, \dots\}$ and data from national telecommunication regulators about penetration rates and market shares we must produce synthetic values for (i) the true population $N_i^{(0)}(t)$ at different time periods[1], (ii) the population $N_i^{\text{Mob}}(t)$ of individuals carrying mobile devices, and (iii) the population $N_i^{\text{MNO}}(t)$ of individuals according to one or several MNOs. Whatever data generation mechanism is put into place, notice that we can assess both the methodology and quality issues not only for a given concrete data set but for as many data sets as we are willing to generate (under diverse hypotheses to test a variety of situations). Notice also how generating aggregated synthetic data allows us to concentrate only on the inference stage not needing to preprocess and aggregate statistical microdata.

At the statistical microdata level, inspired by agent-based simulations (Salamon, 2011), we can also generate (possibly coarse-grained) space and time attributes (i) for each individual according to the official population register, (ii) for each mobile device (according to the national telecommunication regulator), and (iii) for each subscriber of a given MNO (also according to the national telecommunication regulator). With these data we can then proceed to aggregate data allowing us to assess this process step.

These data generation mechanisms can be made more realistic by progressively introducing more and more complex hypotheses (e.g. introducing inbound tourists, ...). Even having access to real data, synthetic data are a highly valuable element to assess the performance of statistical methods. The generation of these synthetic data must be a strategic element in future research.

In the following we provide details on these two proposed parallel tracks.

5.2. Access/Use issues

Access to mobile phone data is a highly entangled question far from being solved across the ESS. The situation is not only different between countries but also in the same country with different MNOs. Even the limited access for research purposes achieved for the present project has involved more efforts than those originally planned for the SGA-1 (see WP5.1 (2016) and WP5.2 (2017)).

Although no universally magic recipe to achieve an agreement with MNOs can be provided as a result of this project (WP5.2, 2017), some lessons for the future must

[1] E.g. using proximity measures between pairs of cells to assign probability of displacement across the territory.

be taken into account. Firstly, as the title of this section points out, it is important to begin talking of data use instead of data access[2]. This subtlety entails that data should never leave the MNOs' information systems and never be transmited to NSIs. Rather on the contrary, a technological solution must be agreed so that official statisticians can use these data in MNOs' premises. Even the integration with official data must be considered in designing such a technological solution. In this regard, we must mention that there already exist hardware solutions to process data in a completely anonymised way (see [WP5Meet1]). These conditions for data use for official statistical purposes will hopefully ease the task to reach an agreement.

In parallel, it may be necessary to clarify (even reinforce in some cases) the legal support through European regulations for NSIs and Eurostat to use these data for the production of official statistics. The situation, to some extent, may be seen as parallel to the original situation with administrative registers. Now even the CoP includes the recommendation for statistical producers to influence on the design of these registers. This initiative must not only seek to reinforce the position of the ESS but also to provide a clear legal protection for MNOs to share their data. Data sharing for official statistics purposes should not be based on a sheer voluntarily basis but as legal obligation so that NSIs can keep on providing a very relevant public service.

From the analysis in this project it seems clear that MNOs, as data generators due to their activity, will play a more involved role in the statistical production process than traditional respondents or data providers. The technological characteristics of this new data source and this involved role introduce the costs of data extraction as an important element to be taken into account. Official Statistics must never pay for their data, otherwise this official information would be immediately unsustainable. A different issue is to develop a data colletion, storage, and processing infrastructure specific for this new data source. Notice that a parallel infrastructure is already present in NSIs for traditional sources (interviewers are hired and trained to go to household across the national territories to collect data, questionnaires are printed and sent by postal mail in some cases, CAWI data collection modes need internal management, ...). Now an important part of the new needed infrastructure must be within the MNOs. This situation, at a much smaller scale, is already present in some European statistics (ESSnetADC, 2012). Thus mobile phone data use is indeed pushing in the direction of already present automatic data collection methods.

All in all, a closer collaboration with data holders (MNOs) will be necessary, a closer collaboration built on mutual trust under a clear legal framework and hopefully with a

[2]This conclusion is not only a consequence of the current ESSnet but also of the parallel complementary activies of the ESS Task Force on Big Data.

common communication strategy regarding the use of mobile phone data of citizens for the production of official statistics arises as the optimal solution.

Thus, the main lines of action regarding data access/use are:

- Reinforce and clarify the legal support for NSIs to access and use mobile phone data to produce official statistics in the ESS.
- Search for a satisfactory technological solution to use data in MNOs' premises possibly integrating official data in a completely anonymised way.
- Assessment in detail the data extraction costs and related infrastructures for their use by NSIs.
- Develop a communication strategy shared by MNOs and NSIs clearly explaining to the citizenship and the public opinion the circumstances under which their mobile phone data will be used to produce official statistics.

Notice that at least three of these lines action should be desirably followed in close collaboration with data scientists from the MNOs, although this is not a necessary condition.

5.3. Methodological issues

In the process depicted in figure 4.1 we have identified several methodological issues which clearly needs further work:

- Given the relevance of the generation of (semi-)simulated data, related methodology must be developed clearly showing how data (official, administrative, from national telecommunication regulators,...) are integrated and what explicit hypotheses are made to generate different types of populations (inbound, outbound, resident tourists, commuters, etc.). This methodology must be developed both at the microdata (mobile device) and aggregated (territorial cells) levels.
- The two-phase life-cycle model by Zhang (2012) must be closely analysed to find an optimal adaptation to the mobile phone data process in the spirit of similar initiatives with administrative data (see e.g. Reid et al. (2017)). Notice that this model entails also both metadata and quality management issues (to be again cited below). Here we focus on the statistical methods (mainly micro-integration techniques) to produce the successive data sets during the process. In particular, the process of transforming original network events into individuals with diverse variables (spatial and time attributes, ...) should be clearly profiled making

explicit what techniques can be reused from other domains and what needs for new methods exist.

- The assignment of spatial attributes to mobile devices is a crucial step in the construction of the statistical microdata. The geolocation of network events is a completely new ingredient in the statistical process and several preliminary techniques are used by MNOs' data scientists and official statisticians already exploring these data (modelling antennae coverage by tessellation or overlapping cells, relating different spatial grids through area proportion, maximum likelihood, Bayesian techniques, etc.).

  Spatial attributes must be duly complemented with time attributes so that so-called anchor points (see WP5.3 (2018)) can be readily identified for home, work, ... Also different possibilities arise depending on the type of data at our disposal (CDRs or signalling): derivation using two detected locations, integration of locations in a timeframe, building a continuous description of locations, ...

- The hierarchical model proposed in WP5.3 (2018) constitutes a first simple proposal which needs further analysis to embrace more realistic situations (detection of more mobile devices than individuals according to a population register or survey data, modelling of cells introducing net changes in the population total like airports, train stations, country border cells, etc.). A detailed study of how more available auxiliary information (e.g. about land use) may be integrated into the choice of priors is also needed.

  In this same line of producing a more sophisticated model, geostatistical considerations must be taken into account to introduce spatial correlations. These spatial correlations should also be considered when assessing the results (see below).

5.4. IT issues

IT issues can be broadly divided into two large groups: (i) those closely related to access/use of data in MNOs' premises especially focused on storing and processing statistical microdata, and (ii) those closely related with the computationally demanding resources needed by the simulation exercise and the implementation of the hierarchical model.

Clearly, the search for a satisfactory technological solution to use data in MNOs' premises must be carried out considering both business and technical perspectives. Technical questions will ultimately depend on a diversity of concrete characteristics of each MNO (volumen and velocity of generation of data, computer facilities, degree of

data monetisation, etc.).

Regarding the computational challenges, we can recognise two related parts. On the one hand, the construction of (semi-)simulated data sets as described above must be a key element to test, modify, and improve many aspects of the process (including the hierarchical model). On the other hand, the Bayesian approach followed in the implementation of model requires a high computation power, although algorithms and routines have been chosen and designed with an eye on parallelization.

So far R packages providing prototyping computations have been developed (see WP5.4 (2018)), but as we get closer to real conditions in terms of volume and velocity, the computational challenges will be higher. Parallel processing is necessary to come close to these real conditions. A distributed computation version of this software must be investigated.

Thus we identify the following potential points for the future research:

- Develop software for the generation of (semi-)simulated populations.
- Along with the business perspective, find and agree on the technical details of a technological solution to use data in MNOs' premises possibly integrating official data in a completely anonymised way.
- Investigate the scalability of the software implementations (both the packages `mobloc` and `pestim` and the software for generation of synthetic data) in order to implement them in a distributed computing platform.
- Revise and optimise some computation algorithms in the implementation of the hierarchical model (including e.g. the investigation of quadratures instead of Monte Carlo techniques to compute certain integrals).
- Include as many visualisation facilities in the software implementations and possibly consider the integration with GIS software.
- Promote prototypes to efficient production tools.

5.5. Quality issues

Quality issues cross-cut many of the issues tackled above. In some sense, they should be included in the preceding suggested points. However, to better appreciate the continuation with the current project we provide a separate list of items to be dealt with in the future:

- In connection with the adaptation of the two-phase life-cycle model and possibly a Total Survey Error model for mobile phone data, identify, analyse, and assess all potential errors (traditionally so-called nonsampling errors) and their sources. This is a key element for the quality of an official statistical product and the fulfillment of the CoP and ESS QAF.

- In this same line of using a given unified model to understand the process, provide standards for different aspects such as the data architecture (as the core data model proposed in WP5.3 (2018)), aggregated or output data files to be exchanged between MNOs and NSIs (e.g. using SDMX with an adequately chosen structure), definitions of new elements (for variables, methods, types of errors,..., i.e. complement the structural metadata system of traditional statistical production), ...

- Provide precision or accuracy measures for all variable assignments, estimates and/or model checking and assessment (if models are used) along the statistical process with mobile phone data. If Bayesian choices are made, then duly integrate these measures into the selection of priors. Since there exist profound changes in the inference methods with respect to target populations, we need to closely revise the related quality indicators (e.g. currently probability sampling does not need to deal with model checking, so this must be a new element in the quality assessment). If spatial correlations are also considered, we must investigate about the adequate indicators to evaluate the final outputs.

- Provide comparison indicators for final estimates from the mobile phone data process and official figures (correlations, cosine similarities, spatial autocorrelation measures, etc.)

- In parallel, develop efficient software implementations for all the preceding measures.

All proposed indicators and quality measures must be tested using the (semi-)synthetic populations and real populations.

5.6. Management issues

As the reader can appreciate, it is virtually impossible that a single NSI could carry out this research proposal and the ESS as a whole needs to be the actor of these activities. This necessarily entails an extraordinary degree of coordination between experts of diverse background from several offices.

So far, the current work package on mobile phone data has shown an exploratory nature in which each WP member country has conducted their own research according to national circumstances and priorities following more or less the guided script on access, methodology, IT, and quality stated above set up by the public tender.

In the future, an assignment of roles associated to the diverse elements of the research proposal (this or other similar) must be closely followed in order to gain efficiency in the development of these activities.

As a final representation of this set of identified elements for the future research on mobile phone data for the production of official statistics, we close this work package with the picture 5.1.

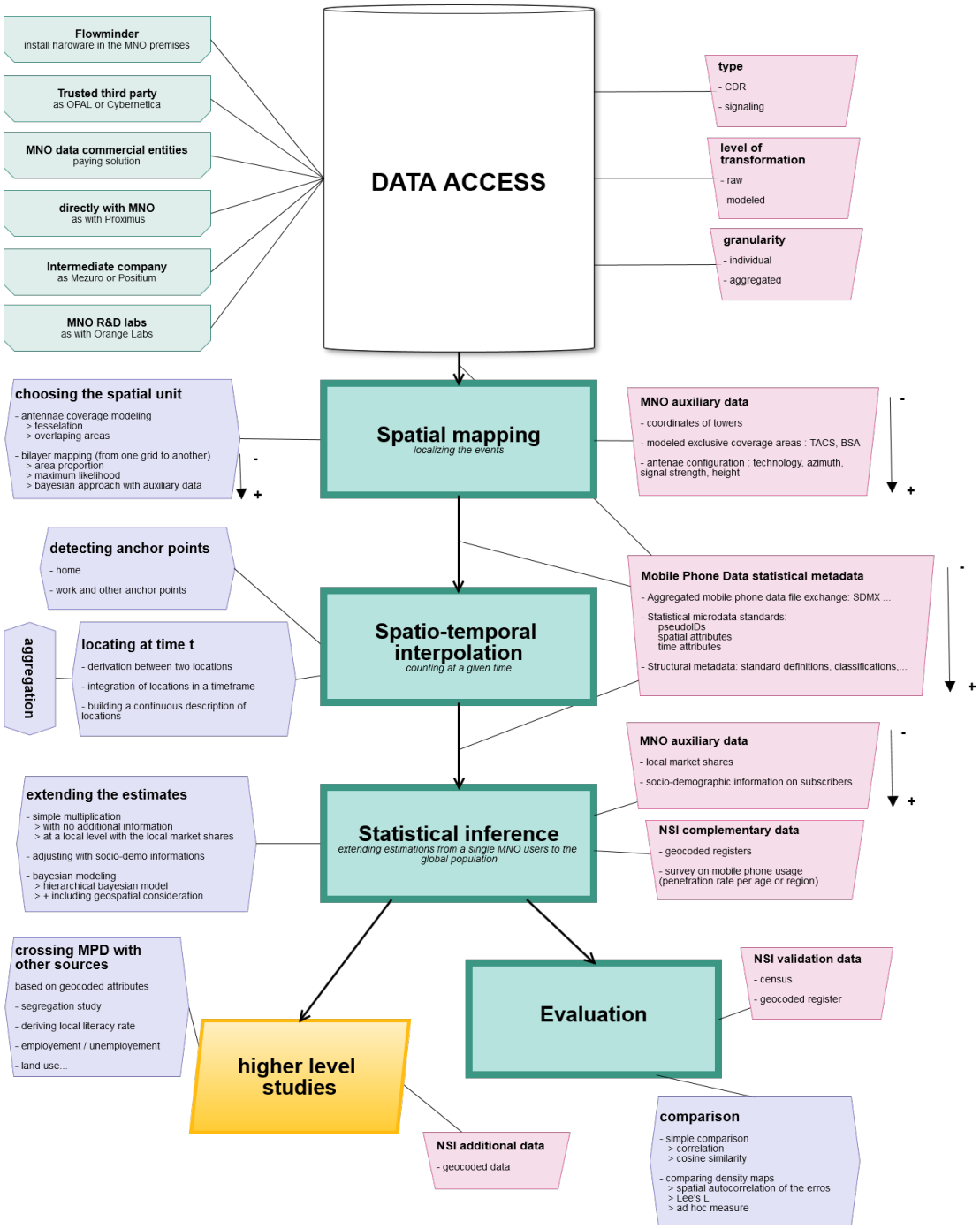

Figure 5.1 Schematic summary of the main elements for future research.

Appendix A

Computational details

A.1. Computation of the model checking indicators

Firstly, we shall focus on the indicators at the initial time t_0. As usual, for ease of notation, we drop out both the cell subscript and the time dependence. To compute the indicators (3.7a) to (3.7f) we need to evaluate integrals of the form

$$I_m(N^{\text{MNO}}) = \int_{\Omega_N} \left(N^{\text{MNO, rep}}\right)^m d\mathbb{P}\left(N^{\text{MNO, rep}} \middle| N^{\text{MNO}}\right), \tag{A.1}$$

where $\mathbb{P}\left(N^{\text{MNO, rep}} \middle| N^{\text{MNO}}\right)$ is the posterior predictive distribution (3.6). This integral can be rewritten as

$$I_m(N^{\text{MNO}}) = \int_{\Omega_{u,v,\lambda}} \left[\int_{\Omega_N} \left(N^{\text{MNO, rep}}\right)^m d\mathbb{P}\left(N^{\text{MNO, rep}} \middle| u, v, \lambda\right)\right] d\mathbb{P}\left(u, v, \lambda \middle| N^{\text{MNO}}\right). \tag{A.2}$$

Then, focusing on the inner integral $I_m(u, v, \lambda)$ we can write

$$I_m(u, v, \lambda) = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} (n_i^{\text{MNO, rep}}(u, v, \lambda))^m, \tag{A.3}$$

where $n_i^{\text{MNO, rep}}(u, v, \lambda)$ denotes a random value generated by the distribution $\mathbb{P}\left(N^{\text{MNO, rep}}(t_0) \middle| u, v, \lambda\right)$. Using again a Monte Carlo approximation, the original integral can then be computed by writing

$$I_m(N^{\text{MNO}}) = \lim_{M_1\to\infty} \lim_{M_2\to\infty} \frac{1}{M_1 M_2} \sum_{i_1=1}^{M_1} \sum_{i_2=1}^{M_2} (n_{i_2}^{\text{MNO, rep}}(u_{i_1}, v_{i_1}, \lambda_{i_1}))^m, \tag{A.4}$$

where the values $u_{i_1}, v_{i_1}, \lambda_{i_1}$ are generated according to the posterior distribution $\mathbb{P}\left(u, v, \lambda \middle| N^{\text{MNO}}\right)$.

In practice, we need first to generate the values $u_{i_1}, v_{i_1}, \lambda_{i_1}$ and then conditional on them we generate the values $n_{i_2}^{\text{MNO, rep}}(u_{i_1}, v_{i_1}, \lambda_{i_1})$ (hence the subscript notation).

Now, for later time periods t_n, to compute the indicators (A.11a) to (A.11f) we need to evaluate integrals of the form

$$I_m(\boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)) = \int_{[0,1]\times I} \left(\boldsymbol{\pi}_i^{\text{MNO, rep}}\right)^m d\mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0, t_n) \middle| \boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)\right), \quad \text{(A.5)}$$

where $\mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0, t_n) \middle| \boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)\right)$ is the multivariate posterior predictive distribution (3.8). Then we can mimic the same development as with the initial time period and write

$$\begin{aligned} I_m(\boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)) &= \int_{[0,1]\times I} \left(\boldsymbol{\pi}_i^{\text{MNO, rep}}\right)^m d\mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}} \middle| \boldsymbol{\pi}_i^{\text{MNO}}\right) && \text{(A.6)} \\ &= \int_{\Omega_{\boldsymbol{\alpha}_i(t_0,t_n)}} \left[\int_{\Omega_{\Omega_{[0,1]\times I}}} \left(\boldsymbol{\pi}_i^{\text{MNO, rep}}\right)^m d\mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0, t_n) \middle| \boldsymbol{\alpha}_i(t_0, t_n)\right)\right] d\mathbb{P}\left(\boldsymbol{\alpha}_i(t_0, t_n) \middle| \boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)\right) \\ &= \lim_{M_1, M_2 \to \infty} \frac{1}{M_1 M_2} \sum_{i_1=1}^{M_1} \sum_{i_2=1}^{M_2} \left(\pi_{i,i_2}^{\text{MNO, rep}}(\boldsymbol{\alpha}_{i,i_1}(t_0, t_n))\right)^m, && \text{(A.7)} \end{aligned}$$

where

- the random values $\boldsymbol{\alpha}_{i,i_1}(t_0, t_n)$ are generated according to the posterior distribution $\mathbb{P}\left(\boldsymbol{\alpha}_i(t_0, t_n) \middle| \boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)\right)$;
- the random values $\pi_{i,i_2}^{\text{MNO, rep}}(\boldsymbol{\alpha}_{i,i_1}(t_0, t_n))$ are generated according to the model

$$\mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_n) \middle| \boldsymbol{\alpha}_{i,i_1}(t_0, t_n)\right).$$

Thus, in both cases (at t_0 and t_n) everything is reduced to the random generation of the model hyperparameters according to their respective posterior distributions.

A.2. Generation of the model hyperparameters

We need to provide algorithms to generate the multivariate vectors (u, v, λ) and $\boldsymbol{\alpha}_i(t_0, t_n)$ according to the distributions $\mathbb{P}\left(u, v, \lambda \middle| N^{\text{MNO}}(t_0)\right)$ and $\mathbb{P}\left(\boldsymbol{\alpha}_i(t_0, t_n) \middle| \boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)\right)$,

respectively. We proceed along similar lines as those used in the appendix of our deliverable on methodology (WP5.3, 2018) profusely using the rejection algorithm (Robert and Casella, 2004) now in a multivariate setting.

For the vector (u, v, λ) we can write

$$\mathbb{P}\left(u, v, \lambda \middle| N^{\text{MNO}}(t_0)\right) \propto \mathbb{P}\left(u, v \middle| \lambda, N^{\text{MNO}}(t_0)\right) \mathbb{P}\left(\lambda \middle| N^{\text{MNO}}(t_0)\right), \tag{A.8}$$

which clearly suggests to generate firstly random values for the posterior marginal distribution for λ and conditional on these values then generate values for u and v. The first step was addressed and solved in the deliverable on methodology. For the second step, we need the unnormalised joint posterior distribution for (u, v, λ) given by

$$\mathbb{P}\left(u, v, \lambda \middle| N^{\text{MNO}}(t_0)\right) \propto f_u(u) f_v(v) f_\lambda(\lambda) \cdot \text{Po}\left(N^{\text{MNO}}(t_0); \lambda\right) \cdot \Phi\left(u \cdot v, u \cdot (1-v), \lambda, N^{\text{MNO}}(t_0)\right),$$

where

- f_u, f_v, and f_λ denote the priors for these hyperparameters;
- $\text{Po}(n; \lambda)$ stands for the Poisson probability function with parameter λ;
- $\Phi(\alpha, \beta, \lambda, N)$ is defined as in the generation of the random values λ:

$$\Phi(\alpha, \beta, \lambda, N) = \frac{B(\alpha + N, \beta)}{B(\alpha, \beta)} \cdot {}_1F_1(\lambda; \beta, \alpha + \beta + N).$$

This joint distribution allows us to apply the rejection algorithm on the conditional distribution $\mathbb{P}\left(u, v \middle| \lambda, N^{\text{MNO}}(t_0)\right)$. As two-dimensional candidate distribution we use the prior $f_u(u) \cdot f_v(v)$. The rest of the algorithm is standard (Robert and Casella, 2010).

For the vector $\boldsymbol{\alpha}_i(t_0, t_n)$, using the model, we can write

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{\alpha}_i(t_0, t_n) \middle| \boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)\right) &\propto \mathbb{P}\left(\boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n) \middle| \boldsymbol{\alpha}_i(t_0, t_n)\right) && \text{(A.9)} \\ &\propto \text{Dirichlet}(\boldsymbol{\alpha}_i(t_0, t_n)), && \text{(A.10)} \end{aligned}$$

where, as argued in section 3.2, we have assimilated $\mathbf{p}_i(t_0, t_n)$ and $\boldsymbol{\pi}_i^{\text{MNO}}(t_0, t_n)$.

A.3. Indicators for the transition probability matrices

The set of indicators for the transition probability matrices for cell i is given by:

- Bias indicator:

$$\mathsf{b}_i(t_0,t_n) = \mathbb{E}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n) - \boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right). \tag{A.11a}$$

The relative counterpart is:

$$\tilde{\mathsf{b}}_i(t_0,t_n) = \mathbb{E}\left(\frac{\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n) - \boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)}{\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)}\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right). \tag{A.11b}$$

- Variance indicator:

$$\mathsf{v}_i(t_0,t_n) = \mathbb{V}\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n) - \boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right). \tag{A.11c}$$

The relative counterpart is:

$$\tilde{\mathsf{v}}_i(t_0,t_n) = \mathbb{V}\left(\frac{\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n) - \boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)}{\pi^{\text{MNO}}(t_0,t_n)}\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right). \tag{A.11d}$$

- Mean square error estimator:

$$\mathsf{mse}_i(t_0,t_n) = \mathbb{E}\left[\left(\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n) - \boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right)^2\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right]. \tag{A.11e}$$

The relative counterpart is:

$$\widetilde{\mathsf{mse}}(t_0,t_n) = \mathbb{E}\left[\left(\frac{\boldsymbol{\pi}_i^{\text{MNO, rep}}(t_0,t_n) - \boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)}{\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)}\right)^2\big|\boldsymbol{\pi}_i^{\text{MNO}}(t_0,t_n)\right]. \tag{A.11f}$$

In the illustrative example included in section 3.2 for the transition probability matrices, the representation for the relative variance and mean square error indicators are:

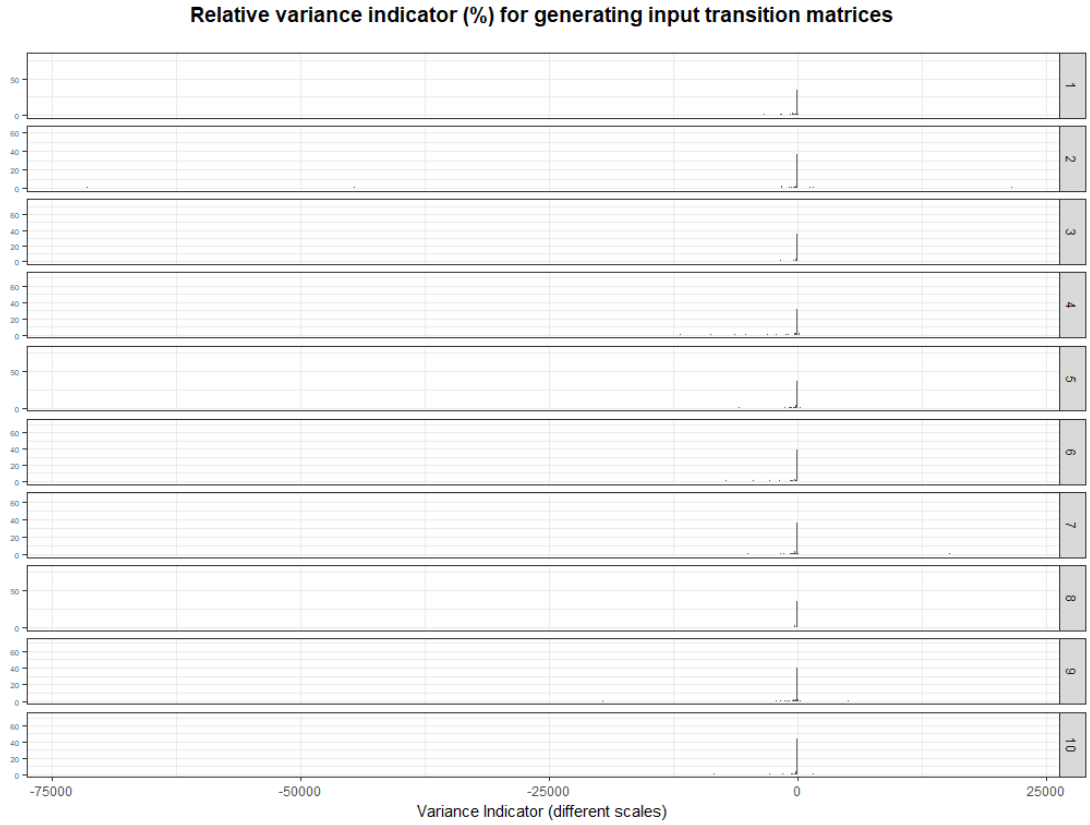

Figure A.1 Relative variance indicators (in percentage) for uniform priors with coefficients of variation equal to 0.10 for all cells.

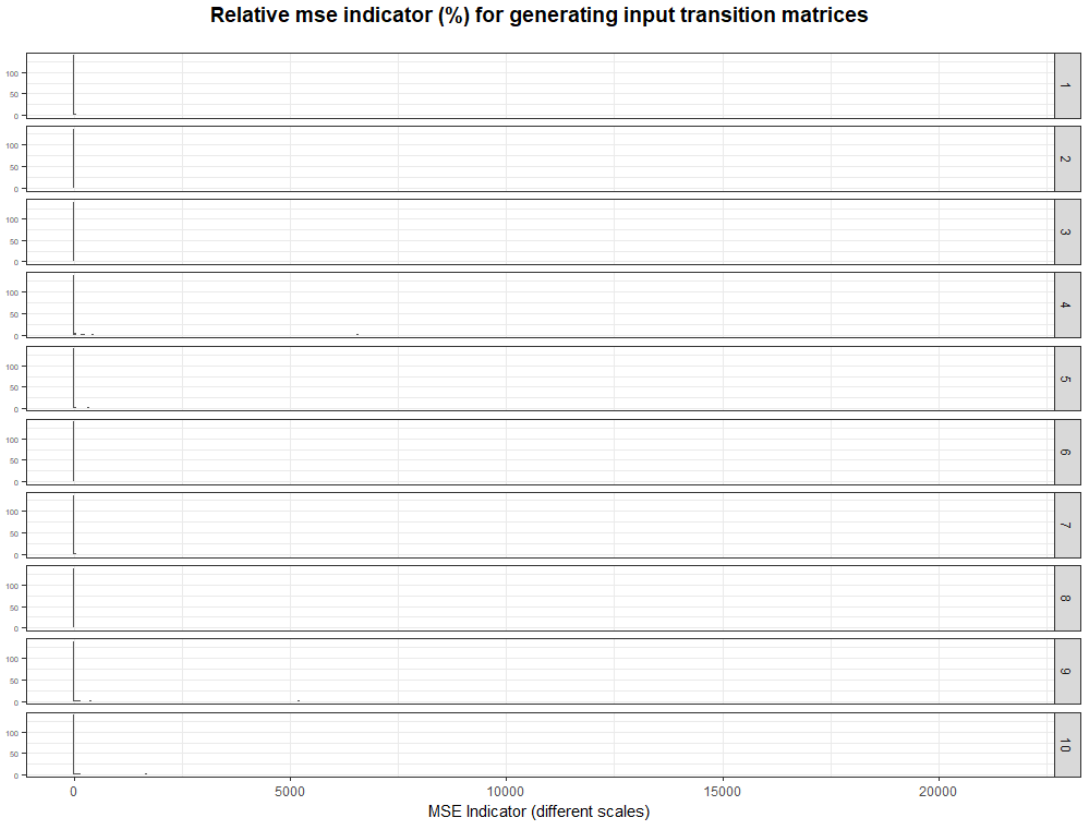

Figure A.2 Relative mse indicators (in percentage) for uniform priors with coefficients of variation equal to 0.10 for all cells.

Bibliography

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury Press.

de Jonge, E., M. van Pelt, and M. Roos (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. *CBS Discussion Paper* 201214. Available at https://www.cbs.nl/NR/rdonlyres/4EDB51ED-927A-4A69-B8F3-4DC57A44DDE4/0/Timepatternsgeospatialclusteringandmobilitystatistics.pdf.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. Stevens, A. Gaughan, V. Blondel, and A. Tatem (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences (USA) 111*, 15888– 15893.

Douglass, R.W, D.A. Meyer, M. Ram, D. Rideout, and D. Song (2015). High resolution population estimates from telecommunications data. *EPJ Data Science 4:4*.

ESS (2011). European Statistics Code of Practice. http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15.

ESS (2012a). ESS Quality Assurance Framework. http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646.

ESS (2012b). ESSnet on Automated Data Collection and Reporting in Accomodation Statistics. https://ec.europa.eu/eurostat/cros/content/tourism.

ESS (2017). ESSnet on Big Data. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php.

ESS (2018). European Master of Official Statistics. http://ec.europa.eu/eurostat/web/european-statistical-system/emos.

ESS (2018). European Statistical Training Programme . http://ec.europa.eu/eurostat/web/european-statistical-system/training-programme-estp.

ESS Task Force on Big Data (2017). Results from the analysis project on legal issues related to the use of Big Data. Doc.DDG.TF.BD 2017 04 27-28-4-legal issues. Internal document.

ESS (2018). ESS Vision 2020 ADMIN (Administrative data sources). https://ec.europa.eu/eurostat/cros/content/ess-vision-2020-admin-administrative-data-sources_en.

European Union (2009). Regulation (EC) No. 223/2009 of the European Parliament and of the Council. Official Journal of the European Union L 87/164 (March 31, 2009). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:en:PDF.

Furletti, B., R. Trasarti, P. Cintia, and L. Gabrielli (2017). Discovering and Understanding City Events with Big Data: The Case of Rome. *Information 8(3)*, 74.

Gelman, A., B. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian data analysis*. CRC Press.

Kang, C., Y. Liu, X. Ma, and L. Wu (2012). Towards Estimating Urban Population Distributions from Mobile Call Data. *Journal of Urban Technology 19(4)*, 3–21.

NetMob (2017). Conference on the scientific analysis of mobile phone datasets. http://netmob.org/.

Rao, J. and I. Molina (2015). *Small area estimation (2nd ed)*. Wiley.

Reid, G. and F. Zabala and A. Holmberg (2017). Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics 33(2)*, 477–511.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods (2nd ed)*. Springer.

Robert, C. and G. Casella (2010). *Introducing Monte Carlo Methods with R*. Springer.

Salamon, T. (2011). *Design of Agent-Based Models : Developing Computer Simulations for a Better Understanding of Social Processes*. Bruckner Publishing.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. Springer.

WP5 of ESSnet on Big Data (2016). Deliverable 5.1. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5_Deliverable_1.1.pdf.

WP5 of ESSnet on Big Data (2017a). Deliverable 5.2. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.2.pdf.

WP5 of ESSnet on Big Data (2017b). Minutes of the 1st Physical Meeting of WP5. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Meeting_2017_06_07-08_Madrid.

WP5 of ESSnet on Big Data (2018a). Deliverable 5.3. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.3.pdf.

WP5 of ESSnet on Big Data (2018b). Deliverable 5.4. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.4.pdf.

WP5 of ESSnet on Big Data (2018c). Mobile Phone ESSnet Big Data. https://github.com/MobilePhoneESSnetBigData.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica 66(1)*, 41–63.