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 5.3

Proposed Elements for a Methodological Framework 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

Introduction

This document proposes elements for a methodological framework to integrate aggregated mobile phone data in the production of official statistics. It corresponds to the third deliverable of the work package on mobile phone data of the ESSnet on Big Data (ESSnetBD, 2017).

In the preceding deliverables (WP5.1, 2016; 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 these data sources in the ESS, to compile enough data sets to conduct 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 routine 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.

The access has been only partially achieved (see WP5.2 (2017) for details). Only different forms of aggregated mobile phone data have been compiled for the subsequent research (with limited exceptions in the cases of CBS, INSEE and Istat). Although this limits the scope of our results, there is still highly valuable information not only regarding the access to these data and the collaboration with mobile network operators (MNOs) but also regarding the new elements and features impinging on the production process at statistical offices.

To begin with, the unsolved issue of access to mobile phone data obliges us to concentrate mainly on proposals to process aggregated mobile phone data. We shall firstly analyze the process of generation of mobile phone data to adopt the most appropriate methodological approach. The nature of this new information source, in our view, introduces important subtleties in the way official statistics have been traditionally produced, but it also shares common features with other data sources already in use as

administrative data. The processing tasks from collection to final aggregate compilation can be divided into three stages. First, the raw telecommunication data cannot be processed for statistical purposes and they must be preprocessed to generate appropriate microdata sets[1]. Second, these microdata must be aggregated according to a prescribed methodology or set of algorithms to provide a kind of intermediate aggregates for each cell of a territorial division and each time division of a given time period. Finally, an inference exercise connecting these intermediate aggregates with the target population under study must be conducted. In chapter 2 we explain in detail this structure of the process and its motivation.

As stated above, we do not have access to microdata. In consequence, we cannot offer a full hands-on bottom-up view of this part of the process. However, partners of the project do have a limited access to call detail records (CDRs) thus allowing us to provide some insight. Complementarily, we have requested technical feedback to an external expert adviser with experience in producing statistics from this data source. As in preceding deliverables, we have asked the Estonian company Positium (Positium, 2018) to write a technical report regarding the microdata. Positium was involved in the feasibility study conducted by Eurostat and other stakeholders for the use of mobile phone data in tourism statistics (Eurostat et al., 2014). We have combined both contributions in chapter 3 to offer an initial view about the treatment of these microdata. In particular, we propose a core data model for the systematic exploitation of mobile phone data for statistical users and some proposed device location and aggregation procedures for these data.

The core of the methodological proposal regarding aggregated data revolves around their treatment in combination with official data to infer official figures for the target population of interest. We propose a hierarchical model to conduct this inference exercise illustrating its use upon toy simulated data with the same structure as actual data. This is undertaken using specific software tools developed for this purpose (although the IT and programming aspects are put off to the next accompanying deliverable). Full details are provided in chapter 4.

Finally, important conclusions and reflections for the future are collected in chapter 5. It is important to underline that our methodological proposal is not intended to be a closed methodology but rather on the contrary to open the way for a methodological framework with many possibilities to include more and more complex elements. We provide some insight not only about potential improvements in our model but also about relevant impingements on strategic decisions for the development of a modernised statistical business process with mobile phone data.

[1]Data at the level of each mobile device.

2

The statistical production process and mobile phone data

Executive summary

This chapter describes the whole generation of mobile phone data from raw telecommunication data to aggregated data for statistical purposes under the two-phase life-cycle model for statistical microdata proposed by Zhang (2012) for the use of administrative registers in the production of official statistics. This model details each of the steps for the constitution of statistical units and their variables (measures) so that every potential error is clearly identified. For mobile phone data, we claim that a third phase (or a double application of the model) is necessary to go from raw telecommunication data to statistical microdata to aggregated data. We identify each step in this process.

Complementarily, we briefly revise the statistical business process model in terms of level-1 subprocesses of the GSBPM (UNECE, 2013) providing a first view of potential changes in this model due to the use of mobile phone data.

The use of standard techniques for the design, development, execution, and evaluation of the production process with mobile phone data will hopefully allow us to normalize its usage for an integration with other Big Data sources, for a better reasoned choice of new methodological proposals, for addressing the issue of data access in a more structured way in terms of agreements with MNOs, and ultimately for a standardised quality assessment of final statistical products irrespective of the input data.

2.1. An overview of the generation of mobile phone data

In our deliverable 5.2 (WP5.2, 2017) we proposed a revised version of the definition of Big Data for use in the production of official statistics which underlines two important features beyond the traditional technical characterization in terms of volume, velocity, and variety. We stressed the extreme relevance for Official Statistics of two facts, namely (i) data do not contain information of the data provider but of third people and (ii) data play a central role in the business of the data provider. These two facts impinge essentially on the access and collection of these data in the production process.

In addition, we claim that there is another feature also impinging specifically on the processing: data are not generated with a specific metadata structure for statistical purposes. In the traditional model, a questionnaire is designed, implemented, and administered to respondents to collect their data under a very specific metadata structure. Every single variable is (ideally) given a rigorous definition and covers a concrete statistical need previously identified in the design of the survey. Indeed, the internationally adopted GSBPM (UNECE, 2013), although being a highly-nonlinear business process model, usually starts with the subprocess *Identify needs* and goes on later to design and develop tools for, among other things, the data collection.

With mobile phone data (as with many other Big Data sources), data have already been generated for very different purposes other than statistical production even before having identified their potential uses in the latter. Therefore no appropriate metadata structure for statistical purposes is included in the mechanism of data generation. Furthermore, data strictly generated for telecommunication services cannot be directly used for producing statistics. It is thus important to have in mind how these data are generated.

The first important feature which we should understand is that data about each mobile device are generated in an extremely complex and broad cellular network spread over a geographical territory (WP5.2, 2017). Raw telecommunication data are generated in Base Station Subsystems as a consequence of the electromagnetic interaction between each mobile device and the antennas (see figure 2.1). They are highly technical data, some of them being only temporarily stored. These data enter into a cascade of larger systems (the Network Subsystem and the Network Management System – see figure 2.1) both to establish a connection with other terminals and for billing purposes. Thus, we need to preprocess these data to generate so-called statistical microdata susceptible of further processing to produce statistics.

These statistical microdata have mainly four elements or attributes for each mobile device detected in the network (see WP5.2 (2017) for details). These are (i) the

pseudonymised identification variables of the mobile device, (ii) the spatial attributes (basically the coarse-grained position coordinates of each telecommunication event between the mobile device and the antennas), (iii) the time attributes (basically the initial and final time coordinates of the event), and (iv) complementary information about the type of event (call, SMS, ping, ...), position of the antennas, and some other data (see WP5.2 (2017) for details).

Statistical microdata can be analysed in many diverse ways. A look at the programmes of the series of conferences NetMob (2017) on the statistical exploitation of mobile phone data immediately suggests how many possibilities arise. Many different techniques can be used to pursue many different conclusions. We are going to concentrate on a particular sort of aggregated mobile phone data, i.e. those providing counts of individuals of a given target population of analysis (general population, inbound tourists, resident tourists, commuters, ...) per territorial cell and time interval together. Notice that the aggregation procedure taking us from microdata to aggregated data will have to be selected according to the target population, the target aggregates, and the statistical methods at hand. In the inference exercise proposed in later chapters we will take for granted this aggregation procedure.

Finally, aggregated data must be somehow linked to the target population under study, i.e. an inference exercise must be conducted. Notice that currently NSIs have access only to data from one or two MNOs at most (WP5.1, 2016). In any case, the link between the data and the target population must be undertaken as an inference exercise as in traditional official statistical production. Now, the way data have been generated and collected will raise the need for new methodology because traditional survey sampling cannot provide a rigorous footing (see section 4.2).

All in all, schematically the process can be represented as in figure 2.2. We follow the convention established e.g. in the Generic Statistical Data Editing Models by the UNECE (2016). Round elements denote datasets and square elements stand for process steps. We have also introduced a colour code to indicate those data sets and process steps upon which WP5 partners have currently some kind of total, partial, or null access for the present project. As seen in figure 2.2 no access whatsoever is currently granted to the raw telecommunication data originated in the network. Similarly, the preprocessing to produce statistical microdata is currently beyond any possibility to be undertaken by NSIs. Even the statistical microdata set itself is hardly accessed by NSIs. In no case has an agreement been reached for standard production conditions and only in some cases statistical microdata have been granted access to for specific research purposes. Furthermore, these have been shared between MNOs and NSIs only in limited conditions (attributes extremely coarse-grained to limit identifiability of mobile devices, CDRs and not signalling data, access exclusively on MNOs' premises

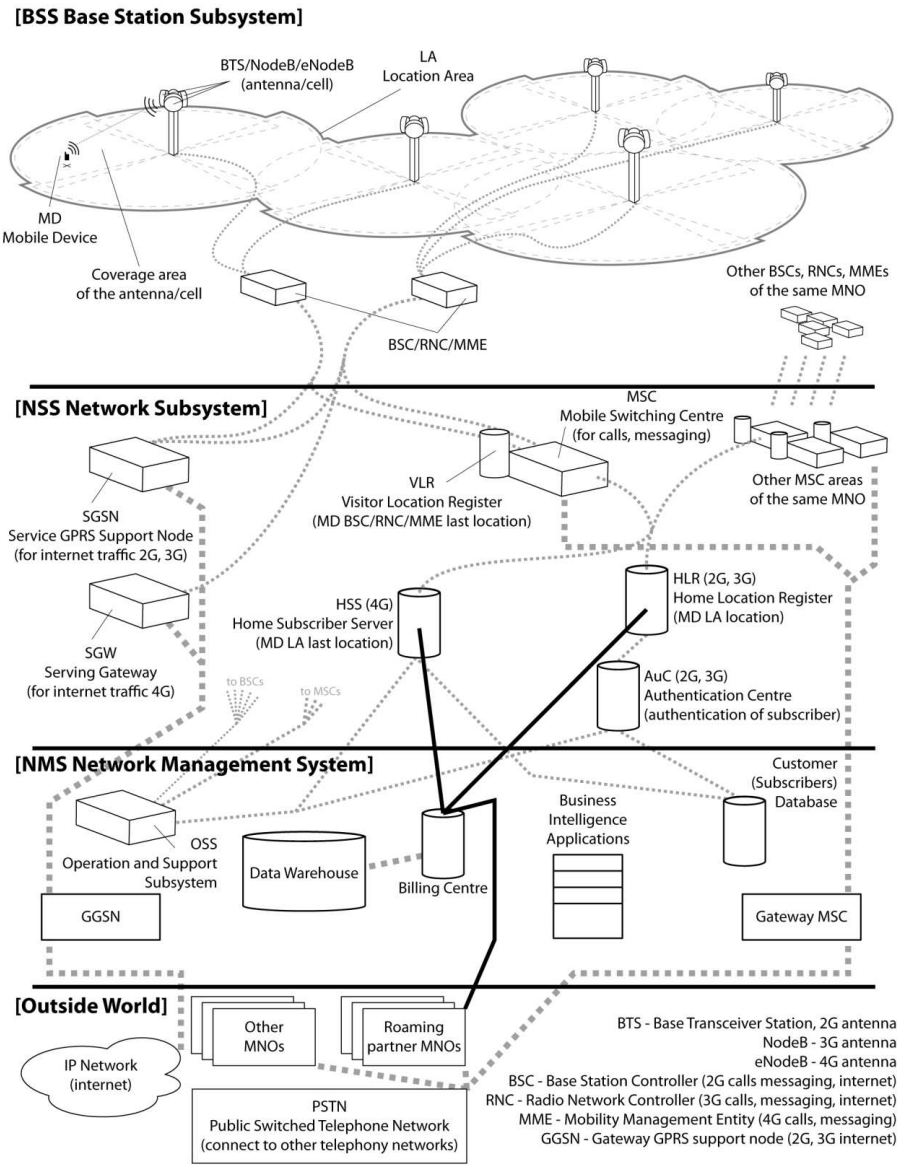


Figure 2.1 Architecture of a telecommunication cellular network (WP5.2, 2017).

or even only by their own staff, . . .). The process of aggregation upon these microdata however shows a slightly wider room for investigation and experimentation by NSIs with highly interesting issues (see chapter 3).

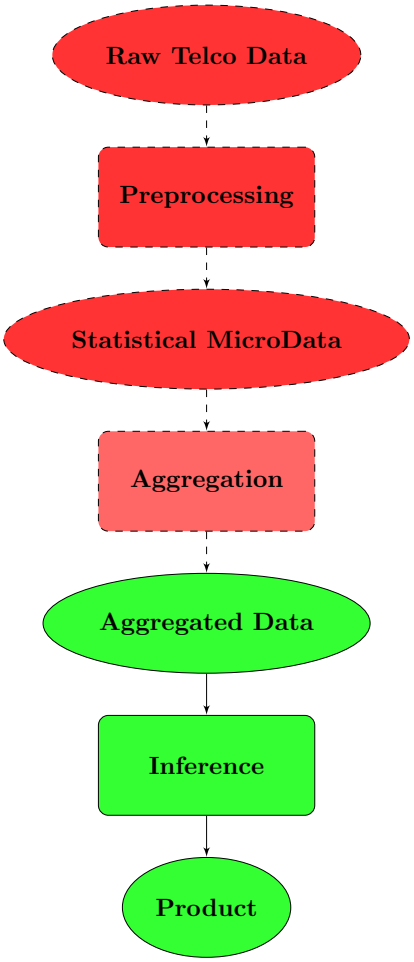


Figure 2.2 Sequence of large processing steps of mobile phone data.

2.2. The generation of mobile phone data and the two-phase life-cycle model

A cautious reader has probably recognised the generation of mobile phone data as a not so unfamiliar situation in NSIs. Administrative data do satisfy the three characteristics cited above: data refer to third-people and not to data providers, are central in their activity (public administration offices), and lack statistical metadata since they are generated for very different purposes. In consequence, one expects common concepts and methods to be shared between administrative data and mobile phone data. In our view it is not necessary to argue about the benefits of using common methodology for all potential data sources in the production of official statistics.

In this sense, we defend the claim that the two-phase life-cycle model of statistical microdata (Zhang, 2012) stands as an excellent tool to describe the generation of mobile phone data for statistical purposes and paves the way for an adequate analysis of their quality. We will not introduce the model, whose details the reader may consult in the work by Zhang (2012). The model entails the phases succinctly represented by figures 2.3 and 2.4.

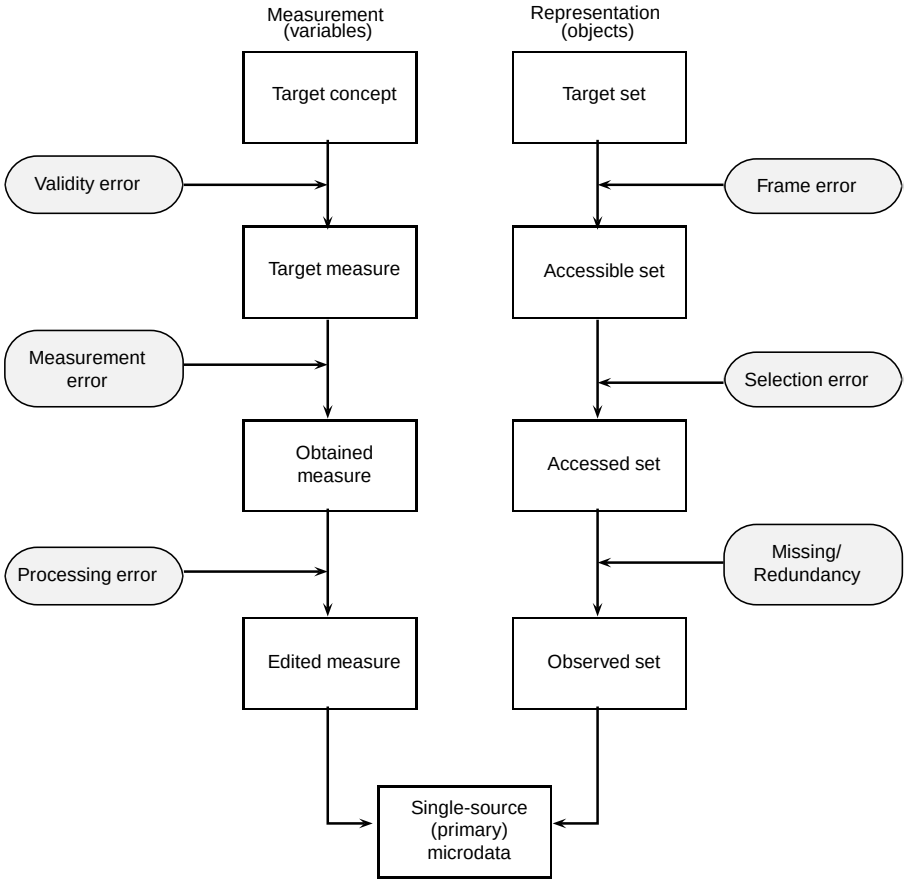


Figure 2.3 Phase I of the two-phase life-cycle model (see Zhang (2012)).

This conceptual model behind the generation of statistical microdata for the production of official statistics is highly adequate to understand the generation of mobile phone data in our context, although, as we shall see, we need a third phase. Our intuition is that even for other Big Data sources this model can be an excellent tool.

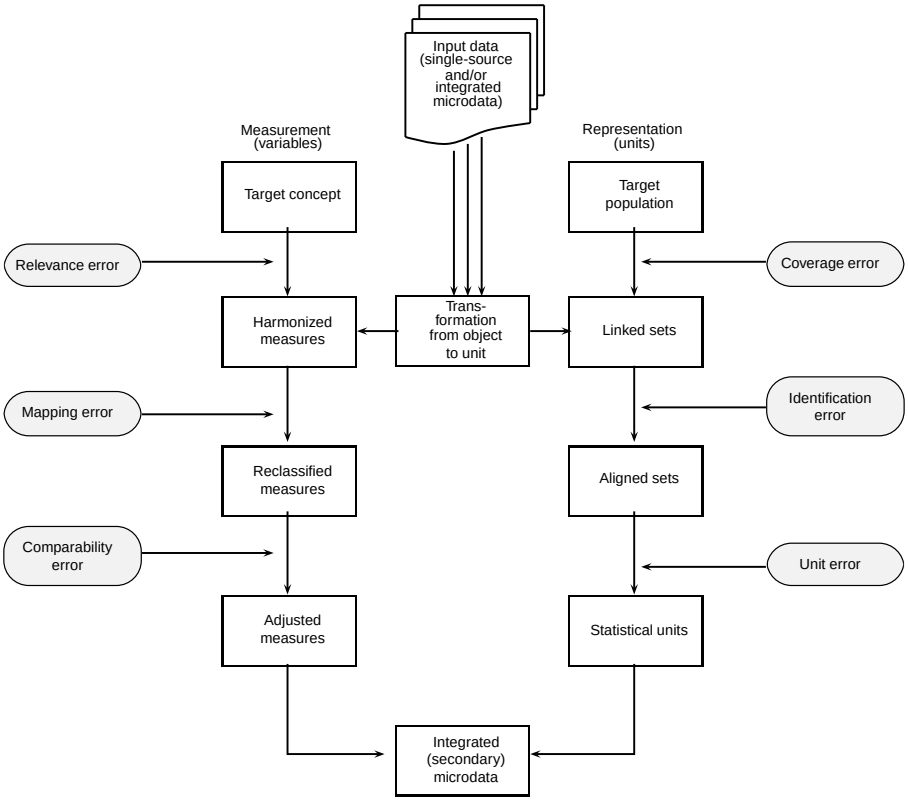


Figure 2.4 Phase II of the two-phase life-cycle model (see Zhang (2012)).

2.2.1. Phase one: raw telecommunication data

In the first phase the generation of the primary microdata amounts in our case to the generation of the raw telecommunication data. In the representation line, the objects are any kind of *intended* electromagnetic interaction between antennas and mobile devices to provide the interconnecting communication between users or between each user and the network. These constitute the *target set*. The *accessible set* comprises however the so-called events in the network, i.e. any interaction registered in digital systems including machine-to-machine communication. Notice that there will also exist elements of the target set not contained in the accessible set such as out-of-coverage attempts to establish the communication. Next one might expect that since no sampling is conducted the *accessed set* coincides with the accessible set. However, selection does occur in a setting for statistical exploitation. Different scenarios are possible depending on diverse factors

related to the access conditions and access agreement. Broadly speaking events can be divided into two sets, namely those originated by an active behaviour of subscribers e.g. initiating a call, sending an SMS/MMS, connecting to Internet through an app, etc. and those not originated by subscribers but only automatically by either the network or the mobile device for diverse technical reasons (handover, network load balancing, ...). The former are usually subjected to billing according to some sort of commercial contract; the latter are not subjected to any kind of billing process. Depending on the infrastructure deployed by the MNO for the statistical exploitation of their data, either only the former are exploited or both on them are processed to compile a gigantic database. In addition, in both cases machine-to-machine events are currently not considered. Thus, selection does occur producing the accessed set. If we now add potential technical problems in compiling this set (antenna shutdown due to punctual technical problem, network connections damaged, etc.), we arrive at the *observed set*.

Regarding the measurement line, the *target concept* comprises the set of attributes allowing the MNO (i) to technically establish the connection between mobile devices and each mobile device and the network and additionally (ii) to accordingly bill its subscribers. For strictly telecommunication purposes this set of concepts is clearly technology-dependent. For example, many services offered with 4G mobile telecommunication cannot be offered with 2G technology. Consequently the set of concepts will differ. However a set of core attributes will be common for the posterior statistical exploitation. These are basically the identification of interconnected users, the spatial attributes (location) of the devices to set up the connection and potentially related attributes to establish this location, the time attributes, and characteristics of the events (call, SMS/MMS, Internet connection, duration, roaming, etc.). The operationalization of these target concepts can be undertaken straightforwardly to constitute the *target measures*. The *obtained measures* result from the application of the technological solutions at stake and depends directly on the reliability of the involved engineering systems. Finally the *edited measures* can be also thought of as part of the whole engineering system to detect and correct potential errors. Thus we arrive at the primary data set comprising raw telecommunication data.

2.2.2. Phase two: statistical microdata

It is important to remark that currently only MNOs have access to their own sets of raw telecommunication data. With high probability, this situation will not change in the foreseeable future due to privacy concerns, legal regulations, and business activity protection by MNOs themselves, among other factors. In this sense, it must be clearly stated that currently using mobile phone data for official statistics production entails the inclusion of MNOs in the statistical production process not only as data providers

but also as active data processors. The situation is similar with administrative data: see principles 8.7, 8.8, and 8.9 of the European Statistics Code of Practice (ESS, 2011) assessing the capacity of MNOs to influence the design of administrative registers for statistical purposes.

Let us elaborate on this statement with the second phase of the life-cycle model. Now raw telecommunication data enter as primary input data into the second phase. As a consequence of the restricted access only MNOs will be able to process their own raw telecommunication data. In this sense, an integration exercise between different primary input data of this sort from different MNOs is currently impossible.

To identify the *target population* in the second phase we set as main statistical output the compilation of statistical microdata for ***persons***. That is, the statistical units will be individuals. For the *target concept* we will extract as many variables as possible for each individual from the raw telecommunication data. Regretfully these data cannot be accessed by official statisticians and we are partially blind about the actual potentiality of generation of these variables. Nonetheless, as a core set of variables we may clearly conceive (i) an identification variable for each individual, (ii) a collection of space-time attributes for each event registered in the network associated to each individual, (iii) all related information regarding each event (type, duration, ...). Technology in the future will probably impinge on more possibilities to compile more information. Occasionally this information can be complemented with sociodemographic information from the personal contract between subscribers and the MNOs.

Now the transformation from objects in the first phase (events) to units in this second phase (individuals) must be undertaken. As already pointed out by Zhang (2012), final statistical units may not be directly obtained in this transformation. In our case we need to produce intermediate units as we shall immediately see. Conceptually the starting point is the primary microdata set in terms of events, which in a first step must be transformed into an intermediate data set in terms of mobile devices. This can be schematically represented by table 2.1.

For rigour's sake it is important to point out that the notion of device ID is a bit more complex than what figure 2.1 may suggest. The interplay among the identification of SIM cards ID, physical device ID, and subscriber ID is a subtle issue which we take for granted here (see WP5.2 (2017)). As *linked sets* we obtain a data set in terms of mobile devices and complementarily additional information per mobile device (sociodemographic variables, data coming from the contract, ...). Again NSIs are currently blind regarding the actual availability and structure of potential information, but a priori the linkage exercise in terms of mobile device ID is clear. The aligment exercise is now conducted with these data sets to obtain the *aligned sets*.

eventID	**deviceID**	**attr1**	...	**attrN**
1	1	...	...	...
2	1	...	...	...
⋮	1	...	...	...
n_1	1	...	...	...
$n_1 + 1$	2	...	...	...
$n_1 + 2$	2	...	...	...
⋮	2	...	...	...
$n_1 + n_2$	2	...	...	...
$n_1 + n_2 + 1$	3	...	...	...
⋮	3	...	...	...

Transformation
object to unit

deviceID	**eventID**	**attr1**	...	**attrM**
1	1	...	...	...
	2	...	...	...
	⋮	...	...	...
	n_1	...	...	...
2	$n_1 + 1$	...	...	...
	$n_1 + 2$	...	...	...
	⋮	...	...	...
	$n_1 + n_2$	...	...	...
⋮				

Table 2.1 Transformation from objects (events) into intermediate statistical units (devices).

Up to this point, currently all the processing must be carried out by each MNO. Now, depending on the access agreement between the NSI and the MNO, aligned sets in terms of mobile device IDs can be made available for processing by NSIs. In this way the final step in the line of representation regarding the constitution of the *statistical units* (individuals of the target population) can be undertaken under control of official statisticians (either on our own premises or on MNOs' premises). This identification step is schematically represented in table 2.2.

deviceID	**eventID**	**attr1**	...	**attrM**
1	1	...	...	...
	2	...	...	...
	⋮	...	...	...
	n_1	...	...	...
2	$n_1 + 1$	...	...	...
	$n_1 + 2$	...	...	...
	⋮	...	...	...
	$n_1 + n_2$	...	...	...
⋮				

Identification
of individuals

individualID	**deviceID**	**attr1**	...	**attrQ**
1	1	...	...	...
	2	...	...	...
2	3	...	...	...
⋮	...	...	...	...

Table 2.2 Identification of statistical units.

Due to the lack of access to these data we cannot provide either fully-fledged recommendations nor complete guidelines for this constitution. However very simple situations invite us to suspect about potential difficulties. Let us consider those individuals in the population with three mobile devices (home and work mobile phones and a personal tablet with a SIM card). Depending on the information provided by the MNO the events produced by these three devices may not be clearly linked to the same individual. Thus identification of units will have to be deduced from the aligned sets themselves (e.g. with analysis of trajectories).

As an additional comment notice that the nature of the identification variables set up in the target concepts may present two extremely different situations. Had we identified each event in the primary data set with an explicit individual identification variable set (name, surname, address, tax identification number, ...), a universe of linkage with many other data sets would potentially arise, especially with external identified official microdata sets. In this case the constitution of units would be partially straightforward and would follow along similar lines to the traditional methodology. However, even in this case mobile devices linked to organizations will keep on posing serious problems. Moreover, this case would also face the difficulty that this linking exercise would have to be conducted by official statisticians under strict privacy and statistical secrecy conditions (the dissemination of identified official microdata sets are strictly prohibited by law). Nonetheless, accessing fully identified mobile phone microdata is currently out of the question in the ESS, thus we discard this possibility.

On the other hand, with pseudonymised identification data, the linkage exercise with external microdata sets in terms of individuals seems to be virtually impossible. However, linked sets for a number of attributes may be relevant for further analysis. For example, the spatial attributes may be linked to data sets with information about the land use which might provide useful information for statistical analyses. In the same vein, the time attributes may be linked to data sets with a work calendar providing further insight into the analysis.

In the measurement line the *target concept* set comprises as many variables as possible for each individual in the target population. The concrete content of this set depends sensitively on the underlying technology and the information made available by the MNO both during the first phase and as external data sets potentially to be used as linked sets. Nonetheless, we can identify the following core set of variables: (i) univocal ID variable for each individual, (ii) space and time attributes for each event associated to each individual, and (iii) auxiliary information about each event (type, duration, ...).

The *harmonised measures* point towards the normalization of these variables. In particular, for the foregoing core variable set the ID variables, the space-time attributes, and

event information should follow suitable international standards. For example, the time attribute may follow the norm ISO8601 (ISO, 2004). Equivalently, the spatial attributes may similarly follow the norm ISO19111:2007 (ISO, 2007). Following internationally accepted standards will not only allow ESS stakeholders to increase comparability, efficiency, ... in their production and/or use of official statistics but will also allow us to integrate more seamlessly future technology-originated information impinging on the generation of these data sets.

The application of these standards upon the primary input data will produce the *re-classified measures* by which attributes in the latter will be expressed under these standards. Finally *adjusted measures* are obtained after detection and correction of errors, especially those arising from the combination of different sources.

As a result of these two phases, an integrated secondary microdata set in terms of individuals is obtained. We will refer to this as the statistical microdata.

To finish this section, let us point out how the definition of individuals in the target population has been intendedly diffuse. By individual in a target population we may well refer to resident tourists, inbound tourists, commuters, ... Conceptually the process is then to be repeated for each different target population which we define. However the situation is similar to that of current frame populations at NSIs for different surveys. For obvious efficiency reasons a generic register either of human population or of business and corporations is maintained in the office. Then each frame population is identified as a subset thereof without the need to compile a completely new one for each survey. In this same vein, an exhaustive complete secondary microdata set of individuals is to be created out of which an appropriate subset thereof according to the definition of target population at stake is selected for producing the statistics. The core data model proposed in section 3.2 points in this direction.

2.2.3. Phase three: aggregated data

Currently, no access to these secondary microdata sets has been granted in the ESS for NSIs (except partially for CBS, INSEE and Istat under very stringent conditions and for a highly limited volume of data). These microdata are further processed by MNOs themselves to produce aggregated data sets with counts of individuals of the target population in an agreed geographical division of the territory and in agreed time intervals. These have been currently shared with NSIs for the investigation of the production of official statistics. A partial goal of this research is to elucidate whether they are enough or access to microdata (either primary or secondary) is necessary. The current description in terms of the two-phase life-cycle model will allow us to provide a

clear reasoning in this sense.

To describe the generation of aggregated mobile data sets, it seems clear that this model is also extremely useful so that a third phase is added to complete the description of the whole generation process for those data entering as input in the NSIs. Now the structure of the second phase of the model is again followed.

In the representation line the target population may be chosen between two options with increasing complexity. On the one hand it may comprise the territorial cells per time interval unit. On the other hand it may focus on the elements of the transition matrix between each pair of cells per time interval unit. Notice that in any case this entails to choose both the territorial division and the time partition. The linked sets comprise the statistical microdata set(s) as well as any other information regarding either the geographical cells (land use, extension, ...) and/or the time partition (work calendar, working hours, TV prime time hours, ...). Now alignment among these data sets is carried out to clarify all the relevant relationships between units thus obtaining aligned sets. Then statistical units are created which amounts to identifying each territorial cell per time interval unit.

In the measurement line, as target concepts we set two sorts of variables. On the one hand, we focus on the totals of individuals of the secondary target population (in the second phase) per statistical unit (either cell per time interval unit or cell to cell per time interval unit). On the other hand, additional attributes for each cell and/or time interval are targeted according to the available information. As in the second phase, now international standards must be pursued as much as possible regarding all geospatial- and time-based information. For instance, for land use there exists *land based classification standards* (APA, 2018). Next, secondary input-source measures are turned into re-classified measures following these harmonised measures. Notice that this includes some kind of aggregation procedure to go from individuals to totals, which is a fairly subtle step (see section 3.4). The familiar editing and imputation activities now will produce the adjusted measures to finally arrive at the aggregated data set.

2.2.4. Combining data from several MNOs

No explicit mention has been made so far to the case when several MNOs provide data to an NSI. The structured approach supplied by the two-phase life-cycle model for the generation of mobile phone data allows us to analyse the combination of data from several MNOs in a seamless way.

According to the modelled process above everything is reduced to choose the step of the process in which we can integrate the several data sources. Notice that it is now

critical to decide when in the entire process will NSIs have access to data. Several possibilities arise.

Firstly we discard the possibility of having access to raw telecommunication data given the present conditions to access. Secondly if only statistical microdata are to be combined, then it is clear that they will jointly enter the third phase as input data so that at the end the final set with statistical units and adjusted measures will contain the information from both sources. Notice now the relevance of the standardization in the choice of harmonized measures in all phases of the generation process.

Thirdly if both microdata and aggregated data are to be combined, we need to generate the corresponding aggregated data set from the former and then reduce this case to that of combining two or several aggregated data sets.

Finally, this combination of aggregated data sets implicitly implies that the respective generation processes are over. If the preceding stages in the generation of each set have been completed successfully, the combination of these two sets should amount to recomputing the adjusted measures using the input aggregated variables (e.g. summing the total number of individuals per cell and time interval unit).

2.3. The statistical business process and mobile phone data

Having described the generation of mobile phone data in a structured way, the next obvious question is how these data enter into the statistical production process to produce official statistics in standard conditions. One of the goals of the project is to follow a hands-on bottom-up approach to produce a concrete statistical output using real data to assess this question, among others. The current agreement on the access to mobile phone data for all NSIs in the project with their respective MNOs clearly limits the use of the results exclusively for research purposes under the current investigation. This entails that many aspects have not been empirically explored. However, diverse facets can already be commented. We shall formally do this using both the GSBPM (UNECE, 2013) and the foregoing model for the generation of data.

A priori we see no reason why the standard GSBPM cannot be used to describe the production process with this new data source. We shall use the model at the first level to comment on the production with mobile phone data. Being a highly modular and nonlinear model whose different functional modules are to be amalgamated by the process designer according to the needs of the statistical operation, we will comment separately on each level-one phase.

Regarding the needs to specify, mobile phone data stand as an extremely promising source to produce both traditional results at an unprecedented geographical and time scale and completely novel statistical outputs. Which needs can be potentially satisfied by this new data source can be fully and empirically assessed only by having access to both statistical microdata and aggregated data. Pursuing the bottom-up approach but keeping some level of generality we have focused on the production of population counts at different geographical and time scales. The target population is not reduced to the general population but it is extended to comprise populations in tourism, mobility and any other potentially identifiable population through mobile phone data. But this is certainly not the unique output we can aspire to. A light view to the programmes of the conference series NetMob from 2012 (NetMob, 2017), in which contributing researches from many centres have access to some kind of statistical microdata, suggests an immediate idea of the openly wide possibilities. Thus here we have another reason why NSIs should have access to statistical microdata.

The design modules mainly concern the development of methodologies for the statistical production. For mobile phone data this is the core content of the present deliverable. Again our focus is completely conditioned by the limited access to data. We have identified two priorities to integrate mobile phone data in the production of official statistics. Firstly, noticing the limitations of traditional design-based methodology to make inferences about a target population with this data source, it is compulsory to address this issue by assessing alternative non-probability sampling techniques. In chapter 4, inspired by ecological sampling techniques, we propose a hierarchical model to estimate population counts as a generic framework to address this issue, but this is certainly not the unique possibility. Secondly, the transformation from objects to units in both the second and third phase of the generation process is absolutely key for the quality of the final aggregated data to be used as inputs in the inference step, hence of the final estimates. Two concrete aspects are addressed in chapter 3. On the one hand, the computation of the geospatial attributes for statistical units (individuals) in the second phase is a delicate question in particular using the limited data provided by MNOs. On the other hand, this limitation again arises in the third phase when computing totals for each statistical unit (cell per time interval unit). This does not mean that there is no further issues to be tackled or that they are trivial. For example, the collection methodology (which should take into account the generation process in the preceding section) has not been addressed because it heavily depends on the (as yet unsuccessful) access agreements with MNOs. Also, some aspects of the design of outputs become newly relevant such as the visualization of results. Achieving detailed breakdowns of outputs bar the dissemination of results in terms of traditional tables and new visualization needs arise. As an illustration let us consider a grid net of $1km^2$ covering the European territory. This will contain over 4.8 million grid cells. If the time interval unit is set to 30 minutes (to observe commuting patterns, for example),

then in a one-month period we will have 1400 time periods per each cell. Tables are useless to visualize and to have a first comprehension of the results. Novel visualization techniques (two-way interactive maps, videos, etc.) will need to be incorporated in standard production conditions. Tables will still be of interest for research and further analysis, however.

The build modules follow accordingly the same trend as the design phase. Each new methodological development (either for inference or for collection or for dissemination) will need the corresponding tools. As with many other Big Data sources the critical issue regarding IT tools revolves around the need of complex architectures involving clusters and parallel computation together with the corresponding software and management tools. IT is the theme of deliverable 5.4. Nonetheless, as a generic principle, we want to express our idea that it should be the statistical methodology which determines the computational needs and hence the optimal IT tools to use and never the other way around. For example, to process aggregated data you may not need a highly distributed system with a ultimate file system[1], thus the complexities derived from new architectures may not be taken into account. Having prioritised the inference issue and the computation of geospatial attributes and corresponding aggregation of units, we will focus on the tools to implement our proposed solutions for these priorities.

The modules about collection again depend heavily on the issue of accessing the data. Currently this is completely open. Notice that our description in terms of the two-phase life-cycle model can facilitate the issue. Are we going to access primary microdata sets? Statistical microdata sets? Only aggregated data sets? Or even some intermediate data set (accessible, accessed or observed sets with their corresponding measures from the first phase or linked/aligned sets from the second or third phase? All these aspects will determine how data collection will be run.

The core of the process phase is the data integration, coding, data editing and imputation, and calculation of aggregates. It is currently very early to assess the full adequacy of the GSBPM for the new statistical methodology (either our proposals or alternative ones), but we call the reader's attention on the fact that some of the modules (*5.6. Calculate weights*) explicitly rely on design-based methods for inference. We will put off our brief assessment of this issue until deliverable 5.5. on quality when the application of our proposals to real data are undertaken and analysed.

In the analyse phase two core tasks arise. On the one hand, the preparation, interpretation, and explanation of outputs depend strongly on the visualization needs and the complexity of the outputs. On the other hand, the statistical disclosure control reaches

[1]Perhaps not even a client-server architecture.

higher levels of intricacy. The reasoning is clear: the more data you have, the higher the risk to identify statistical units. As it is evident, this issue (methodological in nature) cannot be solved until concrete outputs with real data are at hand.

The dissemination phase is completely soaked with the new visualization needs. This phase will need to be executed according to the visualization solutions provided, which will certainly involve a technological twist in the dissemination tools. As this is a methodological document we do not lose the opportunity to advise on the misuse of visualizations to address statistical issues such as precision, accuracy, and quality in general. Modern visualization techniques constitute no method to assess mathematical results. They must just be appropriately used to disseminate and communicate results with an increasing complexity.

Finally the evaluation phase, as in traditional production, is fairly entangled with the overarching quality assessment. This will be undertaken in the deliverable 5.5. monographically devoted to these issues. A challenge arises in quality assessment derived from the new methodology. In any case, quality of official statistical products is and will be the ultimate goal of official statistical production.

3

From statistical microdata to aggregated data

Executive summary

This chapter addresses the process going from the statistical mobile phone microdata, i.e. data in terms of mobile devices and individuals, to aggregated data to be used as input in the inference exercise connecting them with the target populations of analysis.

The main outputs of this chapter are:

- As a key ingredient in the generation of microdata, we include proposals to assign the spatial attributes to each mobile device. After briefly mentioning the widely use of Voronoi tessellation (not taking into account either the directionality of antennae or the overlapping of coverage areas), we present both the Best Service Area approach (taking account directionality) and a Bayesian approach exploiting the signal strength (taking into account both directionality and the overlapping nature of cells).
- A core data model to constitute a normalised database with statistical microdata. This is intended to play a similar role as population and business registers at NSIs do in the traditional production process. These registers allow statisticians to create frame populations to apply the well-known design-based methodology. For the integration of mobile phone data into the statistical production process we propose the creation of a generic database with standardised definitions and variables so that for each different statistical domain of interest a minimal further processing will be needed.
- Once statistical microdata have been duly generated, the key step for the aggregation of these into indicators per territorial cell and time interval unit is to be carried out. We briefly revise these aggregation procedures for two statistical domains (mobility and tourism) complemented with our own experience with data compiled during the first phase of the current project.

3.1. Introduction

As stated in the preceding chapter, in terms of the two-phase life-cycle model no access to the generation of primary microdata is currently possible for NSIs. Only in three cases (CBS, INSEE and Istat) and under limited conditions do we have direct access to these microdata sets. For this reason, technical advice has been requested to an external expert to partially address the generation of statistical microdata and aggregated data. This chapter is mostly based on a technical report composed by Positium (2017) for this specific purpose. This has been complemented with our experience with the few primary microdata sets at our disposal and novel methodological proposals for the computation of some variables (spatial attributes especially) in the generation of both statistical microdata and aggregated data.

The complete process of generation of both the statistical microdata set and the aggregated data set is complex. We shall focus for its remarkable importance on the computation of the spatial attributes in the intermediate linked set in terms of mobile devices and in the computation of its measures. Although the methodology depends on the available data and the background of the data processor, this chapter aims to consolidate the experience from working with mobile phone data into one single methodology that could potentially be used in several countries. Obviously some aspects of the methodology depend on local circumstances, specifics of the data provided by the MNOs, statistical domain specificity and available resources, but in general, the approach presented in this chapter should provide the most simple option for processing the data for several statistical domains (hence different target populations) and be comparable internationally.

3.2. Computation of spatial attributes: geolocation of network events

We begin by focusing on the computation of diverse spatial attributes for each mobile device. There is a strong argument for designing and conducting joint processing of the primary data up to a specific point where different domain-specific processes continue. For example, if the objective is to calculate inbound tourism statistics and de facto population statistics, then both domains include an indicator about foreigners' visits, which should be processed using the same methodology.

In this line we shall call a *core data model* an approach representing a method for processing the data so that the first part of the processing is the same for all target populations, thus including all of the basic data processing steps required for all of them.

The joint processing results in a core data set which becomes the source data for all subsequent specific processes of diverse target populations. This method may be time

consuming in terms of developing the algorithms and data processing but is definitely efficient if we aim at a number of different target populations which can be used in very different domains. Core data should be prepared so as to correspond to the UML schema depicted in figure 3.1.

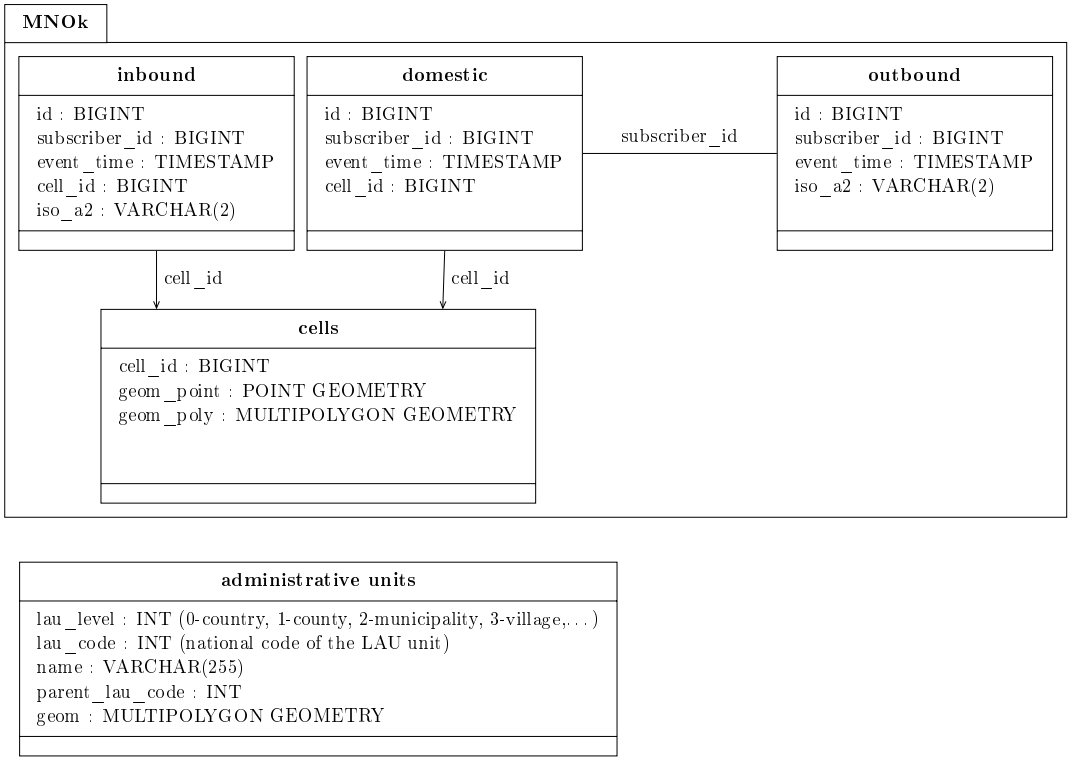


Figure 3.1 UML class diagram of a simplified data structure including three data forms (inbound roaming, domestic and outbound roaming data), antennae table and administrative units table as reference data.

3.2.1. Spatial Interpolation

Before data processing for the creation of the core data set some preparations need to be made. This is of utmost importance with no regular solution among all NSIs since they access different geographical attributes in their data.

3.2.1.1. Deciding the Smallest Geographical Unit and Accuracy Level

Before setting up the system, it is important to decide what smallest geographical units will be used in the data processing. This depends on the smallest geographical breakdowns that the system has to provide as results. As a first working proposal (according to Positium's recommendations) the third level of local administrative units (LAU3) will be used as an example as the smallest geographical unit to which all data will be spatially aggregated during the processing. Alternative options are to use higher LAU units (LAU2, LAU1), or lower level (1 km^2 grid, adaptive grid, etc.). For outbound data, the country will be a smallest geographical unit because usually it is not possible to get more accurate location of the person in a foreign country.

3.2.1.2. Spatial Interpolation Preparations

In the primary data we have the locations of network antennae with coverage areas presented as (multi-)polygons. As reference data, the geographical layer of administrative units should also be prepared. Very many antennae coverage areas might extend over several administrative units, especially with smaller and lower level administrative units. As the geographical accuracy of the location events is limited by the coverage area of the antennae, it is impossible to identify where exactly the location event took place. Therefore, geographical interpolation from antennae coverage area to local administrative units may be necessary. This is justified, if the expected end results are presented in the lowest level local administrative units but it is a subject of decision, as spatial interpolation of the location events is rather complicated and resource-consuming process. There are mainly three options for spatial interpolation:

1. Direct interpolation from coverage areas to lowest level local administrative units using an area proportion method;
2. Direct interpolation from coverage areas to lowest level local administrative units using an area proportion and land use method;
3. Interpolation from coverage areas to grid and then to local administrative units using area proportion and land coverage functionality method.

A single network event can be accurately located with no single spatial interpolation method whatsoever. The effect here is of statistical probability. Options 2 and 3 require additional reference data. Land use data could indicate in which part of the antenna coverage area people are more probably present. With the land use data, one can assume that people are present in some locations more probably than in others (e.g. people are more probably moving in the road than in forests/fields/swamps).

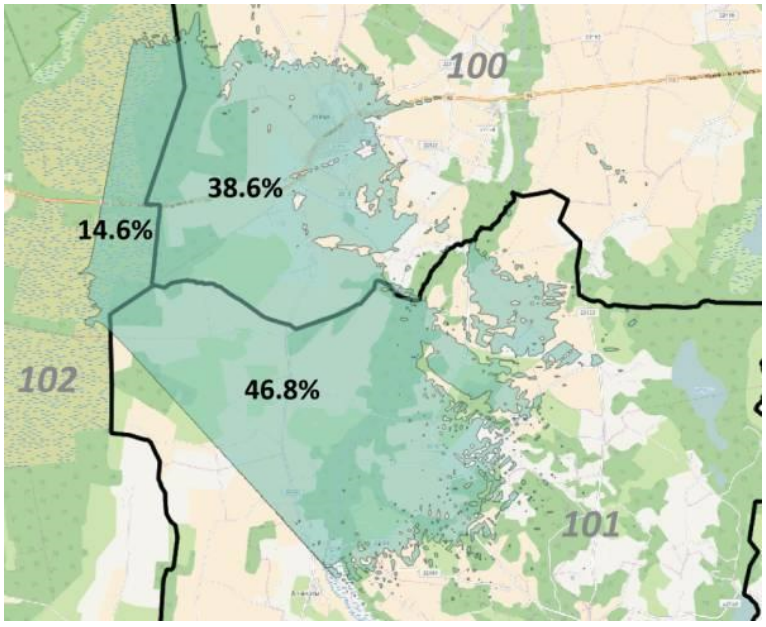


Figure 3.2 Example of spatial interpolation base for one specific antenna covering three different local administrative units (100, 101, 102) with area proportion distribution (option 1).

For option 3, an intermediate grid is used for interpolation target before aggregating to LAU unit. The grid should be kept as a geographic unit throughout the processing before aggregating to LAU units.

With any option, the processor of the data should prepare an interpolation method and the base data for the interpolation before the location event data is processed in next steps. The actual interpolation will take place during the construction of the core data set.

If data is regularly updated, so should the coverage area data, as the network changes, coverage areas change, new antennae are added and existing antennae removed. This spatial interpolation basis should be updated as often as the data updates occur.

As the smallest geographical unit used in this proposal is LAU3, all location events should be interpolated to individual LAU3 codes based on the spatial interpolation principles described above. Obviously the interpolation can be done only for inbound roaming and domestic data, as outbound data does not have identified cells. See figure 3.3 for an illustrating example.

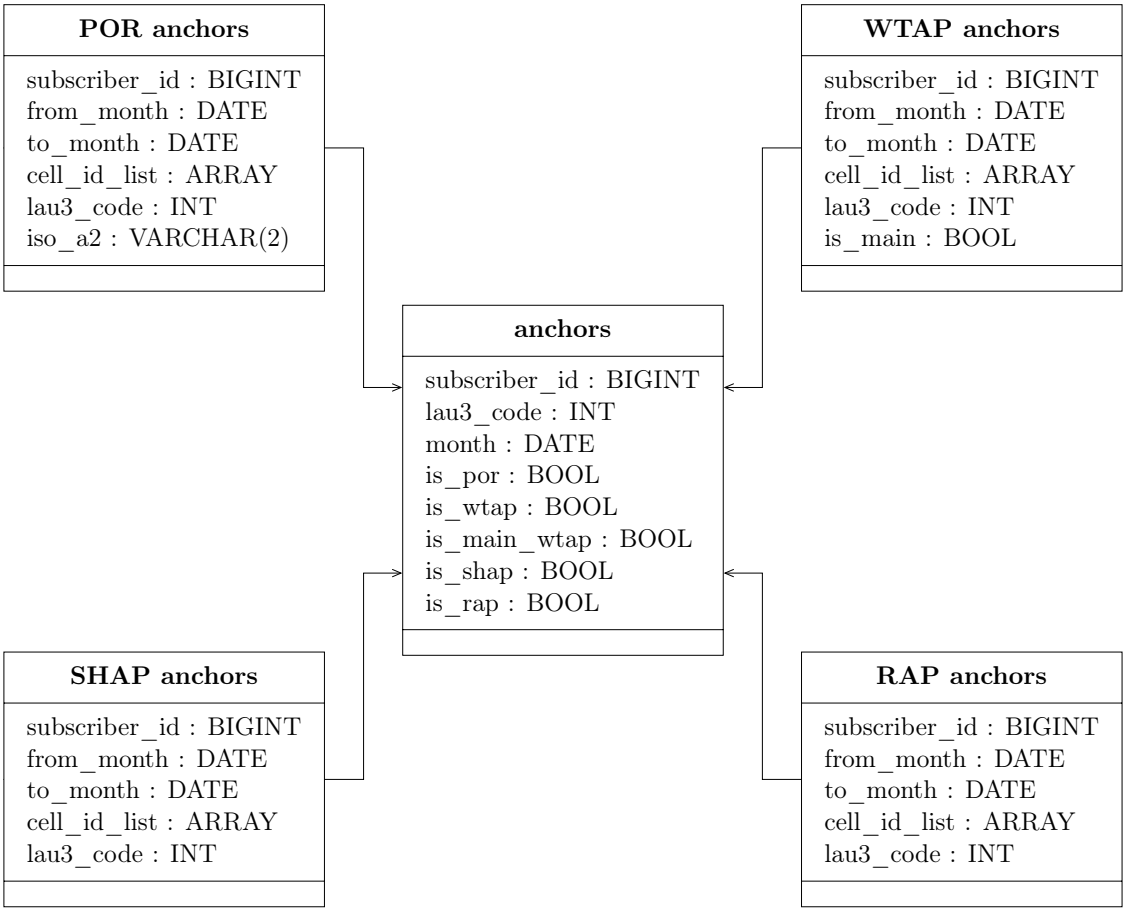


Figure 3.3 UML class diagram of the resulting spatially interpolated location events data.

3.2.2. The Best Service Area approach

This is an approach followed by ISTAT which illustrates how CDRs can be processed to assign the spatial attributes. In general, the CDR schema contains a variable code for the receiving and/or transmitting antenna sector; this information must be accompanied by details relating to the antenna or sector and the transmission technology. This information can be provided at various levels of detail, such as:

1. only the coordinates of the antenna tower, as described in section 2;
2. the parameters that characterize the antenna sector and morphology of the surrounding area, as described below;

3. an expected antenna coverage area calculated by the MNO on the basis of the antenna characteristics, the transmission technology and the morphology parameters of the area covered, as described in immediately below.

The latter information is under the complete control of the MNO and represents a tessellation of the territory, i.e. it does not include overlapping areas (as obtained by the method described in section 3.2.3 below).

3.2.2.1. Description of tessellation via Best Service Areas[1]

The Best Service Areas (BSAs hereafter) are a partition of the territory defined by the MNO in order to plan and manage the radio base stations (BSs) of the mobile phone network in the most efficient way (i.e. guaranteeing a suitable quality of service). Actually, the BSA are defined via the combination of models able to predict the coverage of 3G and 4G networks with computationally-efficient optimizers in order to automatically configure large networks and achieve optimal performances in terms of throughput, served users, and bandwidth re-use. Generally speaking, a BSA represents the area where the signal measurement of a certain antenna sector has the best coverage. In figure 3.4, an example of a BSA derived by three sectors of BTSs is provided.

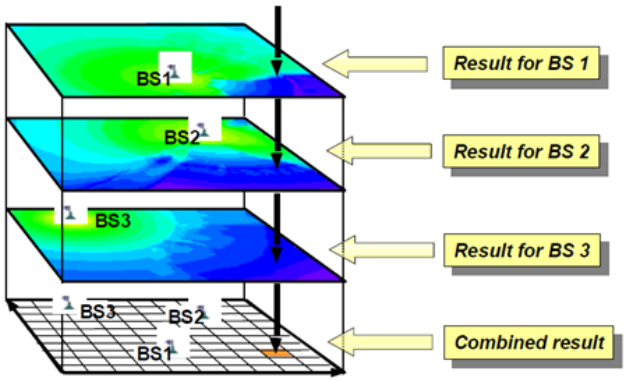


Figure 3.4 Example of BSs and BSA: the method of combining results for BSs determines the BSA.

The antenna cells are divided into several sectors. Each sector is a service characterized by a technology, a direction and a coverage area of antenna, and this area is named Service Area. An example is provided in figure 3.5.

[1]Prepared in collaboration with Francesco Altarocca and Raffaello Martinelli.

It is worthwhile noting that the size of the BSA is a function of the technology and of the land use, as illustrated in figure 3.6.

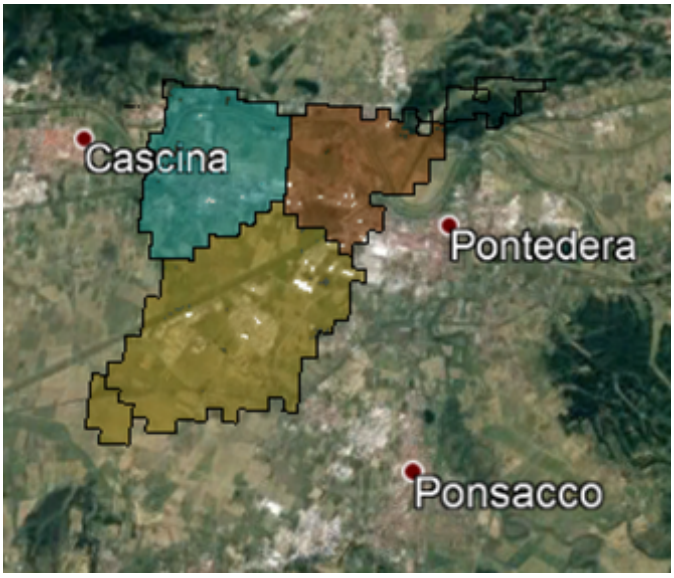


Figure 3.5 Best Service Area of an antenna. Example of three Best Service Areas in 3G technology. The antenna tower is located at the intersection of the 3 areas.

It is interesting to note that the area covered by 4G technology is really wide, as illustrated in figure 3.7, where a BSA for 4G technology covers 4 LAUs.

3.2.2.2. Mapping CDRs over Local Administrative Units via BSAs

As described above, the BSAs provide a tessellation of the territory, distinguished by technology. For instance, in figure 3.8 a LAU is represented as covered by 2G technology, on the left, and 4G technology, on the right.

In the data provided by the MNO, the BSAs are characterized by a unique identifier (ID_sector) and a shape file. The ID_sector, available in the CDR records, allows geo-referencing the CDRs. The shape of each BSA can be compared to the shape of the LAU, so to evaluate the percentage of coverage. This operation can be done for different territorial levels (e.g. LAU2, LAU3, ...) as exemplified in figure 3.9.

Elaborating the shape file with ArcGIS it is possible to calculate (i) the municipality area ($A_{\text{municipality}}$), the BSA area (A_{cell}) and the percentage of overlap (BSA_FRACT) between the BSA and the area of the municipalities:

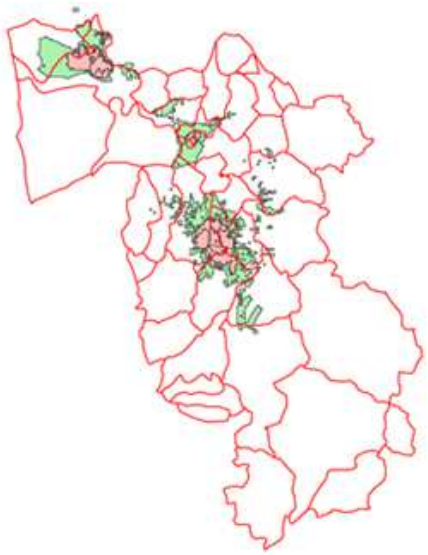


Figure 3.6 Three antennas with different services. Example of three different antennas with their Best Service Areas, the areas with 3G technology in pink (colored) and with 4G in green.

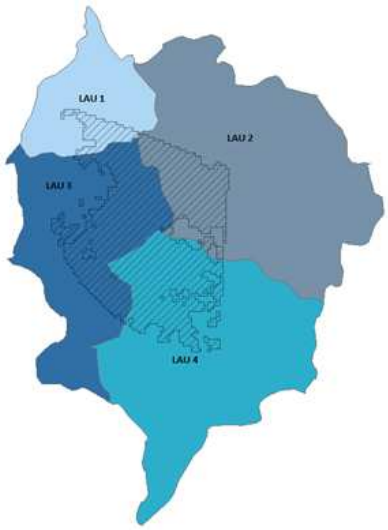


Figure 3.7 Example of BSA covered by 4G technology.

$$BSA\_FRACT(ID_{\text{sector}}, LAU) = \frac{(A_{\text{LAU}} \bigcap A_{\text{IDsector}})}{A_{\text{IDsector}}}.$$

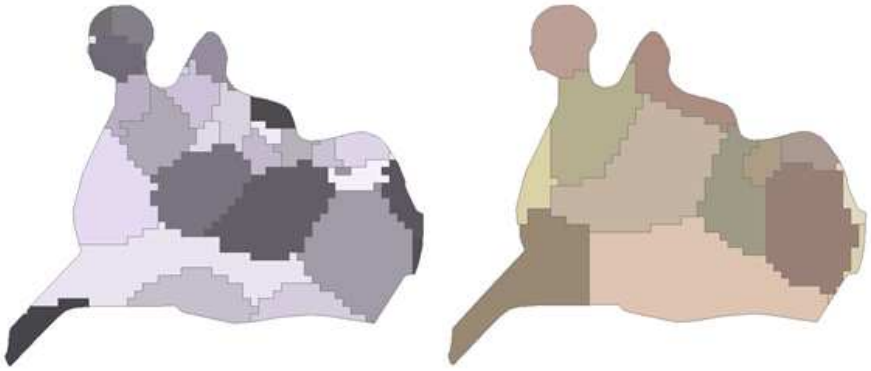


Figure 3.8 BSA tessellation of a LAU by different technologies (2G on the left, 4G on the right).

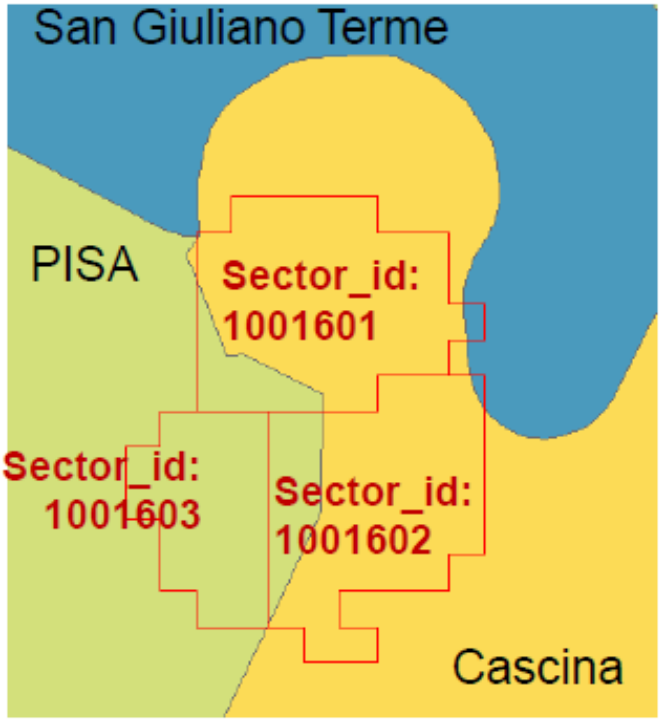


Figure 3.9 Different LAUs coverage for ID_sector.

An example of this operation is reported in table 3.1.

ID_{sector}	Antenna Code	LAU	Overlapping area	BSA_FRACT (%)	Tech Type	Date extraction (MM_AA)
1001601	10016	LAU.1	331400.2411	13.9538	2G	02_17
1001601	10016	LAU.3	34681.97777	1.4603	2G	02_17
1001601	10016	LAU.2	2008905.306	84.5859	2G	02_17
1001602	10016	LAU.1	410446.6632	18.2421	2G	02_17
1001602	10016	LAU.3	13150.95918	0.58449	2G	02_17
1001602	10016	LAU.2	1826401.378	81.1734	2G	02_17
1001603	10016	LAU.1	1187498.644	99.9994	2G	02_17
1001603	10016	LAU.2	7.656187512	0.00064	2G	02_17

Table 3.1 Example of BSA data.

3.2.3. A Bayesian approach using signal strength

3.2.3.1. Introduction

As we can easily see from the core data model in the preceding sections, geographic location is one of the most important variables of the data. However, in many applications, the exact geographic location is either not measured or not stored. Data collected by mobile network antennae are primarily logged for billing customers and for network analysis. For these tasks, only the identification number of the serving antenna is logged rather than the approximated geographic location of the events. There exist advance geographic pinpointing techniques such as triangulation and Timing Advance (Calabrese et al., 2014). However, they are often unavailable since it demands some special infrastructure and data storage and analysis not often found in practice.

In this section we shall avoid the terms antenna and mast as much as possible, since they may cause confusion. Rather, we will use the term *cell*, which refers to both the antenna and the geographic area that is be served by this antenna. Note that this term is also used in cell phones and cellular networks. A *cell site* or shortly a *site* is the location of one of more cells. When we refer to *antenna*, we mean the physical object that receives and transmits signals.

Table 3.2 lists the major site types. The most commonly known site type is the cell tower, which usually contains three cells which have coverage within approximately 120 degrees radius. Cells in other site types are omnidirectional, i.e. the cell operates evenly in all directions.

Type	Description	Number of cells[2]	Range
Cell tower	Tower constructed to support cells	3	500 meters to 40 km
Rooftop site	Cell located on rooftops	3	2 to 40 km
Small cell	Small sized cell	1	500 meters or less
Outdoor DAS[3]	Set of small outdoor cells	1	500 meters or less
Indoor DAS[3]	Set of small indoor cells	1	500 meters or less

Table 3.2 Types of cell sites and their characteristics.

The vast majority of studies on mobile network data use Voronoi tessellation (Okabe et al., 2000) to distribute the geographic location of logged events. The geographic area is divided into Voronoi regions such that each Voronoi region corresponds to the geographic location of a cell and each point in that region is closer to that cell than to any other cell.

There are a couple of downsides to use Voronoi tessellation to estimate the geographic location of devices. First of all, it assumes that all cells are onmidirectional. As described above, most cells are placed in cell towers or on rooftops and are directional. The second downside of Voronoi tessellation is that the coverage range of cells vary across cell types. Table 3.2 shows that the range depends on the cell type and moreover, on the configuration of the cells. Third, cells have overlap, especially in urban areas. This is because of load balancing; if a cell has reached full capacity, neighbouring cells that also have coverage are able to take over communication with mobile devices. This means that a mobile phone is not always connected to the nearest cell with the best signal. In urban areas, a mobile phone switches almost continuously between cells[4].

We present a Bayesian model to estimate densities of mobile phone devices. The likelihood function takes the estimated signal strength of nearly cells into account. Optionally, prior information about where devices are to be expected can be used. This information can be extracted from land use registers, building registers, or OpenStreetMap (OSMF, 2018).

3.2.3.2. Method

In the proposed method, we will use a raster of the geographic area of interest. As an illustration, we use 100×100 raster cells using the Dutch National Grid projection (EPSG, 2018). The main advantage to use raster cells is that different geospatial vector datasets can be combined without the need to calculate spatial intersections, which is a

[2]Usual number of cells per unique location

[3]Distributed Antenna System

[4]There are several smart phone apps that show where the connected cell is located, e.g. Network Cell Info Lite (Wilysis, 2018).

time consuming operation. Besides, the mathematics described below is easier since all raster cells have the constant area size.

The key of the proposed localization method is Bayes' formula, which is used in the following way:

$$\mathbb{P}(i|j) \propto \mathbb{P}(i)\mathbb{P}(j|i) \tag{3.1}$$

where i represents a raster cell and j the polygon of a cell. $\mathbb{P}(i)$ represents prior information about the relative frequency of events at raster cell i. The likelihood term $\mathbb{P}(j|i)$ is the probability that a device is connected to cell j given that it is actually located in raster cell i.

Prior information
The prior function can be used to specify where devices are expected to be. For instance, you would expect more devices on a road than on a grass field next to it. Also, more devices are expected to be inside buildings than outside when normalized per squared kilometer.

Geographic auxiliary information such as land use registers, building registers, and geographic data of roads and railways can be translated into a prior probability of presence of a device per raster cell i.

In the absence of prior information, $\mathbb{P}(i)$ can set to 1. In that case, it is assumed that devices are uniformly distributed across the geographic areas in which they are logged.

Likelihood function
The main advantage of using this Bayesian model compared to the Voronoi tessellation is that it takes the overlap of cells into account. This information is contained in the likelihood, which is defined as

$$\mathbb{P}(j|i) = \begin{cases} 0 & \text{if raster cell } i \text{ is not in cell} j, \\ \frac{s(i,j)}{\sum_k s(i,k)} & \text{if raster cell } i \text{ is in cell} j, \end{cases} \tag{3.2}$$

where $s(i,j)$ represents the signal strength that device i receives from cell j. Before defining it, let us define $S(i,j)$ which is an approximation of the actual signal strength denoted in dBm, which stands for decibels relative to one milliwatt. For onmidirectional cells, it is defined as

$$S(i,j) = S_r(r_{ij}) \tag{3.3}$$

where r_{ij} is the distance between the middle point of raster cell i and the antenna of cell j in meters. The function $S_r(r)$ returns the signal strength as a function of distance r:

$$S_r(r) = S_0 - 10\log_{10}(r^2/r_0^2) = S_0 - 20\log_{10}(r) \tag{3.4}$$

where S_0 is the signal strength at $r_0 = 1$ meter distance from the antenna.

A directional cell has an antenna which is directed at a specific angle. Along this angle, the signal strength is received at its best. However, the signal can also be good in other directions. It is comparable to a speaker which produces sound in a specific direction. The sound will be audible in many directions, but at the sides and the back of the speaker, the sound will much weaker. The directional beam of antenna j can be specified with four parameters:

- The horizontal/azimuth angle α_j is the angle from the top view between the north direction and the direction in which the antenna is pointed. Therefore, in reality this angle can be anywhere between 0 and 360 degrees. Note that cell towers and rooftop cells often contain three antennas with 120 degrees in between.
- The vertical/elevation angle β_j is the angle between the horizon plane and the tilt of the antenna. Note that this angle is often very small, typically only four degrees. The plane that is tilt along this angle is called the elevation plane.
- The horizontal beam width γ_j specifies in which angular difference from the azimuth angle in the elevation plane the signal loss is $3dB$ or less. In other words, the angles in the elevation plane for which the signal loss is $3dB$ correspond to $\alpha_j \pm \gamma_j/2$. In reality, these angles are around 65 degrees.
- The vertical beam width θ_j specifies the angular difference from β_j in the vertical plane orthogonal to α_j in which the signal loss is $3dB$. The angles in which the signal loss is $3dB$ loss correspond to $\beta_j \pm \theta_j/2$. In reality, these angles are around 9 degrees.

Let ϵ_{ij} be the angle from the side view between the line along the elevation angle β_j and the line between the center of antenna j and the center of grid cell i. Let δ_{ij} be the angle in the elevation plane between the azimuth angle α_j and the line between the center of antenna j and the center of grid cell i orthogonally projected in the elevation plane. Then, the signal strength for directional cells is defined by

$$S(i,j) = S_r(r_{ij}) - S_{e\ell}(\epsilon_{ij}, \theta_j) - S_{az}(\delta_{ij}, \gamma_j) \tag{3.5}$$

where $S_{e\ell}$ and S_{az} specify the signal loss based on the angular difference with the elevation and azimuth, respectively.

Each antenna type has its own radiation pattern for both the azimuth and elevation angles. These patterns define the relation between signal loss and the offset angles, i.e., δ_{ij} for the azimuth and ϵ_{ij} for the elevation angles. We used a Gaussian distribution to model the radiation pattern. The result is shown in Figure 3.10. The black line shows the relation between signal loss and angle in the azimuth plane (left) and elevation plane (right). The grey circles correspond to the signal loss; the outer circle means $0dB$ loss (which is only achieved in the main direction), the next circle corresponds to $5dB$ loss, etcetera. The red lines correspond to the angles corresponding to $3dB$ loss. So the difference between the red lines is γ_j in the Azimuth plane and θ_j in the Elevation plane.

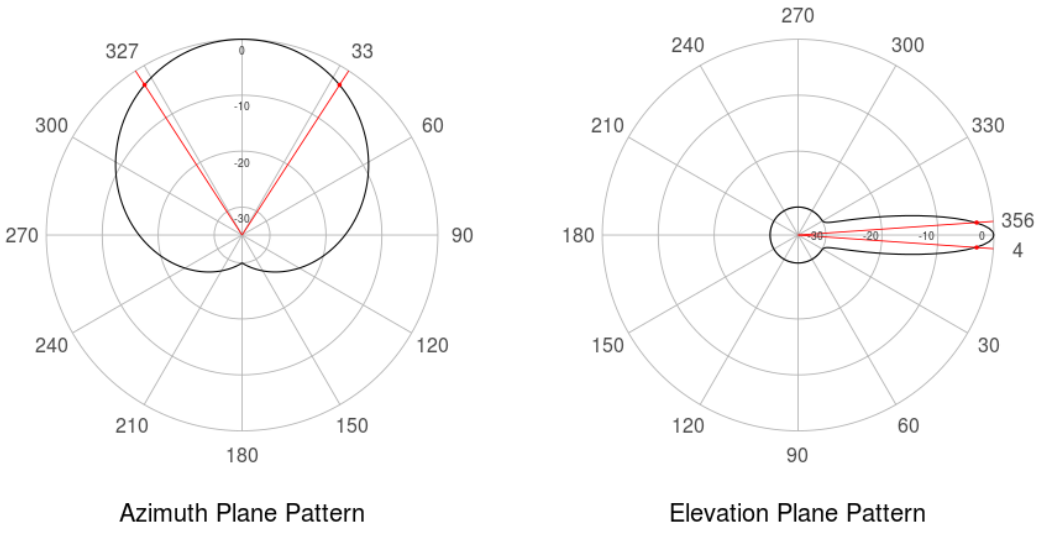


Figure 3.10 Radiation patterns for the azimuth and elevation planes

Although these models approximate the general curve of real radiation patterns, the radiation pattern are more complex in reality, e.g. they often contain local spikes caused by so-called side and back lobes.

Figure 3.11 illustrates the signal strength at the ground level from above for a specific cell. In this case, the cell is placed at $x = 0$, $y = 0$ at 55 meters above ground level. The cell is directed eastwards with an elevation angle (tilt) of 5 degrees, a horizontal beam width of 65 degrees and a vertical beam width of 9 degrees. Table 3.3 describes how the signal strength values are interpreted by the network. Notice that the signal strength close to the cell, which means almost under the cell, is lower than at a couple of hundred

meters distance. This is caused by a relatively large ϵ angles at raster cells nearby the cell.

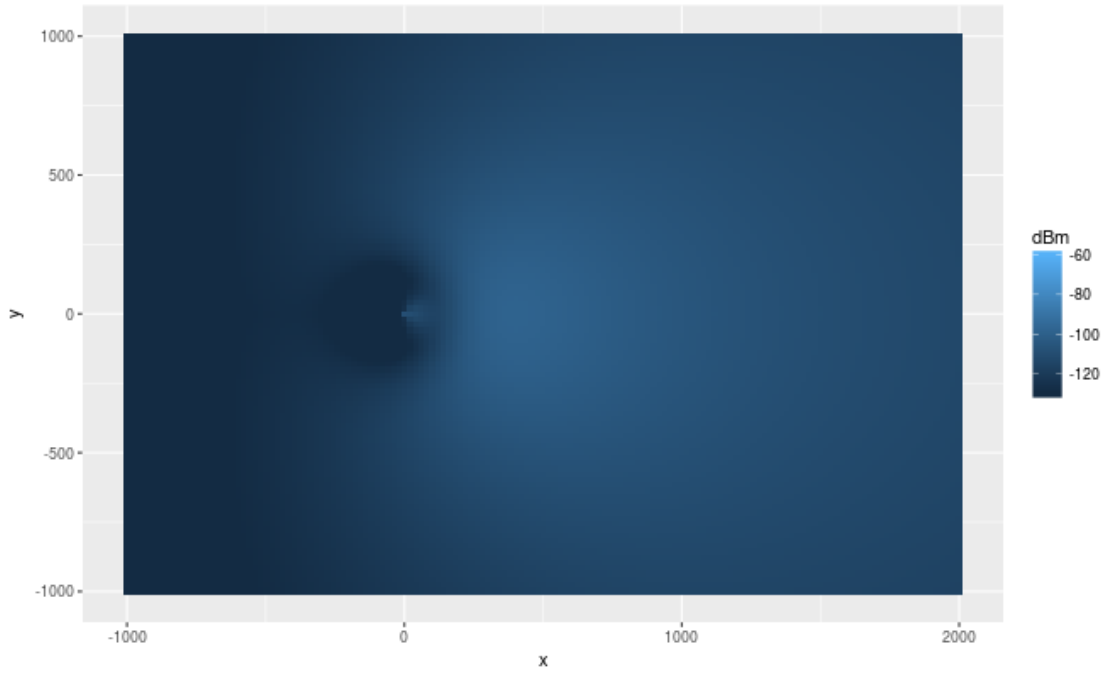


Figure 3.11 Signal strength at ground level

Signal strength (dBm)	Indication
-70 or higher	excellent
-90 to -70	good
-100 to -90	fair
-110 to -100	poor
-110 or less	bad or no signal

Table 3.3 Signal strength indication.

It is often unclear how to load balancing mechanism works in practice. In the connection process between a device and the cell network, it can be assumed that when there are a couple of cells with a signal strength above a certain threshold, say $-100dBm$, the cell is selected that has the highest capacity available. Therefore it is less important if

the signal strength is -70 or -90 than -90 or -110. To model this load balancing mechanism, we have used a logistic function that translates the signal strength $S(i,j)$ to a relative signal strength measure $s(i,j)$ which we used to define the likelihood function (3.2).

$$s(i,j) = \frac{1}{1+e^{-T(i,j)}} \tag{3.6}$$

where

$$T(i,j) = \frac{S(i,j) - S^{mid}}{S^{width}} \tag{3.7}$$

where S^{mid} and S^{width} are parameters that define the mid point and width of the curve respectively. Figure 3.12 shows the relation between the signal strength S_{ij} on the x-axis and the relative signal strength s_{ij} on the y-axis.

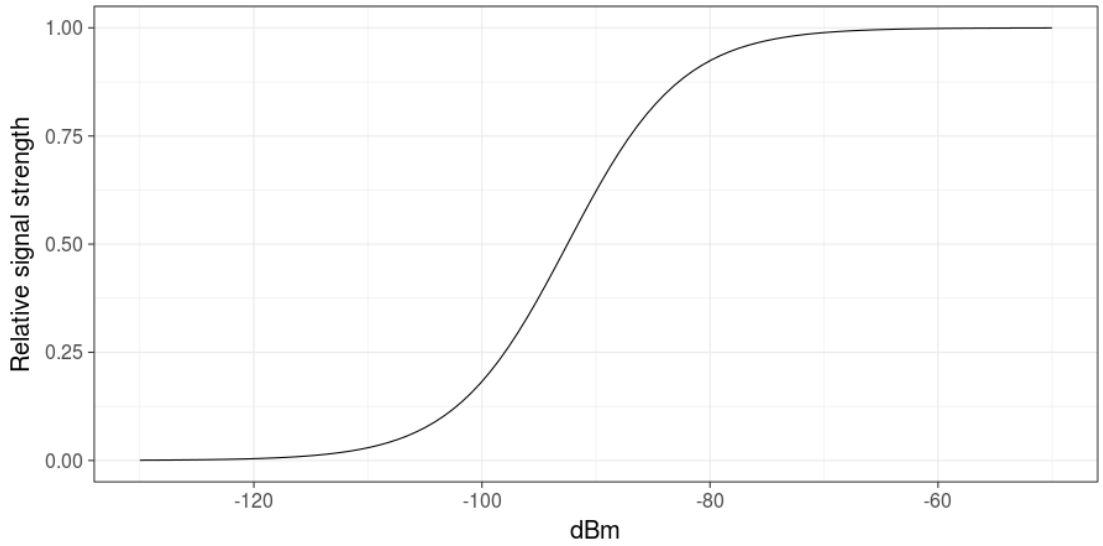


Figure 3.12 Signal strength at ground level

The relative signal strength at the ground level is shown in Figure 3.13. The probability values that are shown are normalized such that they sum up to one. Compared to the absolute signal strength shown in Figure 3.11, this distribution puts more emphasis on the geographic area that is in the spotlight of the cell.

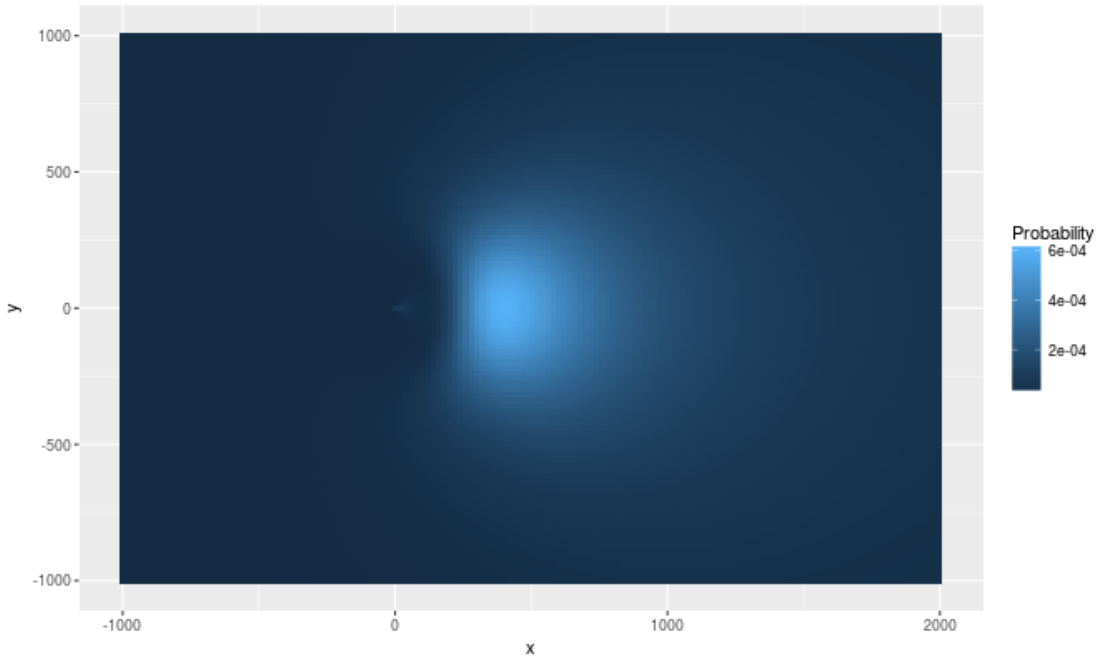


Figure 3.13 Relative signal strength at ground level

3.3. The core data model

This advanced methodology is quite precisely described in the Positium report (Positium, 2017). To implement such a detailed model one should have a very rich access to the individual data in possession of the MNOs. As the goal is to describe a continuous presence of each subscriber it is preferable to detect very regularly the mobile devices. This is why access only to CDRs may not be sufficient to build the entire model. Nonetheless the basic identification of the different most frequented areas is likely to be accessible even with CDRs.

There are mostly two different phases in the core data model. The first one consists of identifying the most frequented places at different scales (section 3.3.1): the country of residence, the place of residence and other anchor points and the usual environment. The second phase is a quite refined model to build a continuous description of movements and locations of all individuals (3.3.2).

3.3.1. Identification of most frequented locations

Identification of the Country of Residence
The Country of Residence (COR) table is part of the core data model. All records should have a COR value. All subscribers must have the COR for all periods of the data. The COR is a dynamic attribute that can change over time.

We can assume that majority of the subscribers identified in domestic and outbound data are residents of the country of reference, and inbound roaming subscribers are residents of appropriate foreign countries. However, in current mobile, transnational society, there are many people who travel and live in different countries, and use SIM cards of different countries. Therefore, it is necessary to identify the COR of the subscribers by looking at their presence patterns. One way to identify the COR relies on the tourism definition of the country of residence:

For a vast majority of persons, there is no problem to determine the country of usual residence. For the small group of persons for whom the place of usual residence is not clear, the recommended interpretation is to determine the place of usual residence according to the length of time spent at the different locations: the place where a person spends the majority of the year shall be taken as his/her place of usual residence (Eurostat, 2014).

This definition is difficult to use straightforwardly with mobile data because there is no time period specified and no suggestion concerning for what period of time this assumption is made. However, an implementation can be to consider as country of residence the country where the person spends majority of time during the consecutive 12 months. The identified COR value is assigned to each month, but not less than for a period of 6 months. If no specific country can be identified, then the country where the mobile phone is from is considered as country of residence.

In order to identify the COR combining data from domestic and outbound roaming visitation days per country is necessary (joined by subscriber_id, which has to be the same in domestic and outbound data). Inbound roaming data is used for the COR identification as it is.

If for some reason the COR was not calculated for specific period of data of the subscriber, these records should be deleted from the merged table (or alternatively not used in following processes).

Identification of Anchor Points and Place of Residence
Meaningful locations (also anchor points) are identifiable locations that the subscriber visits regularly, such as home, work, school, kindergarten, favourite shops, summer-

house, parent's place, sports, free time, relatives home, favourite restaurants, friends places, etc. It is difficult to semantically identify all of those places individually, but using spatio-temporal behaviour analyses and clustering, at least the following groups of locations can be found based on the clusters of antennae:

- Sleeping anchor;
- Work-time anchor;
- Other regularly visited locations.

Figure 3.14 Example of some location's temporal pattern of visitation (based on location events per hour in one month).

In figure 3.14 we can see different patterns to identify these locations. Anchor points should be calculated per month, but longer periods should be used to assign them, similarly to the COR identification. If different places of residence are identified, these should be assigned to at least 6 months of duration (again, similarly to the COR). Several of such locations could be found, but in order to apply realistic semantics, following rules have to be applied:

- Only one home (Place of Residence – POR) per subscriber per period, the most "popular" anchor for sleeping (most days spent over a period), should be assigned. If several sleeping anchors are identified, then only one location is POR, others are assumed second homes (summerhouses, relative's places, etc.) – Second Home Anchor Point (SHAP).
- There can be several Work-Time Anchors Points (WTAP), one should be marked as the main work-time anchor. WTAPs can also be in the same location as POR for persons who work and live at the same place.
- Other Regular Anchor Points (RAP) are identified, but no semantic meaning can be assigned to them (unless more in-depth analysis is conducted).

For identification of these locations, individual cells or clusters of antennae with similar temporal patterns should be found during the day, the month (see figure 3.15).

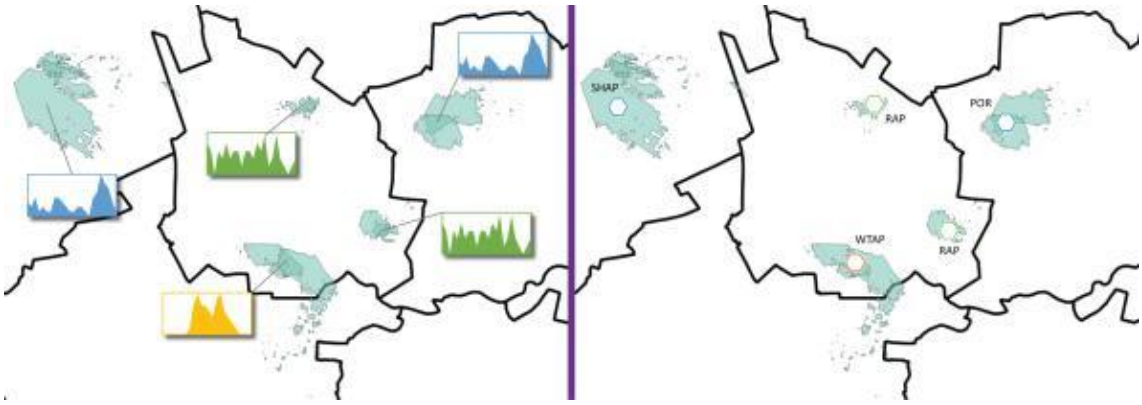


Figure 3.15 Example of identification and classification of anchor points located in three administrative units.

The identified and classified anchor points should now be saved as separate tables (figure 3.3) for further use and are considered a part of the core data model.

Identification of Usual Environment
The Usual Environment (UE) can be defined as

the geographical area, though not necessarily a contiguous one, within which an individual conducts his regular life routines and shall be determined on the basis of the following criteria: the crossing of administrative borders or the distance from the place of usual residence, the duration of the visit, the frequency of the visit, the purpose of the visit (Eurostat, 2014).

For simplicity reasons, we define here UA (usual anchors) as LAU3 units that subscriber visits regularly but countries may have different definitions For example in figure 3.16, the three LAU3 units, where five anchor points are located, should be assigned as UA for the specific subscriber. For subscribers with COR other than the country of reference, UA covers by default also the COR. In addition, if there are anchor points within the country of reference, then these are included in UA – therefore UA can extend for transnational travellers over several countries.

Similarly to anchor points, usual environment (UE) should also be defined based on months. If another country is a part of a UA based on frequency of the visits and/or duration of the stay, this country should be added to the UE table. By default, all inbound

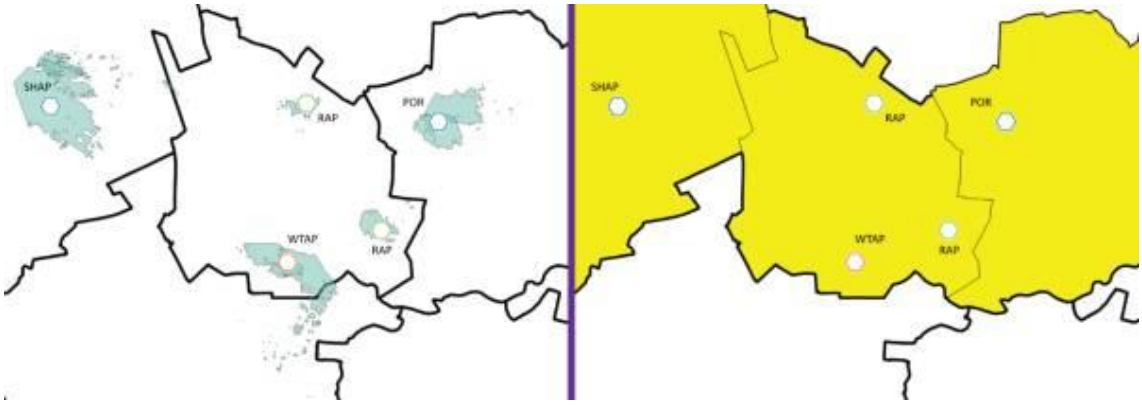


Figure 3.16 Defining usual environment from the anchor points.

roaming data subscribers with COR not in the country of reference (they do not spend more than 183 days in the country) should be added to the UA list.

3.3.2. Continuous description of movements and locations

Now we describe the second phase aiming at a continuous description of movements and locations.

Identification of Stay Sections
Peoples' spatio-temporal behaviours can be reduced to a consecutive sets of two elements: stay sections and movement sections. A person

- ⇝ stays at home, sleeps, wakes up (stay);
- ⇝ goes to work in the morning (movement);
- ⇝ arrives at work (stay);
- ⇝ goes to lunch (movement);
- ⇝ eats at the restaurant (stay);
- ⇝ goes back to work (movement);
- ⇝ finishes workday (stay);
- ⇝ goes to shop (movement);

⇝ does shopping (stay);

⇝ goes home (movement);

⇝ stays at home (stay);

⇝ ...

Identification of stay sections is the initial challenge, as majority of location events indicate presence in some location.

The objective is to identify the stay section locations and time periods when the person was present in the location. As the stay sections geographical representation used here is LAU3, stays in different locations within the same LAU3 are combined. This means that if all activities of the subscriber are within one LAU3, there will only be one long stay section. Each stay section is assigned a special unique stay_id to be able to identify the group of location events within one stay section (see table 3.4 for an example). For outbound data, the stay sections are based on the presence in individual countries. The difficulty comes from the fact that when a single event is recorded, it cannot be decided the duration of the stay section nor how it was generated. This issue will be dealt with later.

subscriber_id	event_time	lau3_code	destination_iso_a2	stay_id
A	2016-02-14 09:22:39	100	FR	1
A	2016-02-14 09:46:12	100	FR	1
A	2016-02-14 10:02:32	100	FR	1
A	2016-02-14 11:04:58	101	FR	2
A	2016-02-14 12:10:17	100	FR	3
A	2016-02-14 14:00:13	100	FR	3
A	2016-03-14 14:17:12		ES	4
A	2016-03-14 20:34:25		ES	4
A	2016-03-15 08:21:37		ES	4
A	2016-03-15 13:32:18		UK	5
A	2016-03-15 19:43:54		UK	5
A	2016-03-15 23:52:13	100	FR	6

Table 3.4 Example of identification of stay sections for one subscriber (domestic and outbound roaming data example).

Linking Anchor Points with Stay Sections

After stay sections have been identified, each of the sections should be linked to any

existing anchor points previously calculated to identify the characteristics of each stay (see figure 3.17).

Identification of Trips
A trip is the collection of movement and stay sections that starts from the POR stay section and ends in a POR stay section – a trip is a journey from home to home. A trip does not include stay sections in the POR. At this point, no movement sections have yet been identified.

For domestic data, including the outbound roaming, identification of trips is based on the POR. For inbound roaming data, the POR is assumed to be their COR, and identification of trips is only possible from the moment they enter the country (i.e. first location event from in the network). If during the trip, the subscriber's POR changes, then the trip starts in one POR stay section and ends in another. Figure 3.18 includes an illustrative example.

Identification of Transit Points
In the core data model, the stay sections are divided into:

- stay sections – places where the person actually stopped for some activity or function for a specific duration (e.g. being at home is a stay section, being in a workplace is a stay section, having lunch away from the workplace for 30 minutes is a stay section, going to a shop is a stay section, staying in a hotel is a stay section);
- transit points – places where people stop for a very short time and have no functional purpose or do not stop at all but they have a location event there (e.g. making a phone call while driving). Transit points, as considered as good guidance for identification of the route that subscribers took to get to the destination.

Because of the quantitative data, and also depending on the nature of the mobile phone data (density of location events, accuracy, etc.), there are challenges to correctly identify the stay sections and the actual length of stays, and to make a distinction between actual stays and transits.

Transit points should be identified using the following rules:

1. If the stay section is located in the same LAU3 where any of the anchor points are located, then it is not a transit point;
2. If the duration of the stay section is longer than 15 minutes, then it is automatically not a transit point;

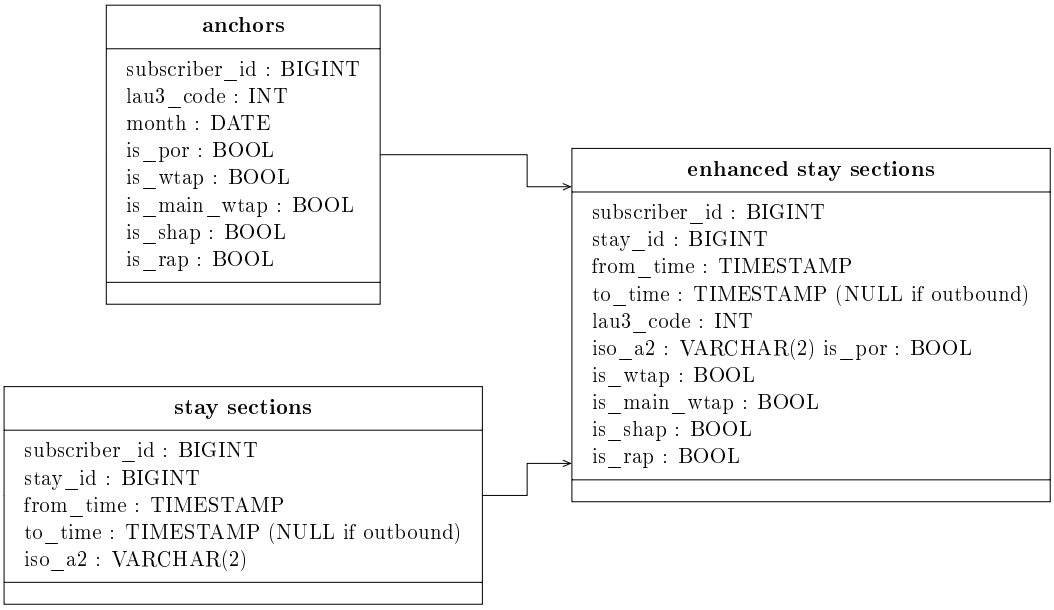


Figure 3.17 UML class diagram describing additional attributes to describe if stay sections are within the same LAU3 as are different anchor points.

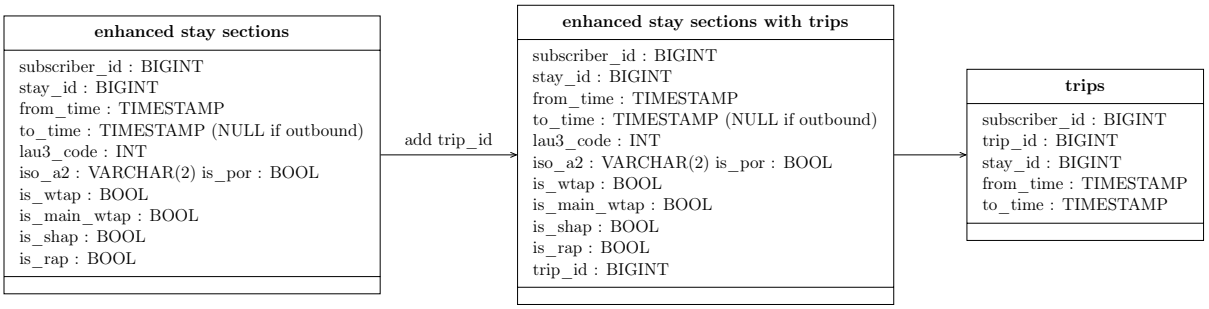


Figure 3.18 UML class diagram of amended stay sections table with trip_id attribute and a new trips table.

3. If the stay section is a part of trip which is longer than 4 hours, and this is the only stay section during the trip, it is not a transit point;

4. If the stay section is a part of trip which is longer than 4 hours, and this is the furthest stay section from the POR (based on distance) during the trip, it is not a transit point;

5. All other stay sections that do not meet the above criteria, are transit points.

For stays in the foreign countries, the following rules should apply:

1. If the duration of the stay section is longer than 4 hours, then it is automatically not a transit point;
2. If there are no other foreign stay sections in this trip, then the furthest stay section during the trip is not a transit point.

The logical path presented in figure 3.19 can be used to distinguish stay sections from the transit points. The 15min/4h criteria are a subject to change based on the country -– e.g. some LAU3 units might be so large, that passing through them under 15 minutes might not be possible.

As a result of this process, "correct" stay sections and transit points are disentangled, which will help to identify the movement sections between stay sections and using transit points as a base.

Calculating Movement Sections

Stay sections and transit points have been identified. The next step is to calculate the intermediate movement sections. This requires trajectory calculations and to first identify the transportation mode. There are five types of transportation modes that could potentially be identified (sea, rail, road, air, on foot). In each country, depending on different data quality from different MNOs, methods, rules, criteria and specific algorithms can be diverse.

Transportation mode identification is closely related to trajectory identification, so that both of these will be conducted together. Trajectory calculations of the movement sections allow identification of pass-through traffic in intermediate LAU3 and foreign countries between stay sections that are far apart. Trajectories can be identified through graph connections between neighbouring LAU3 units (which LAU3 units have to be passed through in order to get from stay section in initial LAU3 to final LAU3) or based on road and other transportation network. The latter requires more complex and resourceful process, with shortest path finding algorithms based on the country's road network. This method is not described here (see Positium (2017)).

We shall then focus on the calculation of trajectories based on graphs between administrative units. First, a generalisation of road networks to LAU3 has to be obtained initially. This can be based on a neighbouring dataset and/or routing algorithms aggregated on an average distance time between neighbouring LAU3 units. This is basically

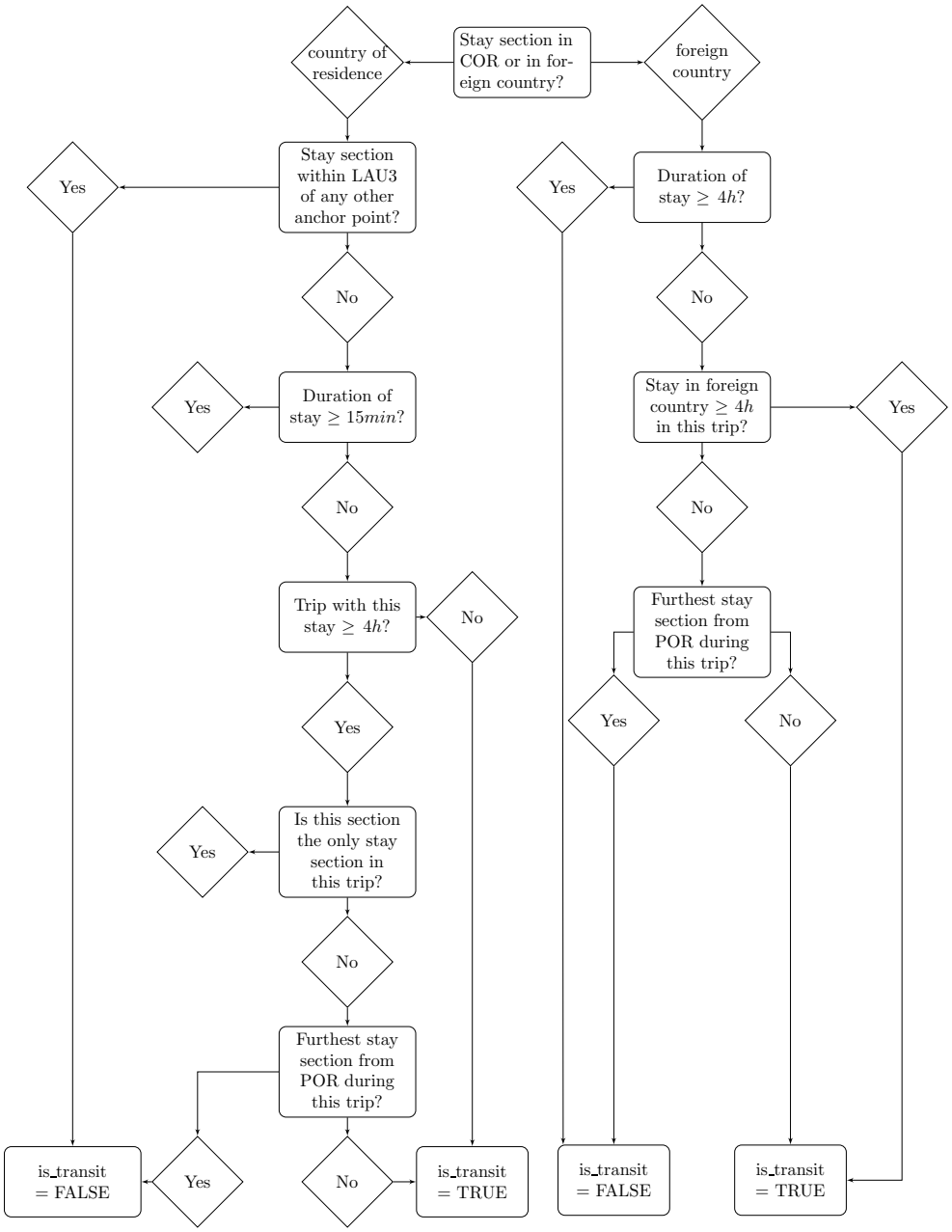


Figure 3.19 Logical steps for identifying if the stay section is transit point or not. In this logical step, being present in a foreign country under 4 hours, is considered a transit visit. Being present in a LAU3 unit under 15 minutes is also considered as a transit.

a graph with LAU3 units as nodes and edges between them representing the average travel time required to travel between them.

A number of shortest path resolving algorithms (e.g. Dijkstra's algorithm (Nemhauser and Wolsey, 1999)) is available for calculating the intermediate transit LAU3 units and travel times required to get from one one stay section in the first LAU3 to the next stay section in another LAU3. If there are transit points between two stay sections (is_transit=TRUE), then the movement between the stay section should be united. An example of several stay sections and movement paths between them can be easily offered (see figure 3.20):

- From stay section in lau_code 100 to next stay section in lau_code 101 – there are no intermediate LAU3-s (100 and 101 are assumed neighbouring units with direct road between them). Travel time is 12 minutes.
- From stay section in 101 to stay section in 108 – the travel should be following: 101 ⇒ 107 ⇒ 108; travel time 11 + 17 = 28 min.
- From stay section in 108 to stay section in 105 with intermediate transit point in 104, the shortest path is following: 108 ⇒ 107 ⇒ 104 ⇒ 105, travel time 17 + 8 = 25 (from 108 to 104 via 107) + 9 (from 104 to 105) = 25 + 9 = 34 min. Although the shortest path from 108 to 105 would be via 109 (19 + 11 = 30), the transit point in 104 suggests that longer path was used.

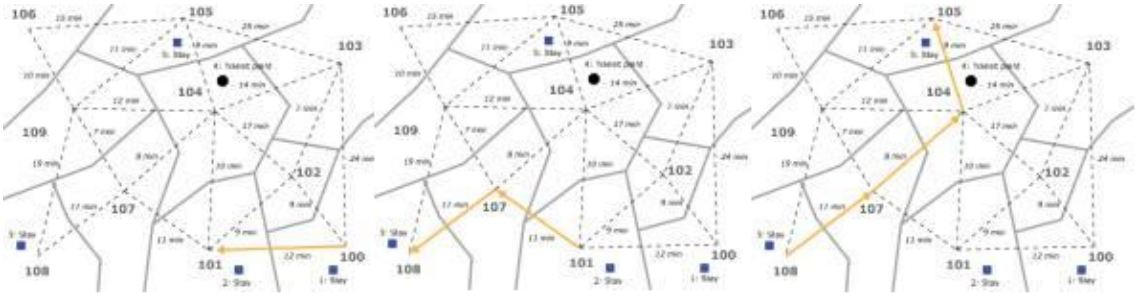


Figure 3.20 Illustration of example of generated 3 movement sections.

A class diagram with the underlying data structure of these movements is given in figure 3.21.

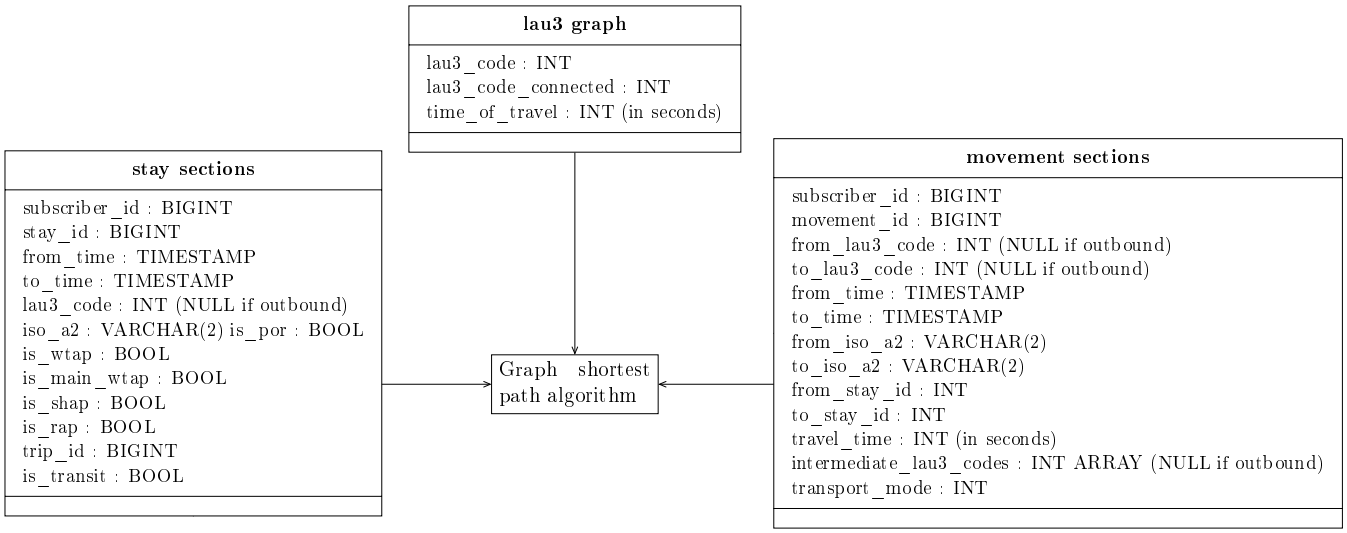


Figure 3.21 UML of the resulting movement sections table.

Adjusting the stay sections' duration based on movement sections
Stay section durations have now to be adjusted to account for the results for movement sections durations. The following logic should be applied:

1. Calculate time periods between stays;
2. Compare the time periods between stay sections to durations of movement sections;
3. If time periods between two stay sections and traveling time differ significantly (e.g. time period between stays 3 hours, travel time 1 hour), then extend the preceding and subsequent stay sections to_time and from_time, respectively (see figure 3.22). If for outbound data calculating traveling times is an issue, simply extend to_time and from_time by half of the movement period time, making two consecutive stays without an intermediate movement section.

Finalising the Core Data Model
To finalise the model, the necessary tables must be brought in order in terms of the system and made accessible for following domain-specific procedures.

The core data model tables include:

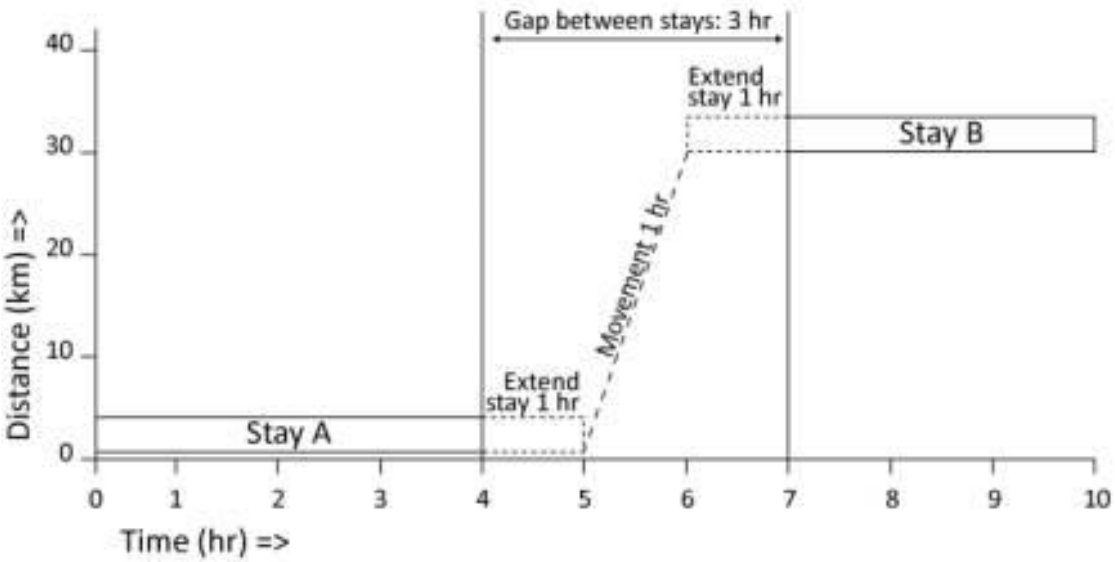


Figure 3.22 Example of extending the consecutive stay sections. Stay A original duration 4hr,extended to 5 (from 0 to 5); stay B original duration 3 hr, extended to 4 (from 6 to 10). Movement between those locations should take 1 hour.

- Country of Residence table;
- Anchor points table;
- Usual environment table;
- Trips (from and to home table);
- Stay sections table;
- Movement sections table;

In addition, at least one reference table is also needed in any processing and querying from the core data model – administrative subdivisions data of the country. Based on that data it is possible to aggregate any data to upper LAU2 and LAU1 units.

There are a number of processes related to specific countries where the processing is conducted based on the specific nature of mobile data, human mobility and geography (eliminating accidental coastal or border roaming, identification of foreigners with local SIM cards, "take home"action).

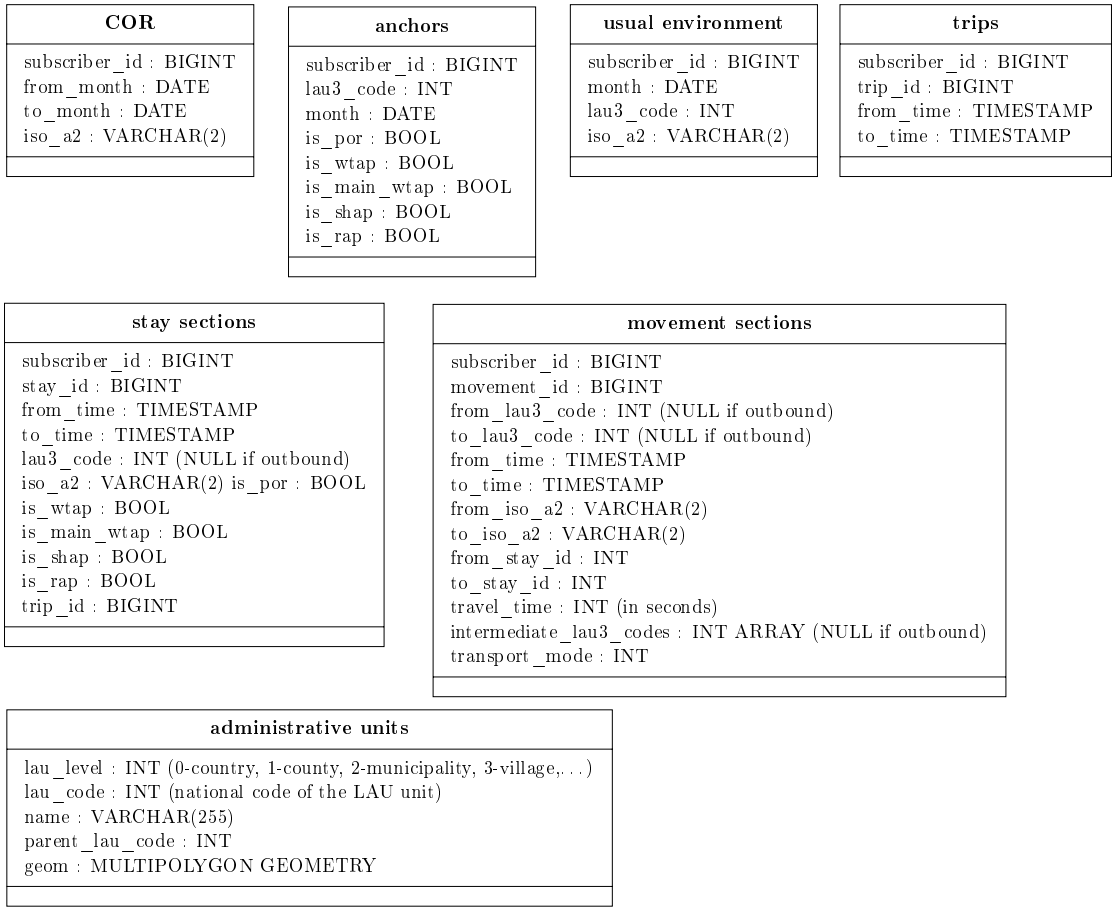


Figure 3.23 UML class diagram of the finalised core data model representing all basic tables for mobility and semantics of the subscribers.

By current stage, the core data model is ready and only processes for specific domains are required in order to generate statistical indicators. Direct queries, aggregation, data mining, machine learning and other statistical processes can be made directly using the core data model as a source.

However, before turning to topics about generating statistical indicators, some words on data revisions should be said. When data for new periods from MNOs is obtained, core data model needs to be updated in a way that all existing microdata time-series elements (stay and movement sections, anchor points, UE, etc.) are extended and possibly

recalculated including “old” and “new” data.

The processes should not simply be calculated for the new data, but the new data should be combined with existing core data model tables and variables (COR, POR, anchor points, etc.). This introduces the issue with data revisions – should the older statistics be changed upon receiving the new data and new information that might suggest even earlier changes in indicators. For example based on the data updates, it might become clear that a subscriber has been the resident of a different country than in current data model (monthly update reveals that inbound roaming subscriber has been present in the country more than 183 days in past 12 months, but in the previous month, the amount of days was less than 183). Will the period of a new COR begin with the data updated time, or should it have effect on historical data (extended to all previous 12 months)? There is no rule here, but one of two options has to be chosen:

1. Historical indicators will not be changed;

2. Historical data will be changed up to some extent; the indicators provided during this extent should be disclaimed as “preliminary” and are a subject for change. Indicators older than the extent (e.g. 6 months), can be marked as final.

3.4. Aggregating the results from the core data model

This section is devoted to examples of aggregations for two different domains, tourism and mobility, derived from the same core model data. The main advantage of the core data model is to allow different aggregation according to the type of indicators needed. As every subscriber has a presence fully described and imputed between the cell phone connections, descriptions of the present population can be obtained by aggregating at every hour the data.

To produce indicators on tourism or mobility additional steps need to be taken.

3.4.1. Aggregation for tourism indicators

For tourism indicators, tourism trips need to be identified, either inbound, outbound or domestic.

A tourist trip in a country of reference is a trip including at least one stay section that is longer than 15 minutes outside the Usual Environment starting and ending at the subscriber’s place of residence. For outbound tourism a tourism trip is a trip including at least one stay section in a foreign country.

Using (i) the COR table to identify the residency of the subscriber, (ii) the UA table combined with the stay sections table to identify stay sections outside the UA, (iii) the trip table, a table of tourism trips classified as domestic, outbound or inbound; and (iv) a table of tourism visits can be produced. Although basically a stay section and a visit can be equalized, in tourism the "visit" will be used to represent stop in a specific location (based on the minimum geographical units, i.e. an LAU3 in the country of reference and or a country code in foreign countries).

The calculation logic for tourism trips and visits is the following:

- A domestic trip is a trip for subscribers with COR outside the UA, when there are stay sections inside the country of reference (stay_sections.destination_iso_a2=COR.iso_a2);
- An outbound trip is a trip for subscribers with COR, outside UA, when there are stay section outside the country of reference (stay_sections.destination_iso_a2!=COR.iso_a2);
- An inbound trip is a trip for subscribers with COR, outside their UA, when there are stay section inside the country of reference (stay_sections.destination_iso_a2!=COR.iso_a2);
- A tourism trip can be a domestic and outbound (is_domestic = TRUE, is_outbound=TRUE) at the same time if there are visits in the country of reference and in foreign countries during the same trip;
- Inbound trip cannot be also domestic and outbound trip at the same time (that defies the logic and is impossible if the core data model was created correctly);
- For each trip, COR and POR data should also be inserted into the tourism trips table;
- All movement sections that include LAU3 that are within the tourism trip and outside the usual environment should be added to the tourism visits table, but with attribute is_transit = TRUE.

3.4.2. Aggregation for mobility indicators

Mobility represents movement between locations. Based on the core data model two different types of indicator can be generated:

- origin-destination (O-D) matrices;

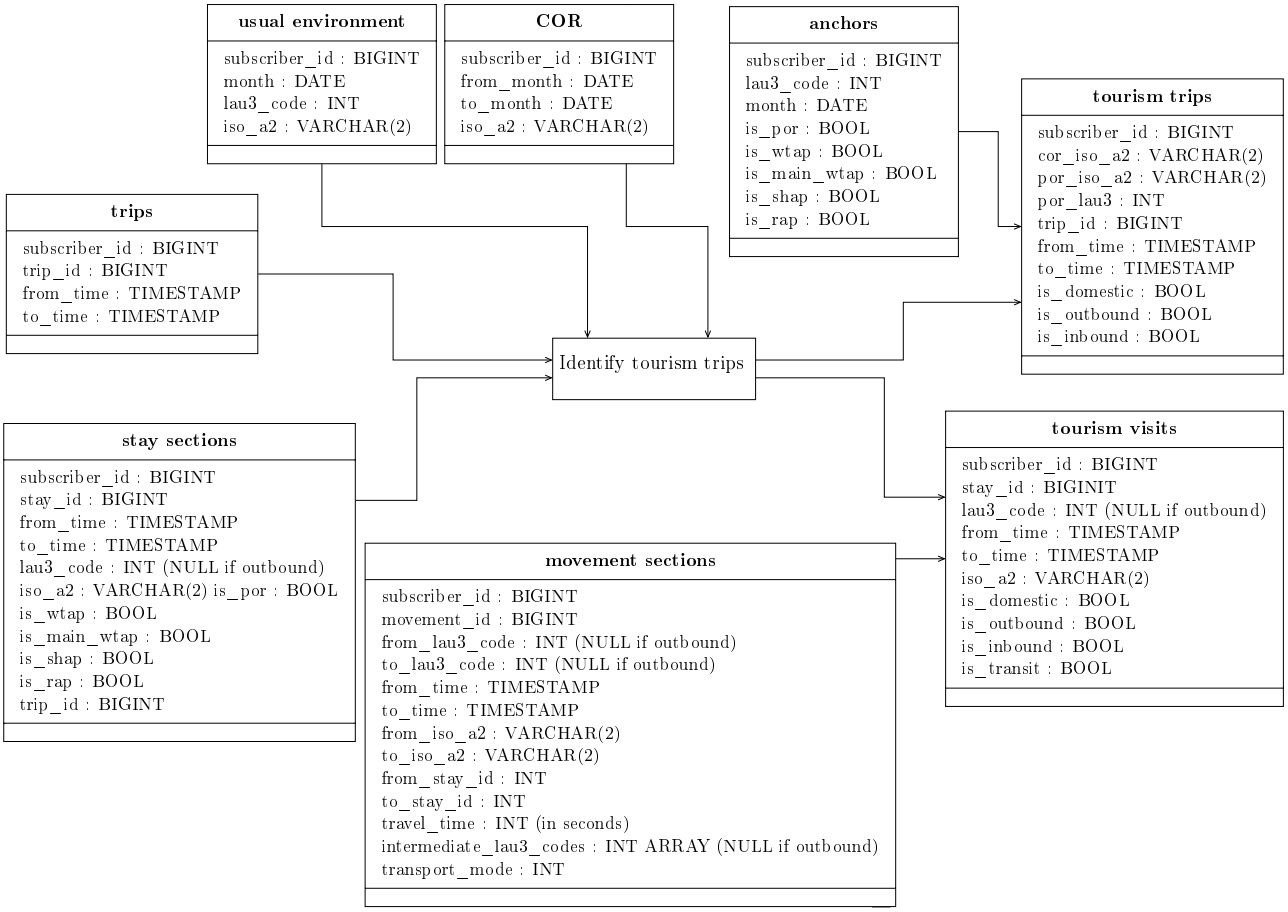


Figure 3.24 UML class diagram representing newly created tourism trips and visits (stay sections) tables.

- everyday commuting indicators.

Both can be used for similar, and also for different purposes. O-D matrices are mostly used in transportation planning as a modelled data based in the traditional four-step transportation forecasting model, steps being 1) trip generation, 2) trip distribution, 3) mode choice, and 4) route assignment. O-D matrices from mobile phone data can cover element in all four of those steps. Everyday commuting is somewhat simpler indicator set describing the patterns of everyday regular commuting between home and work, often also described as work-related travel and used for much wider purposes.

Trips can be tagged as regular or irregular. Besides as the work place has been identified work related trips and daily commuting can also be flagged. From there commuting matrices for instance can be produced through aggregation.

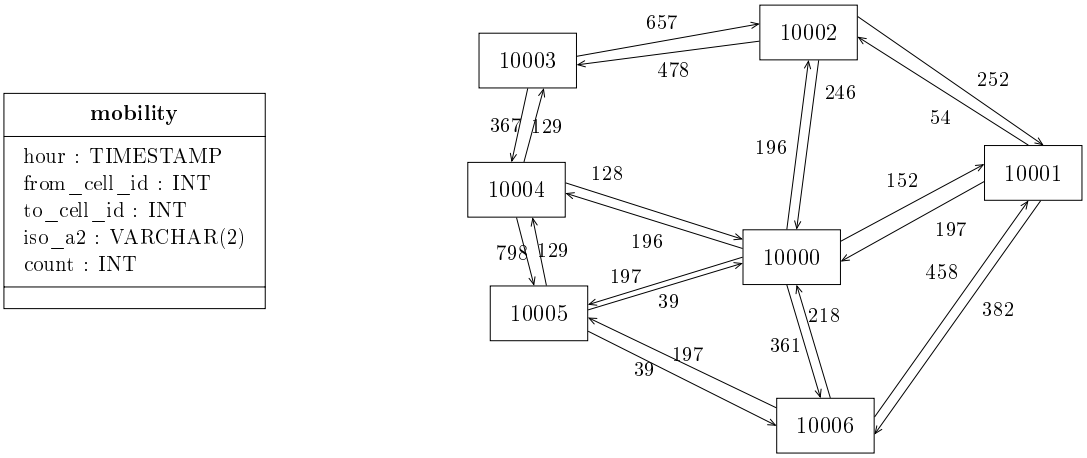


Figure 3.25 Example of aggregated data table provided by MNOs for mobility and graph representation of such data (number of consecutive location events in antennae within a time period).

3.4.3. Aggregation from a less extended access to microdata (CDRs only)

With CDRs only it is possible to define the country of residence, the different anchor points and the usual environment. Yet a continuous model seems difficult to build as the localisations are too sparse with such data. So the aggregation should be done differently.

3.4.3.1. For residential population[5]

Vanhoof et al. (2018) describe different methods for localizing home for every user:

1. The home location is inferred as the location where the highest amounts of calls were made.

2. The home location is inferred as the location that had the maximal number of distinct days with activities.

[5]Prepared in collaboration with Francesco Altarocca and Raffaello Martinelli.

3. The home location is inferred as the location with most activities during hours x and y.

4. The home location is inferred as the location with most activities while aggregating all activities within a range of x meters to this location.

5. The home location is inferred as the location with most activities during hours x and y, while aggregating all activities within a range of xx meter to this location.

Even though it is not easy to assess and relatively sensitive to the heuristics chosen to define the home location, it is quite feasible to estimate where every user is presumably living during a month. The best heuristic, more coherent with official statistics, seems to be the second one: maximal number of distinct days.

From there it is possible to produce aggregates of residential population. By comparing these aggregates from month to month some insights on seasonal variations of the population density are very reachable.

If accessed data included sociodemographic information of subscribers (e.g. the billing address) richer options would be at hand. However, this sort of agreement is not easily attainable.

Complementarily, using the BSA approach above (see section 3.2.2), ISTAT has implemented the following algorithm to compute present population at time t. Let us define:

- m: LAU.
- ts: time slot.
- s: SIM, s= 1, . . . , S.
- c: call.
- CC: number of calls for a SIM in the time slot ts.
- BSAFract(cell): list of BSA fraction of a cell on the LAU as defined in the BSA approach (see section 3.2.2). Notice that for each cell Sum(fract(m, cell))=1 for all LAUs m in BSFract(cell).
- p(s,ts,m): presence score of the SIM s in the LAU m in the time slot ts. Notice that for each SIM s having at least one call in the time slot ts Sum(p(s,ts,m))=1 over all LAUs m.

- pp(ts,m): present population in LAU m in time slot ts. Notice that for each time slot ts Sum(pp(ts,m))$\leq$ S for all LAUs m (it is exactly the number of SIMs having at least one call in the time slot ts).

In pseudocode the algorithm is:

Algorithm 1 Algorithm to calculate pp(ts,m)

```
for SIM s, time slot ts do                    ▷ presence score of SIM s in LAU m
    LC(s,ts) = list of calls for s in ts
    CC(s) = count of LC(s,ts)
    for call c in LC(s,ts) do
        c.cell is the cell of c
        BSAList(c)= BSAFract(c.cell) is the list of BSA fraction of c.cell
        for fract(m,c) in BSList(c) do
            p(s,ts,m) += fract(m,c)/CC(s)
        end for
    end for
end for
pp(ts,m) = Sum(p(s,ts,m)) for all s ▷ count of presence scores in LAU m and time slot
ts
```

3.4.3.2. For O-D matrices[6]

To produce the O-D matrix it is necessary to attribute to each user where he/she resides and where he/she moves during the day. Differently to the heuristics in section 3.4.3.1 to produce aggregates of residential population, in the mobility aggregates the daytime and the day of the week of phone activities is very important for localizing the user. In fact, in the O-D matrix, we need to determine the origin location, that typically refers to the sleeping hours, and the destination location, that typically refers to daily hours of weekdays.

In this case, we used the BSA approach described in section 3.2.2, in particular we shall assign as origin the location with the most activities during hours from 21:00 to 9:00 and as destination the location with the most activities during hours from 9:00 to 21:00 of weekdays. The locations, both the origin and the destination ones, are assigned using the BSA approach. In fact, we will use the proportion of each BSA of the antennas over the LAUs to statistically locate each users. The users will be actually located in the most frequent LAUs.

[6]Prepared in collaboration with Francesco Altarocca and Raffaello Martinelli.

Aggregating these data at the desired level is possible to produce the O-D matrix. As an example, in figure 3.26, we show an O-D matrix for the province of Pisa (Italy), based on mobile phone data for February, 2017.

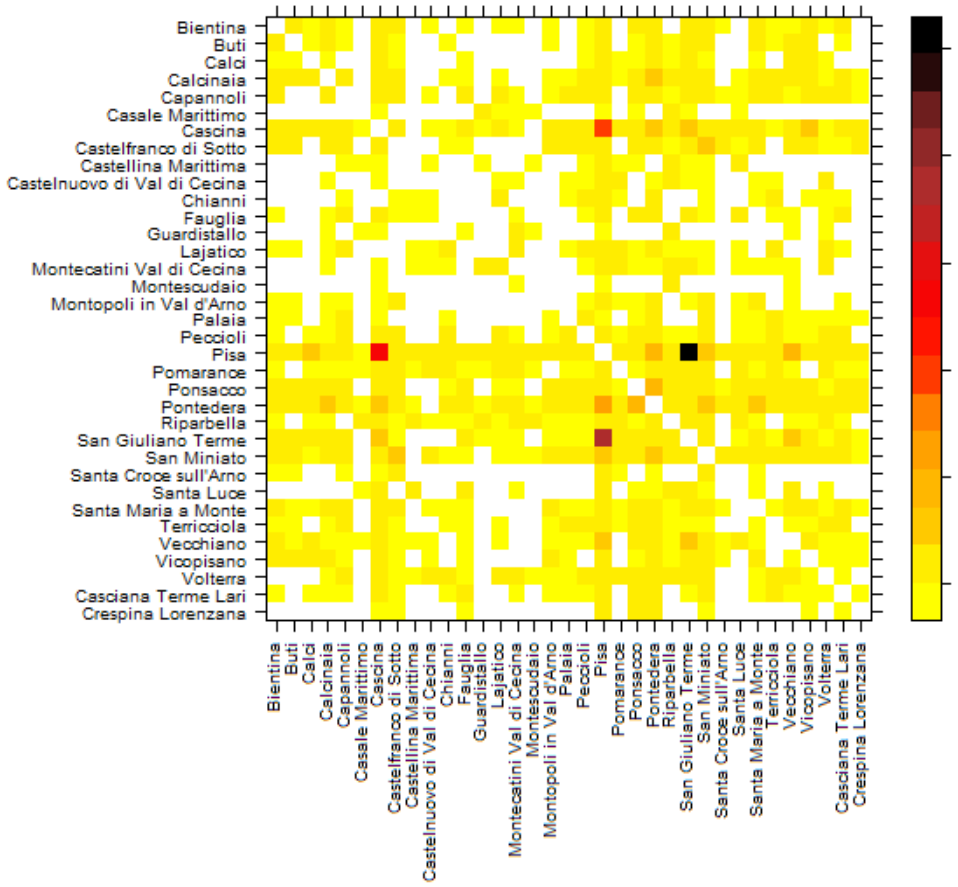


Figure 3.26 Origin-Destination matrix for Pisa province, representing home to work flows.

We define:

- m: LAU, m = 1, . . . , m1, m2, . . . , M.
- s: SIM, s= 1, . . . , S.

- c: call.
- sh: time slot "sleeping hours", from 21:00 to 9:00.
- dh: time slot "daily hours", from 9:00 to 21:00 in weekdays.
- BSAFract(cell): list of BSA fraction of a cell on the LAU as defined in the BSA approach (see section 3.2.2). Notice that for each cell Sum(fract(m, cell))=1 for all LAUs m in BSFract(cell).
- COMM: subset of SIMs identified as commuting. Each SIM in COMM has a single origin and a single destination.
- od(m1, m2): number of SIMs in COMM moving from m1 (origin) to m2 (destination). Notice that Sum(od(m1, m2)) $\leq$ S for all m1,m2 (it is exactly the number of SIMS in COMM).

The algorithm to compute this O-D matrix is given in pseudocode by

3.4.4. What about pre-aggregated data?

Quite often the aggregation is done by the MNO through a process resulting of a negotiation with the NSI. The methods for those aggregation thus depend a lot of the indicators that are aimed at constructing. Regarding present population different MNOs have been providing different types of aggregates.

Proximus example

Statbel (Statistics Belgium), Eurostat and Belgium's former incumbent network operator Proximus (about 41% market share) ran a joint project from December 2015 until March 2017 to explore the possibilities of mobile phone data for commercial and statistical purposes (see Meersman et al. (2016) and several follow-up studies).

The statistical partners had no direct access to the Proximus database derived from network signaling events and consisting of approximately 395 billion mobile device positioning records (13 months, 1 billion mobile phone localisations per day at Voronoi cell level, based on the location of the devices of 5-6 million clients, once every half hour on average); Proximus does not link its customer database to these signaling records, so no attribute whatsoever about the person owning a device can be known directly. Instead of direct access, Statbel and Eurostat defined use cases: the statistical results to be arrived at, together with the query, the selection of data needed to compile them. The output of this query may consist of individual records but in these cases only aggregated or otherwise transformed data were considered; this reduces storing and processing limitations,

Algorithm 2 Algorithm to calculate od(m1,m2)

```
for SIM s do
    LSC(s) = list of calls for s in the time slot sh          ▷ origin LAU of SIM s
    CSC(s) = count of LSC(s)
    for call c in LSC(s) do
        c.cell is the cell of c
        BSAList(c)= BSAFract(c.cell) is the list of BSA fraction of c.cell
        for fract(m,c) in BSList(c) do
            ps(sm) += fract(m,c)/CSC(s)
        end for
    end for
    s.mo=argmax(ps(s,m)) on LAU m
    LDC(s) = list of calls for s in the dh time slot          ▷ destination LAU of SIM s
    CDC(s) = count of LDC(s)
    for call c in LDC(s) do
        c.cell is the cell of c
        BSAList(c)= BSAFract(c.cell) is the list of BSA fraction of c.cell
        for fract(m,c) in BSList(c) do
            pd(s,m)+= fract(m,c)/CDC(s)
        end for
    end for
    s.md= argmax(pd(s,m)) on LAU m;
    if s.mo != s.md then      ▷ Commuting identification and OD matrix increment
        od(s.mo, s.md)++
    end if
end for
```

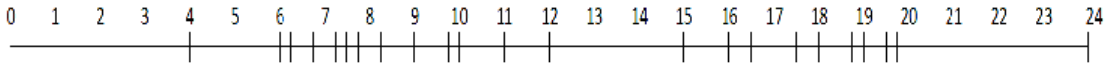
but also avoids privacy issues. It has the disadvantage that the MNO has to be trusted to carry out the query as agreed, and that no tacit assumptions cause misunderstandings.

The study mentioned above was based on counts every 15 minutes of all mobile devices present in each of the approximately 11.000 cells covering the Belgian territory, for one weekday and one Sunday. Some procedure is accomplished by the MNO to deduplicate mobile phones detected in several cells during 15 minutes, yet this methodology is unknown to the institute. Even this limited dataset with somewhat over 2 million records made it possible to estimate the population density correlating at 0.85 with Census results based on the population register. Several similar queries have been devised:

- tourism statistics: SIMs observed in a specific foreign country during a 250-day

period (see Seynaeve et al. (2016));

- actual present population across time and estimation of living place and workplace: 45 counts per Voronoi cell per day, for 12 months, extrapolated for the local market share of Proximus, resulting in about 180 million records;



- detailed living place-workplace matrix: cross-tabulation matrix of 11.000×11.000 Voronoi cells based on individual mobile devices tracked at different times of every day of October and summated per matrix cell.

Telekom example

Due to legal requirements the acquisition of sensitive data is very difficult. Therefore, the German Federal Statistical Office (Destatis) worked in close cooperation with the Data Protection Agency to receive initially simple datasets, where only the daytime- and resident population can be analyzed with this pre-aggregated data.

The long-term goal is the use of dynamic data in form of origin-destination matrices to estimate commuting flows.

Destatis concluded a cooperation agreement with the German Telekom that includes not only access to the mobile phone data as well as further (tailor-made) data provisions, but also a methodological collaboration.

The pre-aggregated data capture anonymized and aggregated signaling events, which depend on Telekom traffic cells with a needed minimum number of mobile events of 30. The traffic cells are of different size. The smallest grid is 500 times 500 meters and the biggest one 8000 times 8000 meters. Moreover, only mobile activities of Telekom contract customers were included.

The German mobile providers Telekom, Vodafone and Teléfonica have a market share of one third each (state 3rd quarter 2017). Furthermore, the dataset is determined by the dwell time of mobile events. A dwell time is defined as the length of stay of a mobile device without movement between locations or grid cells. Since the data includes only signaling data, one cannot say what kind of mobile activity was made only that an activity was made. This can include for example a phone call, sending or receiving a

message or the mere record of the mobile device at the telephone pole. Therefore the choice of an optimal dwell time is important particularly with regard to the analysis of daytime- and resident population. Depending on the length of the dwell time, mobile activities will be counted if they maintained active for the entire time. This means the amount of mobile activities will be determined by the choice of the dwell time. In case of analyzing the daytime- and resident population, longer dwell times may be preferable. The mobile phone data include also some characteristics of mobile phone users like the share age group, gender and nationality per grid cell. To ensure privacy, all mobile activities are available only as aggregated data.

Aggregated data in the ONS

ONS sourced two small samples of aggregated and fully weighted and modelled origin-destination commuter flows, derived from the geo-location traces of mobile phones by observing and making inference on the repeat patterns of movement and dwell time over four weeks. These samples were designed to be equivalent to 2011 Census Travel to Work flows and high level overviews of the methods used were also provided. Each sample had a different study area. The privacy issues involved with the use of mobile phone data led ONS to seek only aggregated data. Moreover, as the infrastructure and methods required to produce commuter flows from mobile phone data (MPD) are complex and unfamiliar to ONS, the first research was to simply compare MPD-modelled outputs with equivalent census data to examine how well they matched. This was facilitated as all three of the major mobile networks in UK, each with excess of 25% market share of mobiles, currently have business operations to produce this sort of statistical output.

Each sample of MPD-commuter flows was provided on two UK based geographies:

- Local Authority (LA): LAs are an UK administrative boundary. There are 346 of them across the England and Wales. They vary greatly in area and in the population that resides in them. Our two study areas each comprised of three contiguous LAs. Each sample contained all LA to LA commuter flows that originated or had a destination within the appropriate study area.

- Middle Layer Super Output Area (MSOA). MSOAs are statistical geographies designed to contain a similar number of UK residents, around 7000 residents of all ages. Naturally, an MSOA in a rural area might cover a large area, whilst highly urban MSOAs are much smaller. Helpfully, the dispersal of cell tower antennas across the UK broadly follows a similar pattern: areas of high population have more cell towers than rural areas, and consequently smaller cell areas (as for MSOAs). Both samples of mpd-commuter flows provided MSOA to MSOA

commuter flows for all flows originating or ending in any MSOA within the appropriate study area.

The attraction of using MSOAs includes the ability to conduct research using various other official data produced for this geography. They are also designed to fit into LAs. Perhaps a drawback, to using this geography is that the shape of MSOAs can be very intricate and convoluted, wrapping around other MSOAs and how this might affect the mapping to mpd cell areas is unknown.

Disclosure threshold
To further guarantee privacy, each data supplier had a general policy to set a threshold for the minimum commuter flow they would release to any third party, including the ONS. These thresholds, of 15 and 20 respectively, were applied to the weighted estimates of mpd-commuters but would equally have been applied to the unweighted mpd-counts (which were not supplied to ONS in any case). Although the methods of applying these thresholds differed between the two samples, there were similar effects at LA and MSOA level flows.

The distribution of commuter flows is highly skewed, with lots of small value flows and only a handful of a more significant magnitude, typically between neighbouring areas. For LA to LA commuter flows the MPD identified flows that, according to Census, represented in excess of 95% of all commuters. However, at MSOA to MSOA level, for both samples, the MPD-flows identified flows representing circa 60% of all commuters according to Census. The missing flows in the MPD were typically small value flows in Census and had been subject to the disclosure thresholds.

A further complication is that there is a great customer need for information on commuter flows by the main mode of transport (i.e. rail, road, cycling, walking etc.). This breakdown was also provided to ONS, although at MSOA to MSOA level virtually all commuter flows were subject to the disclosure threshold as the values were too small.

Research is ongoing to inform on the spatial limitation of MPD to produce commuter estimates, including the complexity of mapping cell areas and considering the relative standard errors of MPD-estimates when the number of commuters to estimate is very small.

4

From aggregated data to official statistical products

Executive summary

This chapter addresses the process going from the aggregated data to the final statistical product to be disseminated. i.e. it concentrates on the inference exercise connecting these data with the target population of analysis.

We briefly discuss the limitations of the traditional sampling design methodology to carry out the inference for these data to the target population under study and argue against some common arguments in detriment of the incorporation of new non-probability sampling techniques.

Inspired by ecological sampling techniques addressing the species abundance problem we formulate a hierarchical model to produce estimates of population counts ofa given target population by combining both aggregated mobile phone data and official population figures.

With a clear pragmatic mind we choose the Bayesian approach for its computational power. We briefly discuss how to use weakly informative priors in the estimation process thus avoiding potential problems with subjectivity in the production of the final estimates.

The model is based on two fundamental assumptions:

- At a point in time, individuals are assumed to be physically in the territorial (administrative) cell appearing in the population register (or auxiliary survey data).
- Mobility patterns of individuals do not depend on the concrete MNO they are subscribed to.

4.1. Sampling design methodology and the curse of representativity

The adoption of Big Data sources in the standard production of official statistics in statistical offices faces many difficulties. In preceding deliverables access has been clearly depicted as a first major obstacle, but this is neither the only one nor the most difficult. Once microdata have been accessed and prepared, how should we further process them? In particular, how is the inference exercise between the collected and aggregated data and the target population to be conducted? Should we restrict ourselves to the traditional methodology based on sampling designs or are we somehow obliged to expand the number of techniques to be used in producing and disseminating new official statistical products? We shall briefly address these questions focusing on mobile phone data thus motivating and justifying our choices for our methodological proposal.

We share the view that the production of official statistics must be firmly rooted on scientific grounds. Indeed, this is the case of the statistical inference methodology based on sampling designs traditionally used in statistical offices allowing us, together with other factors, to fulfil high-level quality standards. Therefore, we must provide strong reasons to why this scheme is not to be followed and in such a case to clearly show how the quality standards are also fulfilled with mobile phone data.

As T.M.F. Smith (1976) already pointed out, the design-based inference seminally introduced by J. Neyman (1934) allows the statistician to make inferences about the population *regardless of its structure*. In our view, this is the essential trait of design-based methodology in Official Statistics over other alternatives, in particular, over model-based inference. As M. Hansen (1987) already remarked, statistical models may provide more accurate estimates *if the model is correct*, thus clearly showing the dependence of the final results on our a priori hypotheses about the population. Sampling designs free the official statistician to make hypotheses sometimes hard to justify and to openly communicate.

This trait appears in the statistical methodology under the use of (asymptotically) design-unbiased linear estimators of the form $\widehat{T} = \sum_{k \in s} \omega_{ks} y_k$, where s denotes the sample, ω_{ks} are the so-called sampling design weights and y stands for the target variable to estimate the population total $Y = \sum_{k \in U} y_k$. A number of techniques exists to deal with diverse circumstances regarding both the imperfect data collection and data processing procedures so that non-sampling errors are duly dealt with (Särndal and Lundström, 2005; Lessler and Kalsbeek, 1992). These techniques lead us to the appropriate sampling weights ω_{ks}. Sampling weights are also present in the construction of the variance estimates and thus of confidence intervals for the estimates.

The interpretation of a sampling weight ω_{ks} is extensively accepted as providing the

number of statistical units in the population U represented by unit k in the sample s, thus settling the notion of representativity on apparently firm grounds. This combination of sampling designs and linear estimators, complemented with this interpretation of sampling weights, stands up as a robust defensive argument against any attempt to use new statistical methodology with Big Data sources. Indeed, one of the first rightful questions when facing the use of Big Data is how the data represent the target population. In particular, for the case of mobile phone data, being aware of the different profiles of MNOs' subscribers, the question is clearly meaningful.

However, before giving due response with new methodology, we believe that it is of utmost relevance to be aware of the limitations of the sampling design methodology in the inference exercise linking sampled data and target populations. This will help stakeholders be conscious about changes brought by new methodological proposals and view the challenges in the appropriate perspective.

Firstly, the notion of representativity is slippery business. This concept was already analyzed by Kruskal and Mosteller (1979a,b,c, 1980) in this line. Surprisingly enough, a mathematical definition is not extensively found providing Bethlehem (2009) an exception with very difficult practical implementation (we would need to know the population distribution). Nonetheless this has not been an obstacle for the extended use of the concept of representativity even in a dangerous way. From time to time one can hear that the construction of the linear estimators is undertaken upon the basis of being ω_{ks} the number of population units represented by the sampled unit k, thus amounting $\omega_{ks} \cdot y_k$ to the part of the population aggregate accounted by unit k, finally being $\sum_{k \in s} \omega_{ks} \cdot y_k$ the total population aggregate to estimate. A strong resistance is partially perceived in Official Statistics against any other technique not providing some similar clear-cut reasoning accounting for the representativity of the sample. This argument is indeed behind the restriction upon sampling weights construction for them to be positive or even greater than 1 (a unit not representing even itself?) in sampling weight calibration procedures (see e.g. Särndal (2007)).

Let us provide our rigorous view on the inference with sampling designs. The randomization approach does allow the statistician not to make prior hypotheses on the structure of the population to conduct inferences, i.e. the confidence intervals and point estimates are valid for any structure of the population. But this does not necessarily entail that the estimator must be necessarily linear. Given a sample s randomly selected according to a sampling design $p(\cdot)$ and the values $\mathbf{y}$ of the target variable, a general estimator is any function $T = T(s, \mathbf{y})$, being linear estimators a specific family thereof (Hedayat and Sinha, 1991).

Ultimately the goal of an estimation procedure is to provide an estimate as close

as possible to the real unknown target quantity together with a measure of the accuracy. The concept of mean square error, and its decomposition in bias and variance components (Groves, 1989), is essential here. Estimators with a lower mean square error guarantee a high-level quality standard estimation. No mention to representativity is needed. Furthermore, not even the requirement of exact unbiasedness is rigorously justified: compare the estimation of a population mean using an expansion (Horvitz-Thompson) estimator and using the Hájek estimator (Hájek, 1981).

What prevents us to use more complex functions? Apparently nothing. A linear estimator may be viewed as a homogeneous first-order approximation to $T(s, \mathbf{y}) \approx \sum_{k\in} \omega_{ks} y_k$, but why not a second-order approximation

$$T(s, \mathbf{y}) \approx \sum_{k\in} \omega_{ks} y_k + \sum_{k,l\in s} \omega_{kls} y_k y_l ?$$

Or even a complete series expansion $T(s, \mathbf{y}) \approx \sum_{p=1}^{\infty} \sum_{k_1,\ldots,k_p\in s} \omega_{k_1\ldots k_p s} \cdot y_{k_1} \ldots y_{k_p}$ (see e.g. Lehtonen and Veijanen (1998))? However, the multivariate character of the estimation exercise at statistical offices provides a new ingredient shoring up the idea of representativity, especially through the concept of sampling weight. Given the public dimension of official statistics usually disseminated in numerous tables, numerical consistency is strongly requested on all disseminated statistics, even among different tables. For example, if a table with smoking habits is disseminated broken down by gender and another table with eating habits is also disseminated broken by gender, the number of total women and men inferred from both tables must be *exactly* equal. Not only is this restriction of numerical consistency demanded among all disseminated statistics in a survey but also among statistics of different surveys, especially for core variables such as gender, age, or nationality. Linear estimators can be made easily fulfilled this restriction by forcing the so-called *multipurpose property of sampling weights* (Särndal, 2007). This entails that the same sampling weight ω_{ks} is used for any population quantity to estimate in a given survey. This elementarily guarantees the numerical consistency of all marginal quantities in disseminated tables.

Notice, however, that this property has to be forced. Indeed, the different techniques to deal with non-sampling errors (e.g. non-response or measurement errors) rely on auxiliary information $\mathbf{x}$ so that sampling weights ω_{ks} are functions of these auxiliary covariates $\omega_{ks} = \omega_{ks}(\mathbf{x})$. Forcing the multipurpose property amounts to forcing the same behaviour in terms of non-response, measurement errors, etc. (thus social desirability or satisficing response mechanisms) regarding *all* target variables in the survey. Apparently it would be more rigorous to adjust the estimators for non-sampling errors on a separate basis looking only for a *statistical consistency* among marginal quantities. However, this is much harder to explain in the dissemination phase and traditionally the former option

is prioritized paving the way for the representativity discourse (now every sampled unit is thought to truly represent ω_{ks} population units).

Secondly, sampling designs are thought of as a life jacket against model misspecification. For example, even not having a truly linear model between the target variable y and covariates $\mathbf{x}$, the GREG estimator is still asymptotically unbiased (Särndal et al., 1992). But (asymptotical) design-unbiasedness does not guarantee a high-quality estimate. A well-known example can be found in Basu's elephants story (Basu, 1971). Apart from implications in the inferential paradigm, this story clearly shows how a poor sampling design drives us to a poor estimate, even using exactly design-unbiased estimators.

Finally, as already well-known in small area estimation techniques (Rao and Molina, 2015) and as R. Little (2012) called *inferential schizophrenia*, sampling designs cannot provide a full-fledged inferential solution for all possible sample sizes out of a finite population. Traditional estimates based on sampling designs show their limitations when the size of the sample for population domains begins to decrease dramatically. With mobile phone data one expects to avoid this problem by having plenty of data, but in the same line one of the expected benefits of this new data source is to provide information at an unprecedented space and time scale. So the problem may still remains in population cells with low market shares of a given mobile phone operator.

In conclusion, sampling design-based inference is a robust methodology providing firm scientific grounds for the production of official statistics but it is not a panacea for all potential situations we face when producing these statistics. An abuse or misuse of the notion of representativity should not be resorted to as an argument to defend this methodology against other alternatives. We believe that the key idea for a high-level quality estimation is not only to use low mean square error estimators, but also to show their robustness against misspecifications of any factor of variability (either the sampling designs or the underlying statistical models or whatever).

4.2. Non-probability sampling and ecological surveys

Can probability sampling still be used with mobile phone data? Let us very briefly remind that probability sampling is essentially the application of a sampling design $p(\cdot)$ on a *finite* population U of *known* size N composed of *identifiable* units u_k (Cassel et al., 1977). This sampling design of our choice will allow us to compute so-called first and second order inclusion probabilities π_k and π_{kl} which together with target variable values y and different procedures to account for nonsampling errors in terms of auxiliary covariates $\mathbf{x}$ will drive us to construct unbiased linear estimators $\widehat{T} = \sum_{k \in s} \omega_{ks}(\pi_k, \mathbf{x}_k) \cdot y_k$.

If we focus on the problem of producing population counts for a partition of the population into territorial cells at a given time period using aggregated mobile phone data, it is fairly clear that the statistician does not have any knowledge at all about the sampling mechanism selecting statistical units appearing in the data set. That is, the sampling design $p(\cdot)$ is completely unknown and this invalidates all the procedure to construct a design-based estimator. An alternative procedure to infer the population total in each cell from the data set must be put in place. Non-probability sampling schemes must be used.

In contrast with probability sampling methodology, which can be found in a small collection of excellent textbooks by Deming (1950); Hansen et al. (1966); Cochran (1977); Särndal et al. (1992) (to name a few), non-probability sampling techniques are dispersed over a set of disciplines developing their own specific methods (clinical trials, epidemiology, ...). In this sense, it seems advisable to look for methods applied in similar circumstances as in the case of mobile phone data.

A word of caution must be made regarding the concept of *sample*. More often that not one can hear an apparently appealing argument in favour of Big Data in strong detriment of any form of statistical inference: we do not need sampling because we have data galore. We will not argue upon the well-known danger of non-sampling errors (we just remind the reader that Yates (1965) himself as early as 1949 (first edition) already pointed out how a census could be more imprecise than a sampling survey because of these non-sampling errors). We shall focus on the subtleties behind the concept of sample. In design-based inference where the problem starts by having a finite population U of statistical units u_k the concept of sample is reduced to that of the selected set of these units according to an adequately chosen probabilistic design $p(\cdot)$ (Särndal et al., 1992). When design-based methodology cannot be applied and we have to resort to some kind of statistical modelling, the notion of *population* itself is different. Now the values of variables are assumed to be realizations of underlying random variables (Valliant et al., 2000) and the notion of population is rigorously formulated in terms of their probability distributions. The standard definition of *random sample* can be given as (Casella and Berger, 2002):

The random variables $X_1, \ldots, X_n$ *are called a* random sample of size n from the population $f(X)$ *if* $X_1, \ldots, X_n$ *are mutually independent random variables and the marginal probability density function of each* X_i *is the function* $f(x)$*. Alternatively,* $X_1, \ldots, X_n$ *are called* independent and identically distributed random variables with probability density function or probability mass function $f(x)$.

This is not to be confused with the ***selection of units*** whose variable values are measured. Thus under a statistical modelling approach we always sample the theoretical

population even if we select all units at hand. The no-sampling mantra in Big Data must be rigorously qualified as no-selection since sampling in the sense of the above definition is always present.

Now to focus on specific methods to be used with mobile phone data we have identified different elements in our problem. Firstly, for the time being we are using aggregated mobile phone data to estimate the size of a given target population (daytime population, tourist population, commuter population, ...). The determination of a population size is a common problem in many disciplines. Secondly, the concept of *detectability* as the probability that a statistical unit of the population is observed (Thompson, 2012) appears also in our problem. From the operations in a commercial telecommunication network it should be clear that only subscribers of the MNO at stake will be detected as potential target individuals (general population, tourists, commuters, etc.). Thirdly there is a strong spatial component in the problem. Not only is the population size estimated for a whole territory but also for a spatial distribution in territorial cells. Finally there is also a time component because the evolution of the population size in each cell is of interest.

The issue of detectability and sampling is treated in the species abundance problem, where the ecologist produces estimates for the number of individuals of a given species of interest across a specified spatial distribution of a geographical territory. Furthermore, the spatial and time components are present because they are of ultimate interest for the study of the evolution of the species abundance at stake. Therefore we find the species abundance problem (Manly and Navarro-Alberto, 2014; Royle and Dorazio, 2014) very similar to our problem. We have focused on this methodology to analyse whether it can be applied directly or after some due modifications to estimate population counts using aggregated mobile phone data.

As Royle and Dorazio (2014) brilliantly show regarding ecological inference, there exist two opposite views on how to make the inference from the sampled data n to the target population size N (for simplicity's sake we drop out subscripts and variables denoting space and time dependence). On the one hand, we have the observation-driven view in which no attempt to model the target variable N is done. The estimation process entirely rests upon the observation procedure. A simplified version of this approach can be exemplified as setting the estimator

$$\widehat{N} = \frac{n}{\hat{p}},$$

where $\hat{p}$ denotes the detection probability of every statistical unit (assumed the same for each unit). No attempt to model the process driving the value of N is carried out. A rigorous footing of this formula can be easily given by modeling $n \simeq \text{Binomial}(N, p)$ under the assumption of homogeneous detection probabilities for all units in the cell

(Thompson, 2012). Then $\mathbb{E}n = N \cdot p$, thus we write $\widehat{N} = \frac{n}{p}$ to have an unbiased estimator for N. Usually the detection probability is unknown and must be estimated, hence $\widehat{N} = \frac{n}{\hat{p}}$. Notice how the whole method focuses only on the observation procedure of the units of interest.

This approach has been experimented at Istat to predict population estimates using mobile phone data. Actually, the target interest of Istat was to evaluate via these data the risk of under/over coverage of administrative population registers as well as an Origin/Destination matrix, so the population estimates obtained via these mobile phone data were an intermediated stage of the analysis rather than the final estimates of interest.

In this application by Istat, let i indicate the area level of the estimates, i.e. the LAU, n_i indicate the aggregated counts from mobile phone data users for location i. The n_i were obtained identifying for each mobile subscriber the location with the most mobile device activities during hours from 21:00 to 9:00, so using as a proxy of the resident population the so-called sleeping population. Finally, let p_i indicate the detection probability of every unit (mobile subscriber) for the LAU i. These probabilities have been estimated by $\hat{p}_i$ via the MNO market share at the LAU level, i.e. via the number of subscriptions of the specific MNO over the overall number of subscriptions, kindly provided at the LAU level by the MNO under the partnership agreement with the statistical agency.

Thus, the population estimates at the LAU level have been obtained by $\hat{N}_i = \frac{n_i}{\hat{p}_i}$. In section 3.4 more details are provided regarding the CDR aggregation procedure to obtain n_i. In the deliverable 5.5 on quality a comparison of population estimates obtained with different localization criteria is provided, highlighting gains of the current method in terms of coverage and accuracy.

On the other hand, a process-driven view can be followed by which a modelling exercise of the target variable N is conducted. Usually this involves the description in terms of statistical models of a complex underlying dynamics dependent on parameters to be estimated using the sampled data. For instance, as an overly simplified example we can pose

$$N \simeq \text{Poisson}(\lambda),$$

where the parameter λ depends on the sociodemographic characteristics of the cell. Notice how no reference to the observation process is made.

In between these two opposite views there exist many possibilities (Royle and Dorazio, 2014). The use of hierarchical models (Gelman et al., 2013) allows us to incorporate elements from both views in the inference process. In the inference model the observed data, the target process, and its underlying parameters must be given a joint proba-

bility distribution $\mathbb{P}(\text{data}, \text{process}, \text{parameters})$. The hierarchical model allows us to decompose this joint distribution as (Royle and Dorazio, 2014)

$$\mathbb{P}(\text{data}, \text{process}, \text{parameters}) = \mathbb{P}(\text{data}|\text{process}, \text{parameters}) \cdot \mathbb{P}(\text{process}|\text{parameters}) \cdot \mathbb{P}(\text{parameters}),$$

which can be conveniently interpreted as the combination of three components:

- an observation process (data given the underlying dynamical process driving the target variable N);
- a state process (the underlying process modelled in terms of its parameters);
- assumptions about the parameters driving not only the state process but possibly also the observation process.

The role of each component in the whole inference will strongly depend on the concrete formulation of the model.

It is also important to mention that specifying a statistical model does not make compulsory to use either the frequentist or Bayesian approach. This decision is up to the analyst. We are aware of how many lively and heated debates are around the frequentist vs. Bayesian approach. However we will adopt a pragmatic philosophy neither entering into nor providing rarely new arguments to the debate. Given the increasing computational power of the Bayesian approach we will use this methodology being aware of the dependence on prior hypotheses which will be made as weakly informative as possible thus bringing estimations very close to maximum likelihood-like frequentist methods. Ultimately we will assess the quality of the inference procedure using simulated populations. A complete analysis of quality will be undertaken in deliverable 5.5.

4.3. A hierarchical model to estimate population counts

Before formulating the hierarchical model to estimate population counts using aggregated mobile phone data let us reflect on the role of traditional official data in the advent of this new Big Data source. Should we produce different statistics according to the different sources at our disposal? In this case, how much consistent must the estimates be made? On the contrary, should we combine them into unified statistics possibly enlarging their spatial and time scope?

The preceding use of the multipurpose property of sampling weights, which pursues the numerical consistency of marginal aggregates among all breakdowns of target variables, invites to consider the proliferation of different estimates of the same population

aggregate as an undesired choice. In this sense, statistics using traditional and new data sources should be made consistent either by combining all data sources to produce the same statistics or by making direct or indirect use of both sources in the production of each separate statistics.

Regarding mobile phone data, in principle we can obtain population estimates at an unprecedented time and geographical scale (e.g. population figures every 30 minutes and on 1km $\times$ 1km cells). On the other hand, we also have population figures either from the Census of Population or from a population register fed by administrative data. All figures must be consistent.

Under these premises we make the first working hypothesis to formulate our hierarchical model: there exists a short time period t_0 in which both the target population $N_i(t_0)$ and the official population N_i^{Reg} of each cell i can be (statistically) equated. Furthermore, at this time period t_0 individuals are detected in the network in their corresponding geographical cells appearing in the population register. For reasons to be made clear below, we shall call this time period the initial time period.

Note that under this assumption, the detected number $N_i^{\text{MNO}}(t_0)$ of individuals in each cell i at the initial time period through the mobile network can be understood as a selection of units of the total number of individuals N_i^{Reg} according to the population register.

In this proposal we shall treat N_i^{Reg} $(i = 1, \ldots, I)$ as fixed external parameters in the model, although in the hierarchical modelling followed herein this could be also incorporated as random variables to be modelled according to some probability distribution with their own parameters.

The choice of the initial time period t_0 will depend on features of the official population and should be made on the basis that at that time period individuals detected with the network must be present in the territorial cell registered in the official population (e.g. very early in the morning or very late in the evening).

To follow the evolution of the population we make a second working assumption regarding its mobility: the mobility patterns are uncorrelated with the specific MNO individuals are subscribed to.

This hypothesis expresses the idea that people move around the geographical territory regardless of the MNO they are clients of. The only potential drawback in this assumption is an MNO operating only in a part of the geographical territory. Using only

these data for inferring the population can be misleading because no single individual will be detected in the rest of the territory. In these circumstances the combination of the observation and process views in the hierarchical model will help us ameliorate this lack of data.

Let us then introduce the notation to formulate the model. The target population to estimate at a given time t_n will be denoted by $N(t_n)$. The population at the initial time t_0 will be denoted by $N(t_0)$. We shall denote by $p_{ij}(t_0, t_n)$ the probability for individuals to move from cell i to cell j in the time interval (t_0, t_n). The number of individuals moving from cell i to cell j according to the network will denoted by $N_{ij}^{\text{MNO}}(t_0, t_n)$. As usual, we denote $N_{i\cdot}^{\text{MNO}}(t_0) = \sum_{j=1}^{I} N_{ij}^{\text{MNO}}(t_0, t_n)$. The number of individuals in cell i according to the population register (or external data source) will be denoted by N_i^{Reg}.

The complete model which we propose is specified by:

$$N_i(t_n) = \left[N_i(t_0) + \sum_{\substack{j=1 \\ j\neq i}}^{I} p_{ji}(t_0,t_n) N_j(t_0) - \sum_{\substack{j=1 \\ j\neq i}}^{I} p_{ij}(t_0,t_n) N_i(t_0) \right], \quad i = 1,\dots,I \tag{4.1a}$$

$$\mathbf{p}_{i\cdot}(t_0,t_n) \simeq \text{Dirichlet}\left(\alpha_{i1}(t_0,t_n),\dots,\alpha_{iI}(t_0,t_n)\right), \quad i = 1,\dots,I \tag{4.1b}$$

$$\mathbf{p}_{i\cdot}(t_0,t_n) \perp \mathbf{p}_{j\cdot}(t_0,t_n), \quad i \neq j = 1,\dots,I \tag{4.1c}$$

$$\alpha_{ij}(t_0,t_n) \simeq f_{\alpha ij}\left(\alpha_{ij}; \frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}\right), \quad i = 1,\dots,I \tag{4.1d}$$

$$N_i^{\text{MNO}}(t_0) \simeq \text{Binomial}\left(N_i(t_0), p_i(t_0)\right), \quad i = 1,\dots,I \tag{4.1e}$$

$$N_i^{\text{MNO}}(t_0) \perp N_j^{\text{MNO}}(t_0), \quad i \neq j = 1,\dots,I \tag{4.1f}$$

$$N_i(t_0) \simeq \text{Poisson}\left(\lambda_i(t_0)\right), \quad i = 1,\dots,I \tag{4.1g}$$

$$N_i(t_0) \perp N_j(t_0), \quad i \neq j = 1,\dots,I \tag{4.1h}$$

$$p_i(t_0) \simeq \text{Beta}\left(\alpha_i(t_0), \beta_i(t_0)\right), \quad i = 1,\dots,I \tag{4.1i}$$

$$p_i(t_0) \perp p_j(t_0) \quad i \neq j = 1,\dots,I \tag{4.1j}$$

$$(\alpha_i(t_0), \beta_i(t_0)) \simeq \frac{f_{ui}\left(\frac{\alpha_i}{\alpha_i+\beta_i}; \frac{N_{i\cdot}^{\text{MNO}}(t_0)}{N_i^{\text{REG}}}\right) \cdot f_{vi}\left(\alpha_i+\beta_i; N_i^{\text{REG}}\right)}{\alpha_i+\beta_i}, \quad i = 1,\dots,I \tag{4.1k}$$

$$(\alpha_i(t_0), \beta_i(t_0)) \perp (\alpha_j(t_0), \beta_j(t_0)), \quad i \neq j = 1,\dots,I \tag{4.1l}$$

$$\lambda_i(t_0) \simeq f_{\lambda i}(\lambda_i; N_i^{\text{REG}}) \quad \lambda_i(t_0) > 0, \quad i = 1,\dots,I \tag{4.1m}$$

$$\lambda_i(t_0) \perp \lambda_j(t_0)), \quad i = 1,\dots,I. \tag{4.1n}$$

where

- $[\cdot]$ denotes the nearest integer function;
- $\perp$ denotes independence between two random variables;
- $f_{\alpha ij}$ stands for the prior probability density function of the parameters α_{ij}. The notation $f_{\alpha ij}\left(\alpha_{ij}; \frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}\right)$ is meant to indicate that $\frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}$ should be taken as the mode of the density function;
- f_{ui} stands for the prior probability density function of the parameter u (see below) in cell i with mode $\frac{N_i^{\text{MNO}}(t_0)}{N_i^{\text{REG}}}$;
- f_{vi} stands for the prior probability density function of the parameter v (see below) in cell i with mode N_i^{REG};
- $f_{\lambda i}$ stands for the prior probability density function of the parameter λ (see below) in cell i with mode N_i^{REG}.

Let us provide the meaning for these terms. Equation (4.1a) states that the number of individuals in a cell i at time t_n equals the initial number of individuals at that cell plus those arriving from other cells in the given time interval minus those leaving for other cells in the same time interval. The number of individuals arriving and leaving are estimated using the transition probability $p_{ij}(t_0, t_n)$ among cells.

Next, to estimate these transition probabilities we model them for a given cell i as a multivariate random variable with a Dirichlet distribution with parameters

$$\boldsymbol{\alpha}_{i\cdot}(t_0, t_n) = (\alpha_{i1}(t_0, t_n), \ldots, \alpha_{iI}(t_0, t_n))^T.$$

These independent parameters, in turn, are given unimodal prior distributions $f_{\alpha_{ij}}$ with mode in $\frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}$ according to our second working assumption.

Equations (4.1e) to (4.1n) specifies how the inference is to be made for the population of each cell at the initial time instant t_0. If in a territorial cell i there are $N_i(t_0)$ individuals and we have an independent detection probability $p_i(t_0)$ for each individual through the mobile telecommunication network, then we will detect $N_i^{\text{MNO}}(t_0)$ individuals according to the aggregated mobile phone data naturally following a binomial distribution.

Now, the number of individuals $N_i(t_0)$ in each cell can be understood as a Poisson random variable (potentially arising from an underlying birth-death Poisson process –

see e.g. Grimmet and Stirzaker (2004)). These variables are pairwise independent and depend on unknown independent parameters $\lambda_i(t_0)$.

Now, the detection probabilities $p_i(t_0)$ in our mobile phone setting differs from the usual ecological setting. In the latter, the field work (observation sites, visual techniques, ...) allows us to propose a model for these probabilities according to the measurement process. In the telecommunication setting, in principle, the measurement process in cell i is always successful provided that a subscriber interacting with the network is within the territorial cell i. Thus at the given instant of time t_0 the detection probabilities $p_i(t_0)$ amount to establishing the proportion of individuals of interest at each cell i being detected by the MNO's cellular network. In other words, $p_i(t_0)$ are the proportions of individuals detected by the MNO at time t_0 in each cell i.

It is interesting to make a short reflection about these proportions $p_i(t_0)$ and the so-called local market shares. The latter are the number of subscribers of a given MNO in a cell i and they are sometimes considered as an important piece of information in performing the inference exercise from mobile phone data to the target population. We must stress that, in our view, it is not the concept of market share which is important but that of the actual proportion of individuals detected by the network. As an illustrative example, a call between a subscriber in a cell i and a non-subscriber in another cell j of a given MNO is certainly detected by the network in ***both cells***, thus potentially being part of the aggregated data N_i^{MNO} and N_j^{MNO}. This is a clear example of why having knowledge of the preprocessing and aggregation procedures from microdata is important for the final results.

We will model the detection probabilities $p_i(t_0)$ to account for the uncertainty we have in these quantities. They are modelled as beta random variables with parameters $\alpha_i(t_0), \beta_i(t_0)$ independently in each cell. The prior distribution of the beta distribution parameters $\alpha_i(t_0), \beta_i(t_0)$ arises from the following reasoning. We assume that $u(t_0) \equiv \frac{\alpha_i(t_0)}{\alpha_i(t_0)+\beta_i(t_0)}$ and $v(t_0) \equiv \alpha_i(t_0) + \beta_i(t_0)$ distribute independently according to $\frac{\alpha_i(t_0)}{\alpha_i(t_0)+\beta_i(t_0)} \simeq f_u\left(\frac{\alpha_i}{\alpha_i+\beta_i}; \frac{N_i^{\text{MNO}}(t_0)}{N_i^{\text{REG}}}\right)$ and $\alpha_i(t_0) + \beta_i(t_0) \simeq f_v\left(\alpha_i + \beta_i; N_i^{\text{REG}}\right)$, where f_u and f_v are respective weakly informative prior distributions for $\frac{\alpha}{\alpha+\beta}$ and $\alpha + \beta$. Notice that we have again made use of the auxiliary information coming from the population register ($\mathbf{N}^{\text{REG}}$). The quantity $\alpha_i/(\alpha_i + \beta_i)$ can be understood as a priori proportions of individuals detected by the MNO in cell i (e.g. should we have no information, then $f_u = \text{Unif}[0, 1]$). The parameters $\alpha_i(t_0) + \beta_i(t_0)$ can be essentially understood as the population size $N_i(t_0)$ of each cell (thus with support in $[0, \infty)$) upon which the detection is executed at that time instant, according to our first working hypothesis. For example, we may assume f_v to be a gamma distribution with parameters $(N_i^{\text{MNO}}(t_0)+1, \frac{N_i^{\text{MNO}}(t_0)}{N_i^{\text{REG}}})$.

In this way, the most probable value for the sample size is N_i^{REG} in consonance with the preceding hypothesis for $N_i(t_0)$.

Finally, the independent parameters $\lambda_i(t_0)$ are modeled with another weakly information prior $f_{\lambda i}$ which may incorporate the information we have from the population register or similar sources. Notice that the only a priori information incorporated is coming from this auxiliary source.

4.4. Getting the flavour of the model

To get a flavour of the model, let us make the following simplifying assumption. Let us suppose that the prior distributions f_u and f_v are degenerate so that equivalently we are assuming that we have full prior knowledge of the proportion of detected individuals[1] $u(t_0) = \frac{\alpha}{\alpha+\beta} = u^*(t_0)$ and of the population cell size $v(t_0) = N^*(t_0) = \alpha + \beta$ whose proportion of subscribers is detected by our MNO.

Then it is straightforward to show that the unnormalized posterior probability density $\mathbb{P}\left(\lambda(t_0)|N^{\text{MNO}}(t_0); N^{\text{REG}}\right)$ is given by

$$\begin{aligned}
\mathbb{P}\left(\lambda(t_0)|N^{\text{MNO}}(t_0); N^{\text{REG}}\right) &\propto f_\lambda(\lambda(t_0); N^{\text{REG}}) \cdot \text{Po}\left(N^{\text{MNO}}(t_0); \lambda(t_0)\right) \cdot \\
&\cdot \sum_{n=0}^{\infty} \frac{\lambda(t_0)^n}{n!} \frac{B(u^*(t_0)\cdot N^*(t_0) + N^{\text{MNO}}(t_0), (1-u^*(t_0))\cdot N^*(t_0) + n)}{B(u^*(t_0)\cdot N^*(t_0), (1-u^*(t_0))\cdot N^*(t_0))} \\
&\propto\!\!\propto f_\lambda(\lambda(t_0); N^{\text{REG}}) \cdot \text{Po}\left(N^{\text{MNO}}(t_0); \lambda(t_0)\right) \cdot \sum_{n=0}^{\infty} \frac{\lambda(t_0)^n}{n!} \cdot u^*(t_0)^{N^{\text{MNO}}} \cdot (1-u^*(t_0))^n \\
&\propto\!\!\propto f_\lambda(\lambda(t_0); N^{\text{REG}}) \cdot e^{-\lambda(t_0)u^*(t_0)} \cdot \frac{(\lambda(t_0)u^*(t_0))^{N^{\text{MNO}}(t_0)}}{N^{\text{MNO}}(t_0)!}, \qquad (4.2)
\end{aligned}$$

where we have used the approximation $\frac{\Gamma(x+a)}{\Gamma(x)} \approx x^a$ (which can be proved using Stirling's approximation) and where $\text{Po}(N;\lambda)$ denotes the probability function of a Poisson random variable N with parameter λ.

In the case of noninformative prior $f_\lambda \propto 1$ the posterior (4.2) corresponds to a gamma distribution for $\lambda(t_0)$ with parameters $N^{\text{MNO}}(t_0) + 1$ and $u^*(t_0)$. The mode of this distribution (thus the most probable value for $\lambda(t_0)$) is $\frac{N^{\text{MNO}}(t_0)}{u^*(t_0)}$. In turn, the most probable value for $N(t_0)$ in the model is $\lfloor \lambda(t_0) \rfloor = \lfloor \frac{N^{\text{MNO}}(t_0)}{u^*(t_0)} \rfloor$. With the due rigorous proviso, $u^*(t_0)$ can be somehow understood as a sampling weight connecting the population of

[1] For ease of notation we drop out the subscripts i regarding the cells, since they are independent.

detected individuals through the mobile phone network with the target population.

Suppose now that we assume a prior gamma distribution $\lambda(t_0) \simeq \text{Gamma}(\alpha + 1, N^{\text{REG}}/\alpha)$, where $\alpha > 0$. Then the posterior (4.2) is again a gamma distribution now with parameters $\text{Gamma}(\alpha(t_0) + N^{\text{MNO}}(t_0) + 1, u^*(t_0) + \frac{\alpha}{N^{\text{REG}}})$. The most probable value then for $\lambda(t_0)$ is $\frac{N^{\text{MNO}(t_0)}+\alpha}{u^*(t_0)+\frac{\alpha}{N^{\text{REG}}}}$ and for $N(t_0)$ is $\lfloor \frac{N^{\text{MNO}(t_0)}+\alpha}{u^*(t_0)+\frac{\alpha}{N^{\text{REG}}}} \rfloor$, which can be written as

$$\begin{aligned}\widehat{N}(t_0) &= \left\lfloor \frac{u^*(t_0)\cdot N^{\text{REG}}}{\alpha + u^*(t_0)\cdot N^{\text{REG}}} \cdot \frac{N^{\text{MNO}}(t_0)}{u^*}(t_0) + \frac{\alpha}{\alpha + u^*(t_0)\cdot N^{\text{REG}}} \cdot N^{\text{REG}} \right\rfloor \\ &\approx \frac{u^*(t_0)\cdot N^{\text{REG}}}{\alpha + u^*(t_0)\cdot N^{\text{REG}}} \cdot \left\lfloor \frac{N^{\text{MNO}}(t_0)}{u^*(t_0)} \right\rfloor + \frac{\alpha}{\alpha + u^*(t_0)\cdot N^{\text{REG}}} \cdot N^{\text{REG}} \end{aligned} \tag{4.3}$$

The estimate is thus an accurately approximate convex combination of both extremes: (i) having no auxiliary information at all about the population register and (ii) using only the information from the population register. The relative weight between these two components is provided by the parameter α.

The full Bayesian approach in the forthcoming sections incorporate our uncertainty in the knowledge of the hyperparameters (especially of $u(t_0) = \frac{\alpha(t_0)}{\alpha(t_0)+\beta(t_0)}$ and $v(t_0) = \alpha(t_0) + \beta(t_0)$), since we do not know with certainty the values of the proportion of individuals and of the actual population size of each cell upon which the detection is executed.

4.5. From the model to the estimation of population counts

Taking advantage of the computational power of the Bayesian approach, we will firstly compute the posterior probability for the population size of each cell at the initial time instant:

$$\begin{aligned}\mathbb{P}\left(N(t_0)\middle|N^{\text{MNO}}(t_0); N^{\text{REG}}\right) &\propto \int_0^\infty d\lambda \quad \mathbb{P}\left(N(t_0)|\lambda\right)\mathbb{P}\left(\lambda\middle|\mathbf{N}^{\text{MNO}}(t_0); N^{\text{REG}}\right) \\ &\propto \int_0^\infty d\lambda \quad \mathbb{P}\left(\lambda\middle|N^{\text{MNO}}(t_0); N^{\text{REG}}\right)\cdot \text{Po}(N(t_0);\lambda), \end{aligned} \tag{4.4}$$

where $\mathbb{P}(\cdot)$ will denote indistinctly a probability density function or a probability mass function. As expected, we need the posterior distribution for the hyperparameters, which moreover will allow us also to practise inference and simulations and to assess the quality of the model. This posterior distribution is readily expressed using the model as (dropping out the time dependence for ease of notation):

$$\begin{aligned}\mathbb{P}\left(\lambda\middle|N^{\text{MNO}};N^{\text{REG}}\right) &\propto \mathbb{P}\left(N^{\text{MNO}}|\lambda;N^{\text{REG}}\right)\\ &\propto \int_0^\infty\int_0^\infty d\alpha d\beta\int_0^1 dp\sum_{n=N^{\text{MNO}}}^\infty \mathbb{P}\left(N^{MNO}|p,N\right)\mathbb{P}\left(N|\lambda;N^{\text{REG}}\right)\mathbb{P}\left(p|\alpha,\beta\right)\mathbb{P}\left(\alpha,\beta;N^{\text{REG}}\right)\mathbb{P}\left(\lambda\right)\\ \propto \mathbb{P}\left(\lambda\right)\int_0^\infty\int_0^\infty d\alpha d\beta\int_0^1 dp\sum_{n=N^{\text{MNO}}}^\infty &\binom{n}{N^{\text{MNO}}}p^{N^{\text{MNO}}}(1-p)^{n-N^{\text{MNO}}}e^{-\lambda}\frac{(\lambda)^n}{n!}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\frac{f_u(\frac{\alpha}{\alpha+\beta};N^{\text{REG}})\cdot f_v(\alpha+\beta;N^{\text{REG}})}{\alpha+\beta},\end{aligned} \tag{4.5}$$

which reduces to

$$\begin{aligned}&\mathbb{P}\left(\lambda\middle|N^{\text{MNO}};N^{\text{REG}}\right)\propto\\ &\mathbb{P}\left(\lambda\right)\sum_{n=N^{\text{MNO}}}^\infty\binom{n}{N^{\text{MNO}}}e^{-\lambda}\frac{\lambda^n}{n!}\int_0^\infty\int_0^\infty d\alpha d\beta\frac{f_u(\frac{\alpha}{\alpha+\beta};N^{\text{REG}})\cdot f_v(\alpha+\beta;N^{\text{REG}})}{\alpha+\beta}\frac{B\left(\alpha+N^{\text{MNO}},\beta+n-N^{\text{MNO}}\right)}{B\left(\alpha,\beta\right)}\\ &\propto \mathbb{P}\left(\lambda\right)\sum_{n=N^{\text{MNO}}}^\infty\binom{n}{N^{\text{MNO}}}e^{-\lambda}\frac{\lambda^n}{n!}I_{N^{\text{MNO}},n-N^{\text{MNO}}}(N^{\text{REG}})\\ &\propto \mathbb{P}\left(\lambda\right)\text{Po}(N^{\text{MNO}};\lambda)\sum_{n=0}^\infty\frac{\lambda^n}{n!}I_{N^{\text{MNO}},n}(N^{\text{REG}})\\ &\propto \mathbb{P}\left(\lambda\right)\cdot\text{Po}(N^{\text{MNO}};\lambda)\cdot S\left(\lambda,N^{\text{MNO}},N^{REG}\right),\end{aligned} \tag{4.6}$$

where we have defined

$$I_{N^{\text{MNO}},n}(N^{\text{REG}})=\int_0^\infty\int_0^\infty d\alpha d\beta\,\frac{f_v(\frac{\alpha}{\alpha+\beta};N^{\text{REG}})\cdot f_v(\alpha+\beta;N^{\text{REG}}}{\alpha+\beta}\frac{B\left(\alpha+N^{\text{MNO}},\beta+n-N^{\text{MNO}}\right)}{B\left(\alpha,\beta\right)}, \tag{4.7}$$

$$S(\lambda,N^{\text{MNO}},N^{\text{REG}})=\sum_{n=0}^\infty\frac{\lambda^n}{n!}I_{N^{\text{MNO}},n}(N^{\text{REG}}). \tag{4.8}$$

Everything is thus reduced to the computation of the expression $S(\lambda,N^{\text{MNO}},N^{\text{REG}})$. Computational details have been put off to appendix A. Notice that so far we have made use only of equations (4.1e) to (4.1n) in the hierarchical model corresponding to the initial time period t_0.

With the unnormalized probability distribution (4.6) we can generate as many values of the parameter λ according to the model as we want, hence also of population counts N. As an illustration, let us consider a cell with $N^{(0)}=100$ individuals. The population register reports $N^{\text{Reg}}=97$ due to non-sampling errors. Let us consider that $N^{\text{MNO}}=19$

individuals are detected by the network at this initial time instant.

Next we need to choose the prior distributions. For the initial population counts we only need to specify f_u, f_v, and f_λ. As a weakly informative prior for u we can reason as follows. A priori $u^* = \frac{N^{\text{MNO}}}{N^{\text{Reg}}} = \frac{19}{97}$ appears as a highly probable proportion of detected individuals. Assuming an uncertainty of up to $\pm 0.15 \cdot u^*$, we set $f_u \simeq \text{Unif}(0.85 \cdot u^*, 1.15 \cdot u^*)$. As a weakly informative prior for v, the population register figure $v^* = N^{\text{Reg}}$ appears also as a highly probable population size for the cell. Assuming an uncertainty of up to $\pm 0.05 \cdot v^*$, we set $f_v \simeq \text{Unif}(0.95 \cdot v^*, 1.05 \cdot v^*)$. Notice that this uncertainty can be motivated by the uncertainty in the estimate N^{Reg}. Finally, as a weakly informative prior for λ, we choose a gamma distribution with mode in N^{Reg} and a large standard deviation. We set $f_v \simeq \text{Gamma}(\alpha + 1, \frac{N^{\text{Reg}}}{\alpha})$, with $\alpha = 1$. This value of α amounts to assuming a coefficient of variation for λ given by $\text{CV}(\lambda) = \frac{1}{\sqrt{\alpha+1}} \approx 0.71$, large enough not to force specific values for λ.

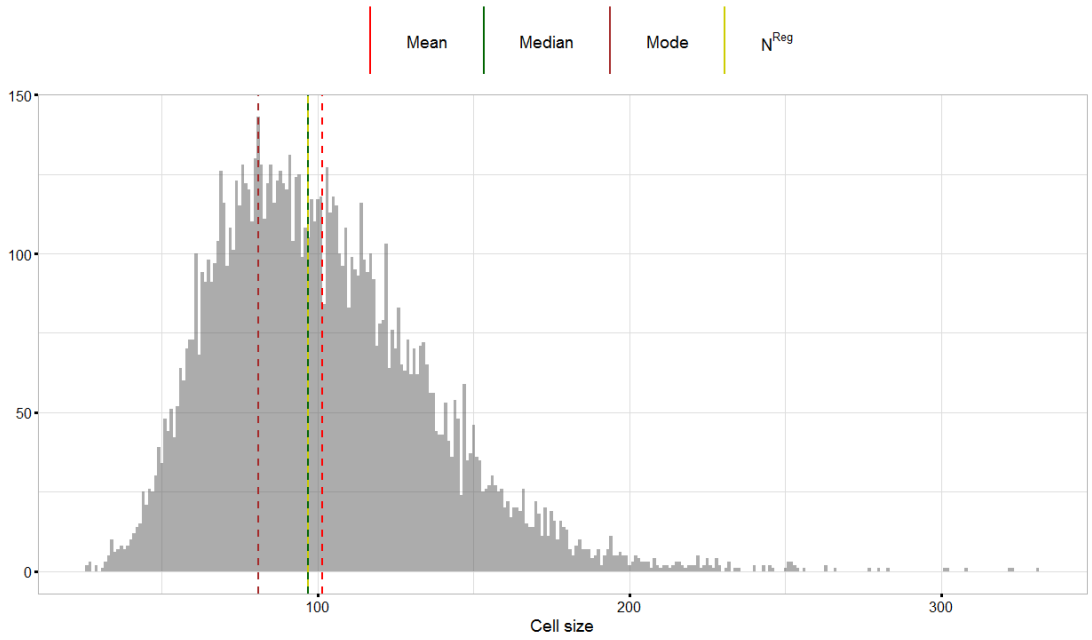


Figure 4.1 10000 simulated populations for $N^{\text{MNO}} = 19$, $N^{\text{Reg}} = 97$ and weakly informative priors.

In figure 4.1 we have depicted 10000 simulated populations according to the hyper-

parameter distribution given by (4.6). Once we can generate any arbitrary number of populations, we can empirically compute statistics such as the mean, the median, or the mode to provide a point estimate for the population count.

Notice how the mean and the median recovers quite accurately the population reported in the population register. The mode is more sensitive to outlying frequencies and will produce in general worse estimates.

The second step in the inference exercise is to estimate the time evolution of the population count of each cell. Now we need to make use also of equations (4.1a) to (4.1d) of the model. The philosophy is similar. We choose weakly informative priors for the hyperparameters $\alpha_{ij}(t_0, t_n)$, which allow us to generate probability vectors $\mathbf{p}_{i\cdot}(t_0, t_n)$ for each cell i. These are transition probabilities for an individual to go from cell i to any other cell j in the time interval (t_0, t_n). Notice that these transition probabilities are estimated using the transition matrix $N^{\text{MNO}}(t_0, t_n) = [N_{ij}^{\text{MNO}}(t_0, t_n)]_{1\leq i,j\leq I}$ of individuals detected by the network moving from each cell i to each cell j, according to our second working hypothesis.

With the initial population $N_i(t_0)$ simulated as above and the transition probabilities $p_{ij}(t_0, t_n)$, the equation (4.1a) allows us to compute the population count at time t_n for each cell i.

Let us consider a simplified example for illustrative purposes. Let us consider 12 cells in time instants t_0 to t_{672} (7 days at intervals of 15 minutes). Individuals move from cell to cell according to an unrealistic displacement (basically with higher probability to closer cells at each time instant). We represent schematically these movements in figure 4.2. These are the populations counts we must estimate.

The input data coming from the telecommunication network comprise the transition matrix on individuals detected from each cell i to each cell j in the successive time intervals (t_0, t_n), $n = 1, \ldots, 672$. To arrive at the estimates $N_i(t_n)$ we need to compute $N_i(t_0)$ and the transition probabilities $p_{ij}(t_0, t_n)$. The former is carried out as explained above whereas the latter requires to choose prior distributions for the hyperparameters $\alpha_{ij}(t_0, t_n)$. Again we choose weakly informative priors based upon our second working assumption. We consider as priors for $\alpha_{ij}(t_0, t_n)$ uniform distributions with their midpoint in $\frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}$ and coefficient of variation up to 20%. The priors for the initial population counts are constructed for each cell as above. For u_i we choose uniform distributions with midpoints at $\frac{N_{i\cdot}^{\text{MNO}}(t_0)}{N_i^{\text{Reg}}}$, $i = 1, \ldots, 12$ and interval lengths up to 30% of their midpoint, respectively. For v_i we also choose uniform distributions with mid-

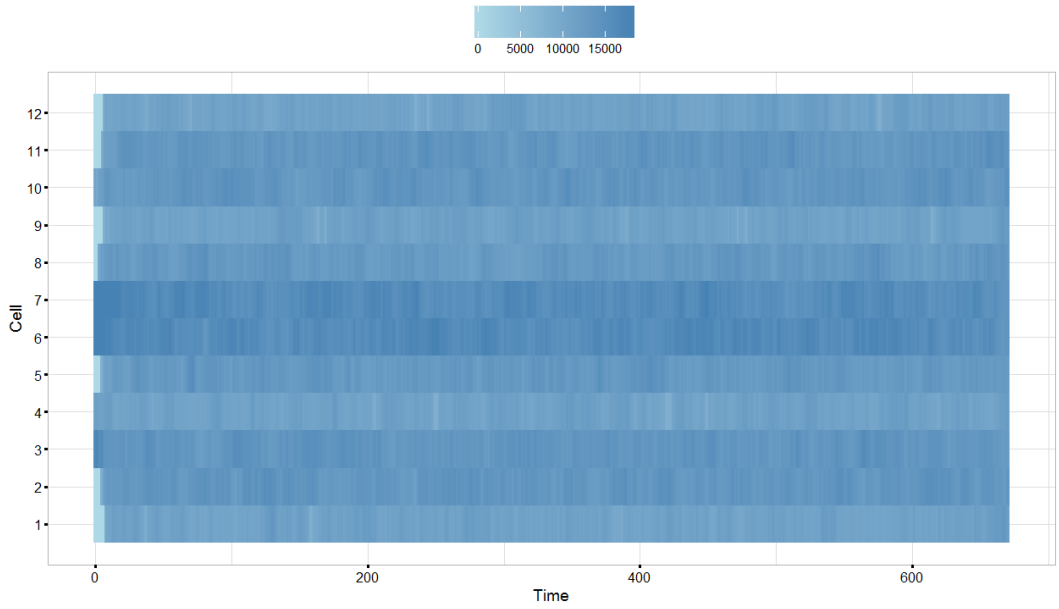


Figure 4.2 Simulated population counts among 12 cells along 672 consecutive time periods of 15 minutes for 7 days.

points at N_i^{Reg} and interval lengths up to 10% of their midpoint, respectively. The prior distribution of λ_i will be a gamma distribution with mode at N_i^{Reg} and a coefficient of variation of 60%.

The time evolution of the distributions of the generated posterior population size for each cell is depicted in figure 4.5 together with the evolution of the true (simulated) population size. We also include a comparison of actual and estimated population counts for each cell in figure 4.6 using the posterior mean and posterior median as estimators. The relative bias $\frac{\widehat{N}_i - N_i}{N_i}$ is similarly depicted in figure 4.7. The evolving population in each cell is reproduced with a relative bias of around $\pm 5\%$. Notice that estimates recover the true (simulated) population figures even despite the unrealistic displacement of individuals among cells. That is, estimates depend on the input data from the mobile telecommunication network only and not on particular characteristics of the mobility patterns of individuals.

It is important to pay attention to the fact that input data comprise essentially transition matrices $N^{\text{MNO}}(t_0, t_n)$ between cells for a ***fixed*** initial time t_0. This strongly conditions the aggregated data to be extracted from the statistical microdata sets. One

can immediately ask whether this is too stringent a condition and perhaps the more usual format in terms of transition matrices in consecutive time periods $N^{\text{MNO}}(t_{n-1}, t_n)$ could instead be used. Thus is it possible to adjust the estimation procedure for the hierarchical model to this kind of input data?

The answer is positive to a certain degree of approximation. Let us consider equation (4.1a):

$$N_i(t_n) = \left[N_i(t_0) + \sum_{\substack{j=1 \\ j \neq i}}^{I} p_{ji}(t_0, t_n) N_j(t_0) - \sum_{\substack{j=1 \\ j \neq i}}^{I} p_{ij}(t_0, t_n) N_i(t_0) \right], \quad i = 1, \ldots, I.$$

We can replace $N_i(t_0)$ by $N_i(t_{n-1})$ so that the population size at time t_n is computed using the population size at time t_{n-1}. This, in turn, is computed using the population size at time t_{n-2} and so on until arriving at the initial time period t_0 which is computed as above. Schematically this can be represented as

$$[N(t_n)] = [N(t_n)|N(t_0)] \cdot [N(t_0)] = [N(t_n)|N(t_{n-1})] \cdot \cdots \cdot [N(t_1)|N(t_0)] \cdot [N(t_0)].$$

Notice however that both procedures are not equivalent. To illustrate this let us consider the simplified case of t_0, t_1, t_2. Then we can write (we drop out cell subscripts for ease of notation)

$$\begin{aligned} \mathbf{N}^T(t_2) &= \left[\mathbf{N}^T(t_0) \cdot p(t_0, t_2)\right] \approx \mathbf{N}^T(t_0) \cdot p(t_0, t_2) \\ &= \left[\mathbf{N}^T(t_1) \cdot p(t_1, t_2)\right] \approx \mathbf{N}^T(t_1) \cdot p(t_1, t_2) \\ &= \left[\left[\mathbf{N}^T(t_0) \cdot p(t_0, t_1)\right] \cdot p(t_1, t_2)\right] \approx \mathbf{N}^T(t_0) \cdot p(t_0, t_1) \cdot p(t_1, t_2). \end{aligned}$$

Apart from numerical rounding errors, it is highly improbable that in general we have $p(t_0, t_2) = p(t_0, t_1) \cdot p(t_1, t_2)$ for all cells. The underlying mobility patterns of individuals are certainly expected to be complex enough to avoid this equality.

We could push the theory a bit further to find conditions under which this equality can be assured. However we find the empirical approach more interesting at this point. Let us illustrate the comparison of both approaches with another simple example. From a small-scale simulated population of individuals, both alternative aggregated data sets have been compiled. The population comprises 12223 individuals across 12 cells. We analyse their displaments along 48 time periods. They move from cell to cell according to an unrealistic pattern in which the closest cells are the most probable to move to for each individual and each time period. We compute the estimated population counts

for both methods and compare the estimates with the true (simulated) population size. Results for the relative bias with the posterior median estimator are depicted in figure 4.3 (for the posterior mean estimator they are similar).

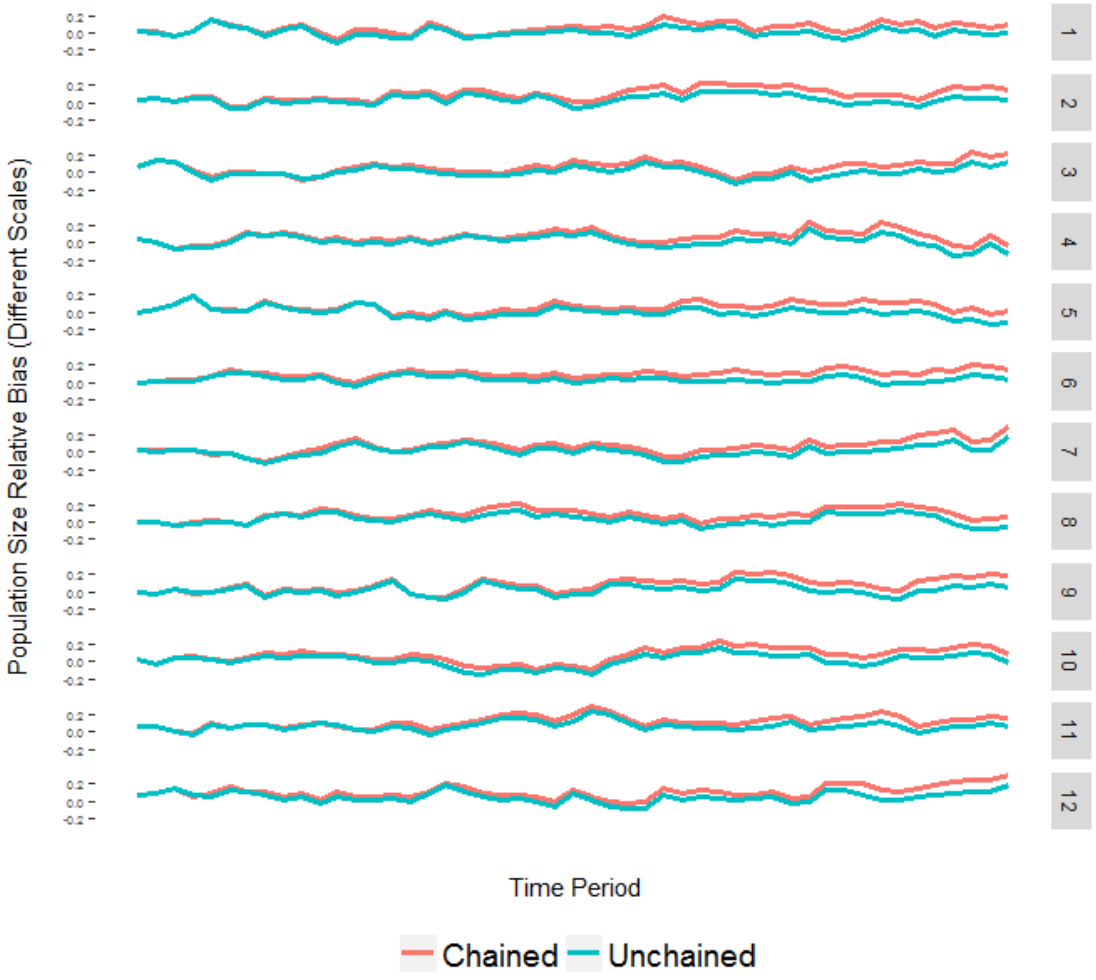


Figure 4.3 Comparison of estimated population counts using both approaches.

We observe a close similarity in the estimates both using the chained and unchained approaches. Notice however how the difference grows in time and how the chained approach overestimates with respect to the unchained method. A further comparison will be conducted in the deliverable 5.5. in terms of quality assessment, especially regarding

the accuracy and inference issues.

The remaining and highly relevant question is the choice of prior information. This is probably the most controversial issue in adopting such a methodology to produce official statistics especially in comparison to traditional survey sampling under design-based inference. We focus on this in the next section. The whole methodological proposal can be succinctly summarised by figure 4.4, where the different elements are represented as input and output objects.

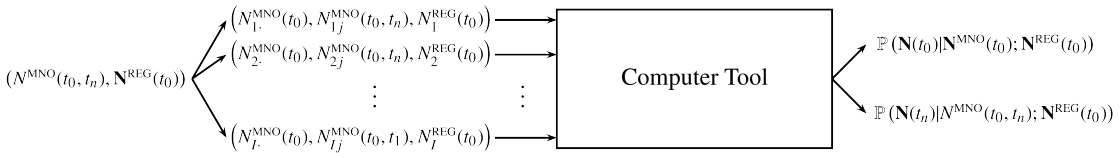


Figure 4.4 Schematic representation of the methodological proposal.

4.6. Prior information

As we argued in section 4.2 we have chosen the Bayesian paradigm under a pragmatic spirit mainly because of its computational power. As a beneficial by-product we get estimates under a strict rigorous inferential paradigm in contrast to the traditional design-based approach. As a potential drawback we are forced to choose prior distributions for the parameters in the model, thus possibly paving the way for an undesired degree of subjectivity in the production of official statistics.

However the argument against Bayesian inference based only on the need to choose prior distributions is just a caricature of this paradigm (Royle and Dorazio, 2014). In the frequentist approach, when using e.g. the maximum likelihood method the statistician is indeed obliged to formulate a probability distribution giving rise to a likelihood function to be optimized.

Nonetheless, this choice must be carefully analysed, justified, and disseminated when producing official statistics. Furthermore, we are not necessarily condemned to choices driving final estimates towards one direction or another. It is well-known (Gelman et al., 2013) that highly weakly information priors do produce similar results to the frequentist approach. Notice that in no case we are entering into the eternal Bayesian-Frequentist debate. Should we choose weakly informative enough priors, numerical estimates would be similar under both approaches and both (irreconcilable)

paradigms provide a firm footing for the quality assessment.

More concretely then, to keep under objective terms, we must provide a wide range of choices for weakly informative enough priors so that final estimates are free of subjective discretion.

For the prototyping model explored in this project we have focused on the following distributions:

- The uniform distribution Unif(x_m, x_M), where the bounds x_m and x_M will be selected according to the objective information at hand.
- The triangular distribution Triang(x_m, x_M, x^*) with support in $[x_m, x_M]$ and mode in x^*. Again the selection of parameters will be made according to the objective information at hand.
- The gamma distribution Gamma(k, θ) where the shape and scale parameters will also be selected according to the objective information at hand.

A combination of these distributions will be used to choose all priors in the model. Let us discuss each hyperparameter in turn. We drop out cell subscripts for ease of notation.

For the hyperparameter u with support on $[0, 1]$ the uniform distribution is easily justified as posing no prior information on the proportion of individuals detected by the network in the cell. This must be qualified by the choice of the range of the distribution. It seems natural to choose the a priori most likely value $u^* = \frac{N^{\text{MNO}}}{N^{\text{Reg}}}$ as the midpoint of the interval $[x_m, x_M]$ and an adequately interval radius expressing the uncertainty on this value. In this same line the triangular distribution can be also used. The choice for its mode seems to be more naturally the value $u^* = \frac{N^{\text{MNO}}}{N^{\text{Reg}}}$. For the support interval $[x_m, x_M]$ the same recipe applies. Notice now that we are giving less probability as we move away from u^*.

For the hyperparameter v with support on $[0, \infty)$ every of the three foregoing distributions can be considered a possible choice. For the uniform and the triangular distributions the same considerations as above can be made with the proviso $v^* = N^{\text{Reg}}$. For the gamma distribution we make use of its unimodality to naturally choose N^{Reg} as its mode. This sets just one of the parameters; we have freedom to set the other. As with the other two distributions, this second parameter is fixed according to the uncertainty on this modal value. This drives us to trivially set $k = \alpha + 1$ and $\theta = \frac{N^{\text{Reg}}}{\alpha}$, where $\alpha > 0$

expresses our uncertainty. Indeed the coefficient of variation of such a gamma distribution is given by cv $= \frac{1}{\sqrt{\alpha+1}}$: the greater α, the more certain the value of v around N^{Reg}.

For the parameter λ with support on $[0, \infty)$ the same kind of reasoning is valid. In particular, we have favoured for ease of computation the gamma distribution with a low value for α (high-degree of uncertainty).

For the hyperparameters α_{ij} again the same considerations can be made with the proviso that their most probable value can be naturally chosen as $\frac{N_{ij}^{\text{MNO}}}{N_{i\cdot}^{\text{MNO}}}$ for each cell i.

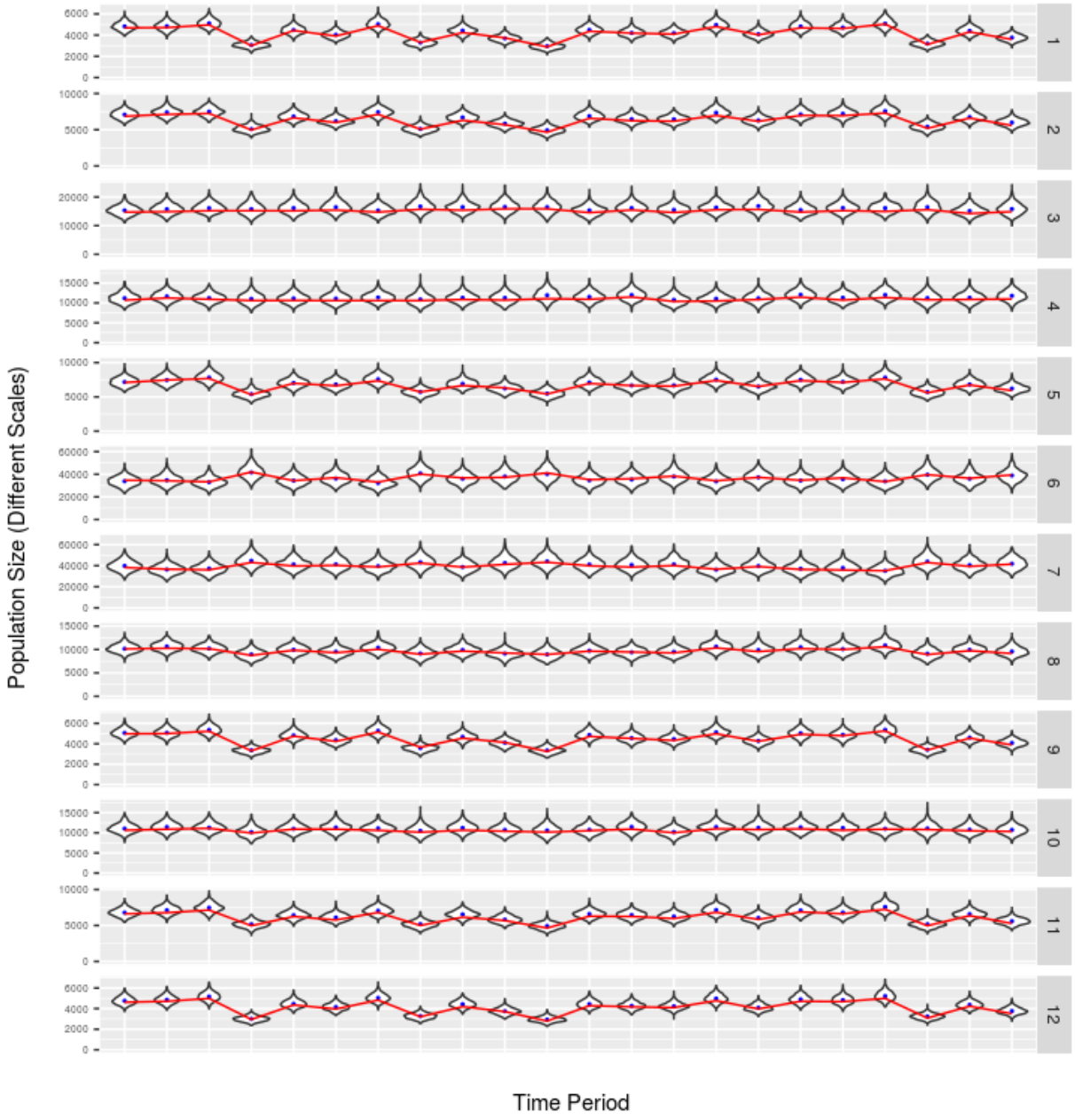


Figure 4.5 Distributions of the simulated population size of 12 cells along 672 consecutive time periods of 15 minutes for 7 days compared to the true (simulated) population (only every 30 time periods shown in the graph).

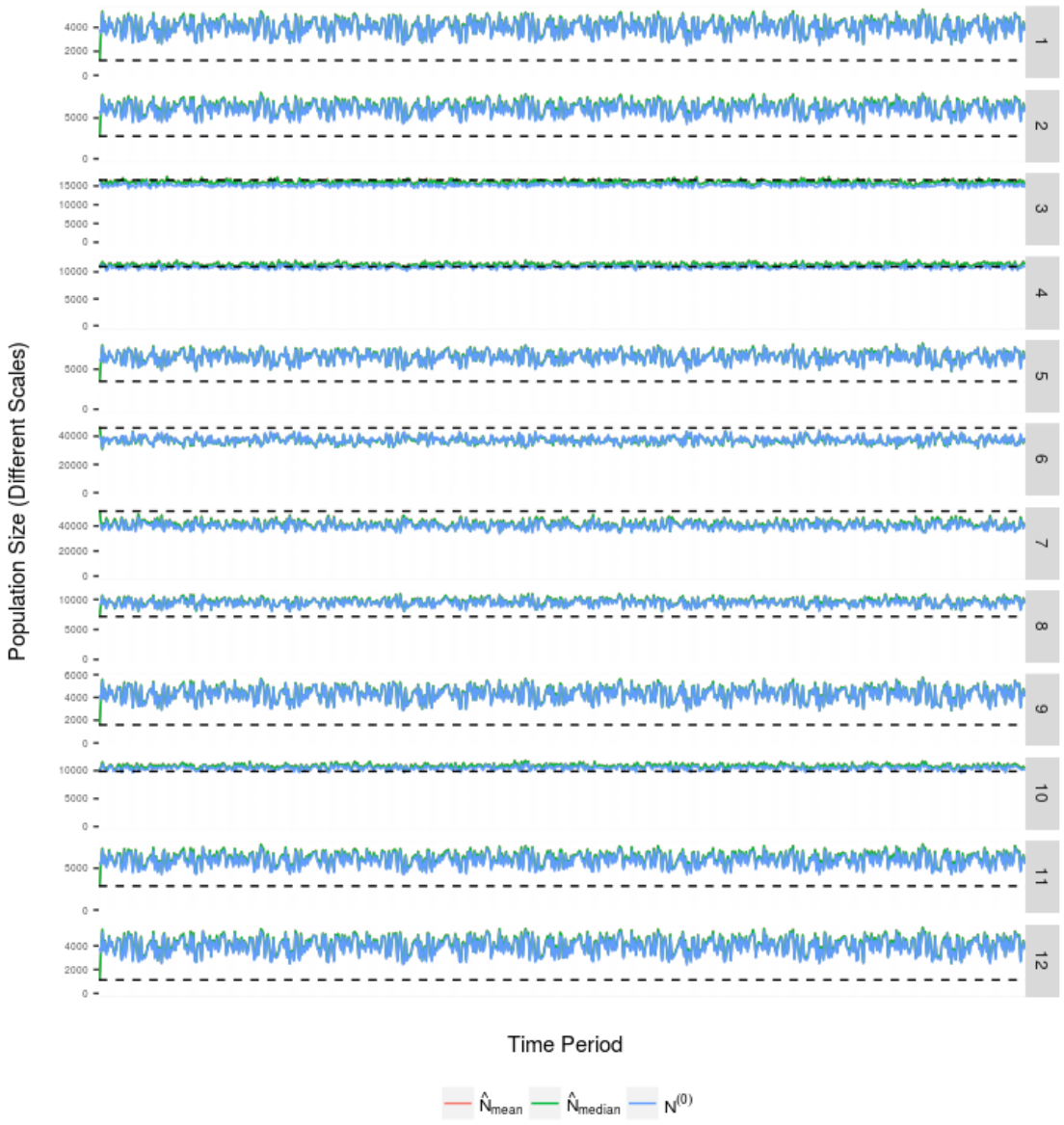


Figure 4.6 Estimated population size of 12 cells along 672 consecutive time periods of 15 minutes for 7 days.

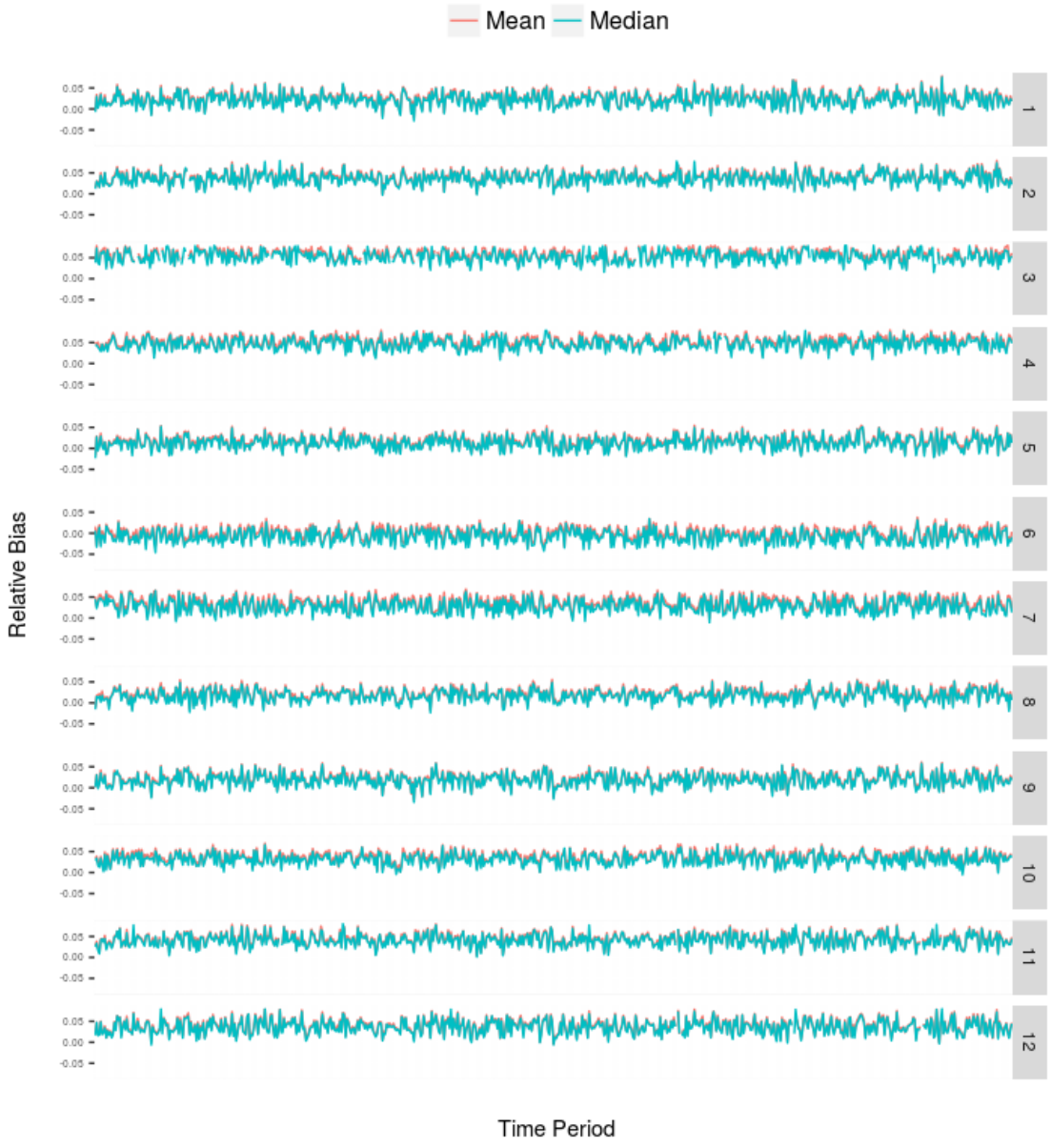


Figure 4.7 Relative bias of the estimated population size of 12 cells along 672 consecutive time periods of 15 minutes for 7 days.

5

Conclusions and proposals for the future

Executive summary

This document, far from proposing definitive solutions for the use of mobile phone data in the production of official statistics, aims at providing the first steps towards a standardised methodological framework for the integration of this new data source into the daily official statistical production.

Many aspects need to be tackled for this integration and subsequent use (not all of them dealt with here). As main conclusions we can state the following:

- A core data model building an organic database with mobile phone statistical microdata should be constituted in each NSI playing the role of traditional business and population registers.
- Further investigation in the procedures to generate variables (measures) for mobile devices and/or individuals and the aggregation of microdata should be undertaken, especially regarding important attributes as the geospatial information of each unit.
- The use of standard techniques in other data sources such as administrative registers (here we use the two-phase life-cycle model) should be promoted to gain in standardization and normalization.
- The hierarchical model proposed to produce estimates about the target population out of the aggregated data is just a first proposal which gives ample room for more complex elements (geostatistical models, selection bias correction techniques, ...).

As a first global conclusion it seems fairly clear that NSIs must have access to statistical microdata to advance in the development of a methodological framework like the one proposed here.

However, the access to data, as we described in preceding deliverables, is currently blocked for many reasons (essentially because other social agents have

begun to produce statistics competing with official statistics, something which clearly brings an economic profit). The ESS should not wait for this issue to be solved to go on developing these methodological frameworks with Big Data sources.

We propose to use simulated data at a large scale to keep on investigating using as much of real data as possible. When access will be finally granted, then simulated data will be duly substituted with real data. Additionally, simulated data will be of great value in the assessment of the new statistical methodology to be used with these new and forthcoming data sources.

Several conclusions can be drawn both from the methodological and the business process points of view. We see them interrelated driving us to some proposals for the future.

Firstly, after recognising common elements between mobile phone data and administrative data, we claim that already developed tools as the two-phase life-cycle model for statistical microdata can be similarly used to understand the generation of data and their preparation for statistical production. Indeed the adaptation of this model to the mobile phone data source reveals three phases in the generation of these data for official statistical production. In the first phase, raw telecommunication data are generated with the purpose of providing telecommunication services, which then enter into the second phase as input to produce statistical microdata at the level of individuals. As intermediate data sets we have the information in terms of mobile devices. These microdata, in turn, get further elaborated to arrive at the aggregated data, which will be used to make the final inference exercise to connect these data with the target population of interest (inbound tourists, resident tourists, commuters, ...).

Having a common framework to describe the generation of data is relevant not only for understanding the whole process for the correct methodological approach and the quality assessment but also to take strategic decisions regarding the business process and the unsolved question of the access to this new data source. Furthermore, the abstract formulation of such a model might invite us to consider the possibility of describing an integrating modernised process for all kind of data sources (survey data, administrative data, Big Data, smart statistics, ...).

As an immediate proposal derived hereof we suggest that a more detailed analysis of the application of the two-phase life-cycle model to *empirical* situations with statistical microdata and ideally raw telecommunication data should be pursued. Clearly this points necessarily towards a close collaboration with MNOs.

Secondly, a core data model has been proposed to create an organic normalized database for statistical microdata for all planned statistics which just needs a final customised subprocess to adapt the database to the statistical domain at stake (tourism, mobility, ...). The situation runs parallel to the compilation and maintenance of central business and population registers in the statistical offices. These registers have played, are playing, and will play an essential role in the production of official statistics. In the traditional design-based inference methodology they provide the source for constructing frame populations for any survey allowing statisticians to apply this inferential approach. Now a normalised central database fed with mobile phone data should also be pursued. However the current experience with mobile phone statistical microdata is short and much work in this direction has to be done.

Thirdly, in the whole generation process of both statistical microdata and aggregated data there exists a key step in producing final data to be used in the final inference with respect to the target population. This is the aggregation step on going from microdata to aggregated data. Many difficulties arise mainly because data were not originated for statistical purposes (e.g. the territorial divisions in the telecommunication industry are different to the usual administrative divisions for official statistical purposes). A systematization of the aggregation procedure with many proposals and techniques to be empirically compared is needed. Again we meet with the access to microdata to fully address this question. Some first proposals have been included in the text.

Finally, new methodology regarding the inferential step between aggregated data and the target population is needed. Sampling designs cannot be rigorously used any more. We have adapted ecological models addressing the species abundance problem to produce population counts using the aggregated data as input together with official population figures. The adaptation rests upon two assumptions. On the one hand, an initial time period is accepted to exist in which individuals are assumed to be physically in the territorial cell appearing in the official population figures. On the other hand, mobility patterns of MNOs' subscribers are assumed to be uncorrelated with the MNO in particular they are subscribed to.

Apart from rich technical details we want to underline important aspects regarding the business process and the strategy for incorporating mobile phone data in the standard production of statistical offices. Since the methodology is completely new, efforts must be made stressing the emphasis on the quality of the final product at least with the same quality standards as traditional design-based inference.

Connecting all these foregoing aspects as a strategic proposal we claim that the use of simulated populations should be boosted in the research and development phase. We

provide two strong reasons for this. On the one hand, from the purely methodological point of view, the fact of having a true (simulated) situation can be of utmost help in assessing the performance of the inferential techniques, be it Bayesian, frequentist, or whatever. Apart from the obvious fact of using statistical methods rooted on firm grounds, the new methodology proposed for new data sources should be illustrated with synthetic data as close as possible to real conditions. On the other hand, from the strategic point of view, the development of a modernised business process adapted to all kind of new data sources cannot wait for the issue of data access to be fully solved. The present work package (and the rest of this ESSnet project) follows a common plan to advance in the use of Big Data sources: first, access to real data; second, development of methodological proposal to process these data; third, development of computer tools implementing these solutions with as many technological novelties as needed; and fourth, an exhaustive assessment of the quality. The wealth of data in society and the data monetization process is already a reality and the access to many new data sources is currently blocked by diverse reasons. This work package on mobile phone data is an empirical demonstration of this fact. The European Statistical System should avoid this barrier by working on simulated data in parallel to addressing the data access issue. A repository of simulated data both at the micro and aggregated level and at the whole European scale would be a very useful tool. As the access issue is progressively solved, then immediately simulated data should be substituted by real data and daily production in standard conditions may start.

The current methodological proposal in the present document is not a final closed one. It is just a first proposal towards a methodological framework where new elements should be progressively added. We can point out the following:

- In the core data model there exist a number of parameterisations for diverse goals (4h, 6 months, ...). Alternative choices adapted to each country and the robustness of these choices should be studied in detail.

- The core data model has been presented with a clear tilt towards mobility and tourism. More statistical domains must be analysed so that the model is progressively completed.

- In the hierarchical model, independence among pairs of cells has been assumed in the specification of the model. This is clearly a simplifying assumption which can (and should) be dropped to explore more realistic hypotheses as the geospatial correlations among cells. This entails the use of hierarchical modelling techniques with geospatial data. There already exists methodology in other contexts (Banerjee et al., 2015).

- Apparently there is a clear selection bias in the use of mobile phone data to produce population counts. There might be population sectors with restricted access to mobile devices (children, elderly people, ...). This should be taken into account when constructing the prior distributions of model. Techniques as the Heckman correction (Heckman, 1979) may be useful at this point.
- An exhaustive comparison with other techniques should be conducted (e.g. Deville et al. (2014); Doyle et al. (2014)).
- An exhaustive search of adequate prior distributions to account for all possible situations with real data should also be undertaken.
- The model can be made progressively more complex by e.g. modelling also the number of individuals according to the official population register N_i^{Reg}. Equally, the specification for N_i in terms of a Poisson distribution is a simple hypothesis (although a very generic one). If some underlying dynamical process is assumed for N_i, this specification can certainly be made more complex.

In a more technically-minded realm, a cautious reader with some insight with data will notice that immediate needs for improvement can be easily detected. Just to mention a few:

- The candidate distribution for the rejection method to sample from the posterior distribution of the parameter λ has been chosen as a Cauchy distribution. More efficient choices should be explored.
- In this same line, the algorithm finding the mode of the posterior distribution of the parameter λ has been choosing exclusively in pursuance of having a prototyping framework up and running. More efficient algorithms should be investigated.
- If estimates based on posterior means or posterior medians are to be used, an alternative computation in terms of Monte Carlo integrals could be explored instead of producing simulated populations.

All methodological proposals and the core data model should be adequately implemented with appropriate IT tools. This is the goal of the next deliverable.

Appendix A

Computational details

We have put off all computational details to this appendix not only to provide a clearer statistical discourse for the inference exercise in the main text but also to concentrate all these details in the same section clearly providing the methodological input for its software implementation. In this same line, only mathematical/statistical contents are presented here. The software implementation will be one of the main points of deliverable 5.4.

Our choice of the Bayesian paradigm was motivated mainly by the computational power associated with this approach. The challenge is evident given the high degree of spatiotemporal breakdown in the data. To avoid computational overheads in this sort of analysis we have decided to carry out analytical developments as much as possible for the proposed model. In this sense instead of using general-purpose tools as Stan (Stan, 2018), JAGS (JAGS, 2018), ... we have undertaken the first step towards specific tools for the proposed model.

Therefore, following standard procedures for Bayesian computation we will sample values according to the model posterior distributions in order to find their mean, median, mode, ... as point estimates for the population counts. The collection of results in terms of the model specifications (4.1a) to (4.1n) is given by the following items:

- Either to find an analytical expression or a computational routine for the unnormalized density probability function $\mathbb{P}\left(\lambda|N^{\text{MNO}}; N^{\text{Reg}}\right)$ given by equation (4.6).
- To find a computational routine to sample values from this posterior distribution $\mathbb{P}\left(\lambda|N^{\text{MNO}}; N^{\text{Reg}}\right)$.
- To find a computational routine to sample values from the posterior distribution $\mathbb{P}\left(N|N^{\text{MNO}}; N^{\text{Reg}}\right)$ given by equation (4.4).

Once values can be sampled from the posterior distribution for N we can compute the posterior mean, the posterior median, the posterior mode, or any other position indicator to produce a point estimate. Furthermore, this posterior distribution will also allow us in deliverable 5.5. to approach the quality assessment of the estimates in terms of credible intervals, coefficients of variation,...

Since the model specifies independence between cells, subscripts and time dependence will be dropped for ease of notation throughout the whole appendix unless strictly necessary.

A.1. The unnormalized density probability function for λ

The unnormalized density probability function for λ is given by the expression

$$f(\lambda|N^{\text{MNO}}; N^{\text{Reg}}) = f(\lambda) \cdot \text{Po}(N^{\text{MNO}}; \lambda) \cdot S(\lambda, N^{\text{MNO}}, N^{\text{Reg}}) \tag{A.1}$$

The key point is the computation of the function $S(\lambda, N^{\text{MNO}}, N^{\text{Reg}})$. Both an analytical and a numerical approach have been attempted. The former drove us to a blind alley (which we include for completeness' sake) and the latter has been finally implemented in the software. Next we describe both of them.

A.1.1. Analytical approach

In this approach we first compute the integral $I_{N^{\text{MNO}},n}(N^{\text{Reg}})$ and then we sum up the series. We perform the change of variables $u = \frac{\alpha}{\alpha+\beta}, v = \alpha + \beta$ so that the integral transforms into

$$\begin{aligned} I_{n,m}(N^{\text{REG}}) &= \int_0^\infty dv f_v(v) \int_0^1 du\ f_u(u) \frac{B(u \cdot v + n, (1-u) \cdot v + m)}{B(u \cdot v, (1-u) \cdot v)} && \text{(A.2a)} \\ &= \int_0^\infty dv\ f_v(v) \frac{\Gamma(v)}{\Gamma(v+n+m)} \int_0^1 du\ f_u(u) \frac{\Gamma(u \cdot v + n)}{\Gamma(u \cdot v)} \frac{\Gamma((1-u) \cdot v + m)}{\Gamma((1-u) \cdot v)} \\ &= \int_0^\infty dv\ f_v(v) \frac{\Gamma(v)}{\Gamma(v+n+m)} \int_0^v dt\ f_u(t/v) \frac{\Gamma(t+n)}{\Gamma(t)} \frac{\Gamma(v-t+m)}{\Gamma(v-t)} && \text{(A.2b)} \end{aligned}$$

We pursue the analytical computation using expression (A.2b). The inner integral can be computed expressing the integrand in terms of Stirling numbers of the first kind (see e.g. Graham et al. (1996)):

$$\begin{aligned}\int_0^v dt f_u(t/v)\frac{\Gamma(t+n)}{\Gamma(t)}\frac{\Gamma(v-t+m)}{\Gamma(v-t)} &= \int_0^v dt f_u(t/v)\prod_{k=0}^{n-1}(t+k)\prod_{l=0}^{m-1}(v-t+l)\\ &= \int_0^v dt f_u(t/v)\sum_{k=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}t^k\sum_{l=0}^{m}\begin{bmatrix}m\\l\end{bmatrix}(v-t)^l\\ &= \sum_{k=0}^{n}\sum_{l=0}^{m}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}v^{k+l+1}\int_0^1 f_u(x)x^k(1-x)^l dx\\ &= \sum_{k=0}^{n}\sum_{l=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}\bar{B}(k+1,l+1)v^{k+l+1}, \qquad \text{(A.3)}\end{aligned}$$

where $\begin{bmatrix}n\\k\end{bmatrix}$ denotes the unsigned Stirling numbers of the first kind and $\bar{B}(k+1,l+1) = \int_0^1 f_u(x)x^k(1-x)^l dx$ (notice that for $f_u =$ Unif(0,1) the function $\bar{B}$ reduces to the beta function). Then we can write

$$I_{n,m}(N^{\text{REG}}) = \sum_{k=0}^{n}\sum_{l=0}^{m}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}\bar{B}(k+1,l+1)\int_0^\infty dv f_v(v;N^{\text{Reg}})\frac{\Gamma(v)}{\Gamma(v+n+m)}v^{k+l+1} \qquad \text{(A.4)}$$

Now denoting

$$J_{n+m,k+l}(N^{\text{REG}}) = \int_0^\infty dv \cdot f_v(v;N^{\text{REG}})\cdot\frac{v^{k+l}}{\prod_{i=1}^{n+m-1}(v+i)}, \qquad \text{(A.5)}$$

we have

$$\begin{aligned}I_{n,m}(N^{\text{REG}}) &= \sum_{k=0}^{n}\sum_{l=0}^{m}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}\bar{B}(k+1,l+1)J_{n+m,k+l}(N^{\text{REG}})\\ &= \sum_{p=0}^{n+m}J_{n+m,p}(N^{\text{REG}})\sum_{q=0}^{p}\begin{bmatrix}n\\q\end{bmatrix}\begin{bmatrix}m\\p-q\end{bmatrix}\bar{B}(q+1,p-q+1)\\ &= \sum_{p=0}^{n+m}J_{n+m,p}(N^{\text{REG}})\cdot a_{n,m}(p) \qquad \text{(A.6)}\end{aligned}$$

Thus we have reduced the integral to the computation of $J_{n+m,p}(N^{\text{REG}})$ and

$$a_{n,m}(p) = \sum_{q=0}^{p} \begin{bmatrix} n \\ q \end{bmatrix} \begin{bmatrix} m \\ p-q \end{bmatrix} \bar{B}(q+1, p-q+1). \tag{A.7}$$

The integral $J_{n+m,p}(N^{\text{Reg}})$ can be further computed analytically resorting to the residue theorem (see e.g. Brown and Churchill (2004)). Applying this theorem to $g(z) = f_N(z) \cdot \frac{z^p}{\prod_{k=1}^{n+m-1}(z+k)} \cdot \log(z)$ in the closed path around the origin composed by a straight path γ_1 along and above the positive real axis (from $+\epsilon$ to $+R$), a counterclockwise circular path γ_R at radius R, a straight path γ_2 along and below the positive real axis (from $+R$ to $+\epsilon$) and a clockwise circular path γ_ϵ at radius ϵ. We place a branch cut at the positive real axis. Then it is easy to prove (via Jordan's lemma) that $\int_{\gamma_R} g(z)dz \to 0$ and $\int_{\gamma_\epsilon} g(z)dz$ when $R \to \infty$ and $\epsilon \to 0$, while

$$\int_{\gamma_1} g(z)dz \to \int_0^\infty dx\ f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} \log(x), \tag{A.8}$$

$$\int_{\gamma_2} g(z)dz \to -\int_0^\infty dx\ f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} \left(\log(x) + 2\pi i\right). \tag{A.9}$$

The poles of $g(z)$ are simple and located at $z = -k$, $k = 1, \ldots, n+m-1$ and the residues can be computed easily:

$$\begin{aligned} \text{Res}(g, -k) &= f_N(-k) \cdot \frac{(-k)^p}{\prod_{\substack{i=1 \\ i \neq k}}^{n+m-1}(i-k)} \log\left(ke^{i\pi}\right) \\ &= f_N(-k) \frac{(-1)^p k^p}{(-1)^{k-1}(k-1)!(n+m-1-k)!} \left(\log(k) + i\pi\right) \\ &= f_N(-k) \cdot \frac{(-1)^{p-k} k^{p+1}}{(n+m-1)!} \binom{n+m-1}{k} \left(\log(k) + i\pi\right) \end{aligned} \tag{A.10}$$

Substituting on the residue theorem and focusing on the imaginary part of the expressions we have

$$-2\pi i \int_0^\infty dx \quad f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} = 2\pi i \sum_{k=1}^{n+m-1} f(-k) \cdot \frac{(-1)^{p-k} k^{p+1}}{(n+m-1)!} \binom{n+m-1}{k} \left(\log(k) + i\pi\right), \tag{A.11}$$

thus arriving at

$$J_{n+m,p}(N) = \int_0^\infty dx\ f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)}$$
$$= \frac{1}{(n+m-1)!} \sum_{k=1}^{n+m-1} (-1)^{p-k-1} \binom{n+m-1}{k} f(-k) \cdot k^{p+1} \log(k). \quad \text{(A.12)}$$

Setting $n = N_i^{\text{MNO}}$, $m = n_i - N_i^{\text{MNO}}$ and $N = N_i^{\text{REG}}$, we have

$$J_{n_i,p_i}(N_i^{\text{REG}}) = \frac{1}{(n_i-1)!} \sum_{k_i=1}^{n_i-1} (-1)^{p_i-k_i-1} \binom{n_i-1}{k_i} f_{N_i^{\text{REG}}}(-k_i) \cdot k_i^{p_i+1} \log(k_i). \quad \text{(A.13)}$$

In contrast, we have found no way to further simplify expression (A.7), not even using recursive relations between Stirling numbers. Moreover these numbers are computationally demanding.

In any case we arrive at a dead alley so that the numerical approach is clearly favoured.

A.1.2. Numerical approach

In this second approach we compute the integral using Monte Carlo techniques Robert and Casella (2004, 2010). We first sum up the series and then we compute the integral so that we write

$$\begin{aligned}
S(\lambda, N^{\text{MNO}}, N^{\text{REG}}) &= \int_0^\infty \int_0^\infty d\alpha d\beta\ \frac{f_u(\frac{\alpha}{\alpha+\beta}; N^{\text{REG}}) \cdot f_v(\alpha+\beta; N^{\text{REG}})}{\alpha+\beta} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \frac{B(\alpha+N^{\text{MNO}}, \beta+n)}{B(\alpha,\beta)} \\
&= \int_0^\infty \int_0^\infty d\alpha d\beta\ \frac{f_u(\frac{\alpha}{\alpha+\beta}; N^{\text{REG}}) \cdot f_v(\alpha+\beta; N^{\text{REG}})}{(\alpha+\beta)\cdot B(\alpha,\beta)} \int_0^1 dx\ x^{\beta-1}(1-x)^{\alpha+N^{\text{MNO}}-1} \sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} \\
&= \int_0^\infty \int_0^\infty d\alpha d\beta\ \frac{f_u(\frac{\alpha}{\alpha+\beta}; N^{\text{Reg}}) \cdot f_v(\alpha+\beta; N^{\text{REG}})}{(\alpha+\beta)\cdot B(\alpha,\beta)} \int_0^1 dx\ e^{\lambda x} x^{\beta-1}(1-x)^{\alpha+N^{\text{MNO}}-1} \\
&= \int_0^\infty \int_0^\infty d\alpha d\beta\ \frac{f_u(\frac{\alpha}{\alpha+\beta}; N^{\text{Reg}}) \cdot f_v(\alpha+\beta; N^{\text{Reg}})}{\alpha+\beta} \frac{B(\alpha+N^{\text{MNO}}, \beta)}{B(\alpha,\beta)} \cdot {}_1F_1(z; \beta, \alpha+\beta+N^{\text{MNO}}) \\
&\equiv \int_0^\infty \int_0^\infty d\alpha d\beta\ \frac{f_u(\frac{\alpha}{\alpha+\beta}; N^{\text{Reg}}, \mathbf{z}) \cdot f_v(\alpha+\beta; N^{\text{Reg}})}{\alpha+\beta} \Phi(\alpha, \beta; \lambda, N^{\text{MNO}}, N^{\text{REG}}),
\end{aligned} \quad \text{(A.14)}$$

where we have defined $\Phi(\alpha,\beta;\lambda,N^{\text{MNO}},N^{\text{REG}}) = \frac{B(\alpha+N^{\text{MNO}},\beta)}{B(\alpha,\beta)}\cdot {}_1F_1(\lambda;\beta,\alpha+\beta+N^{\text{MNO}})$ (${}_1F_1$ stands for the confluent hypergeometric function). Now to compute this integral let us change variables as in the preceding section so that

$$\begin{aligned} S(\lambda,N^{\text{MNO}},N^{\text{REG}}) &= \int_0^\infty dv\ f_v(v)\int_0^1 du\ f_u(u)\cdot\Phi(u\cdot v,(1-u)\cdot v;\lambda,N^{\text{MNO}},N^{\text{REG}}) \\ &= \int_0^\infty dv\ f_v(v)\int_0^1 du\ f_u(u)\cdot\bar{\Phi}(u,v;\lambda,N^{\text{MNO}},N^{\text{REG}}) \end{aligned} \tag{A.15}$$

where we have defined $\bar{\Phi}(u,v;\lambda,N^{\text{MNO}},N^{\text{REG}}) = \Phi(u\cdot v,(1-u)\cdot v;\lambda,N^{\text{MNO}},N^{Reg})$.

Denote $\bar{\Phi}(\mathbf{x}) = \bar{\Phi}(x_1,x_2;\lambda,N^{\text{MNO}},N^{\text{Reg}})$ and generate M bidimensional random variables $\mathbf{x}\in[0,1]\times\mathbf{R}^+$ according to the bidimensional distribution $f_u\times f_v$. Then, using $f(\mathbf{x}) = f_u(x_1)f_v(x_2)$ as importance function, we can write as a first option

$$S(\lambda,N^{\text{MNO}},N^{\text{REG}}) = \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\bar{\Phi}(\mathbf{x}_i). \tag{A.16}$$

To accelerate the convergence we make use of stratified importance sampling (Robert and Casella, 2004). To introduce the stratification let us define $H_1\cdot H_2$ strata as the rectangular domains $[a_{h_1-1},a_{h_1}]\times[b_{h_2-1},b_{h_2}]$, where $a_{h_1} = F_u^{-1}(h_1/H_1)$ $(h_1=1,\ldots,H_1)$ and $b_{h_2} = F_v^{-1}(h_2/H)$ $(h_2=1,\ldots,H_2)$, and F_i stands for the distribution function corresponding to the density function f_i. Defining the importance function in each stratum by $f_{h_1h_2} = H_1\cdot H_2\cdot f_u\cdot f_v$ truncated at $[a_{h_1-1},a_{h_1}]\times[b_{h_2-1},b_{h_2}]$ and taking equal-size strata $M_{h_1h_2} = \frac{M}{H_1H_2}$, then we finally write

$$S(\lambda,N^{\text{MNO}},N^{\text{REG}}) = \lim_{M\to\infty}\frac{1}{M}\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\sum_{i_{h_1}=1}^{M/H_2}\sum_{i_{h_2}=1}^{M/H_1}\bar{\Phi}(\mathbf{x}_{i_{h_1}i_{h_2}}) \tag{A.17}$$

where the random values $\mathbf{x}_{i_{h_1}i_{h_2}}$ are generated with the corresponding density function $f_{h_1h_2}$. Expression (A.17) provides the basis to compute the function $S(\lambda,N^{\text{MNO}},N^{\text{Reg}})$.

Finally the computation of the unnormalized density function $f(\lambda|N^{\text{MNO}};N^{\text{Reg}})$ is completed multiplying by standard functions $f(\lambda)$ (prior density of λ) and

$$\text{Po}(N^{\text{MNO}};\lambda) = e^{-\lambda}\cdot\frac{\lambda^{N^{\text{MNO}}}}{N^{\text{MNO}}!}.$$

In figure A.1 we represent $f(\lambda|N^{\text{MNO}};N^{\text{Reg}})$ for a combination of values of N^{MNO} and N^{Reg} with uniform prior distributions for u and v with an interval length of $\pm15\%$ and gamma prior distribution for λ with a mode in N^{Reg} and a coefficient of variation of

0.40.

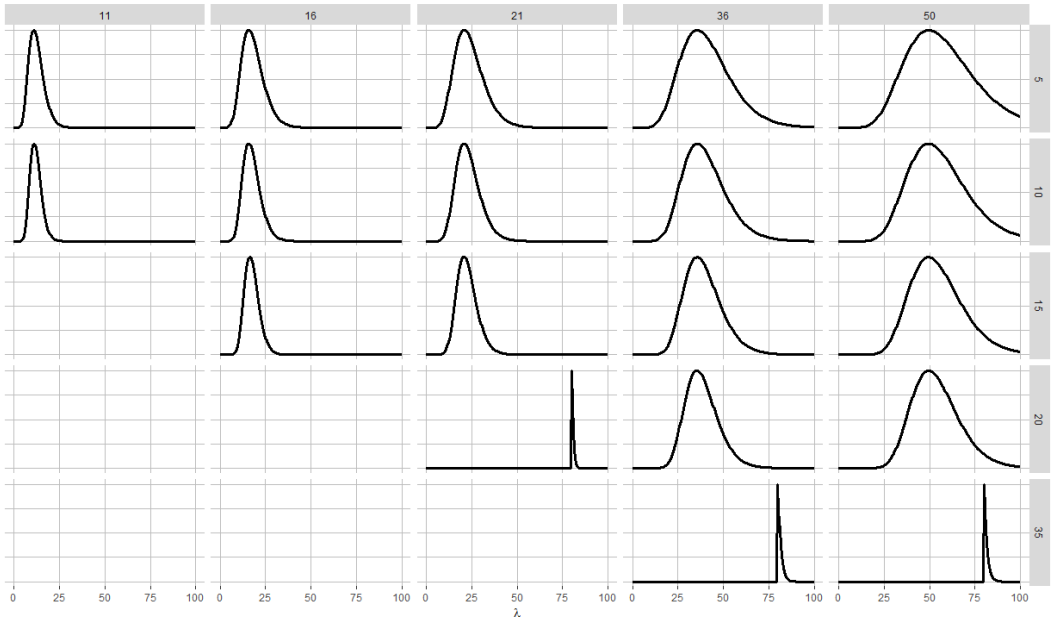


Figure A.1 Posterior density for λ for several values of N^{MNO} (vertical) and N^{Reg} (horizontal). Modes are normalize to 1 for a better joint visualization.

Notice that the Monte Carlo technique used here does not exhaust the possibilities to numerically compute the involved integral. Quadrature methods have not been explored and remains as an alternative option.

A.2. Sampling from the posterior distribution of λ

To conduct simulation studies and carry out the estimation on the number of individuals per cell we need to generate random variables according to the posterior distribution $f\left(\lambda|N^{\text{MNO}}; N^{\text{Reg}}\right)$. This does not allow us to find easily the corresponding posterior distribution function to apply the inverse method to generate random variables (see e.g. Devroye (1986)). We have chosen the acceptance-rejection method (Robert and Casella, 2004). Indeed this method is appropriate to use with unnormalized probability functions.

As a first candidate distribution $g(\lambda)$ we have focused on the Cauchy distribution $g(\lambda) = \text{Cauchy}(\lambda; \lambda_0 = \lambda^*, \sigma)$ truncated at $\mathbb{R}^+$ with $\lambda^* = \text{argmax}_{\lambda \geq 0} f(\lambda|N^{\text{MNO}}; N^{\text{Reg}})$ (i.e. the mode of $f(\lambda|N^{\text{MNO}}; N^{\text{Reg}})$). For rigor's sake we need to prove that f is majorized

by this candidate distribution g for appropriate values of σ. We have prioritized computing tests which have been for the time being satisfactory (see however conclusions and proposals in chapter 5). Also we do not have a general recipe for the scale parameter σ.

Next we must find $c \in \mathbb{R}$ such that

$$\inf_{\lambda \geq 0} \frac{c \cdot g(\lambda)}{f(\lambda | N^{\text{MNO}}; N^{\text{Reg}})} \geq 1. \tag{A.18}$$

Taking the minimal c for sampling efficiency reasons we have

$$c = \sup_{\lambda \geq 0} \frac{f(\lambda | N^{\text{MNO}}; N^{\text{Reg}})}{g(\lambda)}.$$

To generate random values λ according to $f\left(\lambda | N^{\text{MNO}}; N^{\text{Reg}}\right)$ we generate values according to $g(\lambda)$, and values v according to $\text{Unif}(0, 1)$ so that we accept those λ such that $v \leq \frac{f(\lambda | N^{\text{MNO}}; N^{\text{Reg}})}{c \cdot g(\lambda)}$.

In figure A.2 we illustrate the construction of a candidate distribution with location in λ^* and scale as $\sigma = \sqrt{(1+\alpha)} \cdot \frac{N^{\text{Reg}}}{\alpha}$.

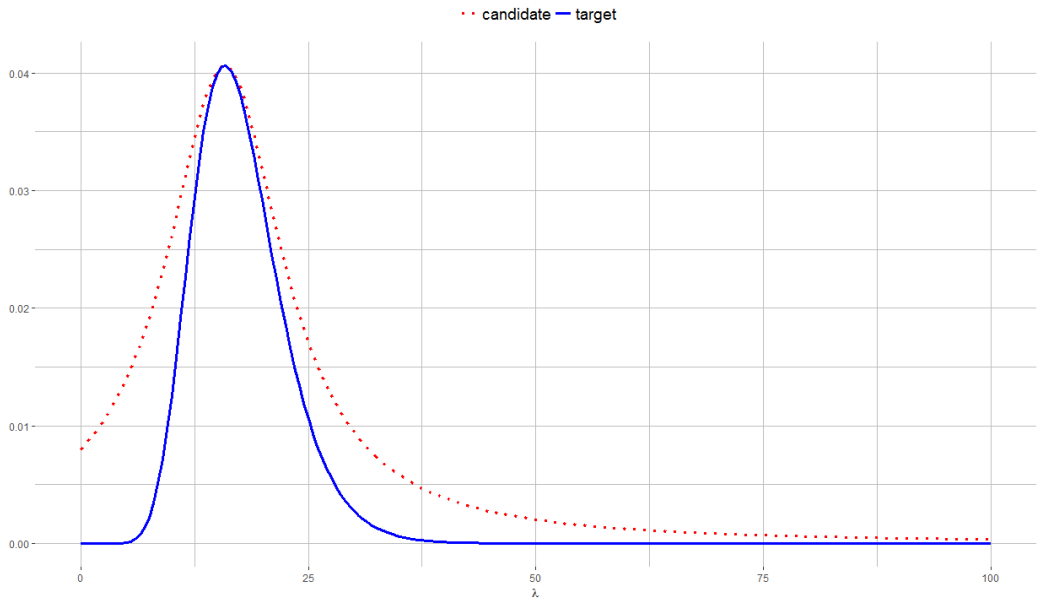


Figure A.2 Candidate distribution for the rejection method to sample values from the posterior $f(\lambda | N^{\text{MNO}}; N^{\text{Reg}})$.

A.3. Sampling from the posterior distribution of N

To sample from the posterior distribution of N we make use of the hierarchical model itself. In particular we use the specification (4.1g) to generate each value N from the corresponding parameter λ generated in the step above. This is elementary. In figure A.3 we illustrate the generation of 1000 values according to this posterior distribution.

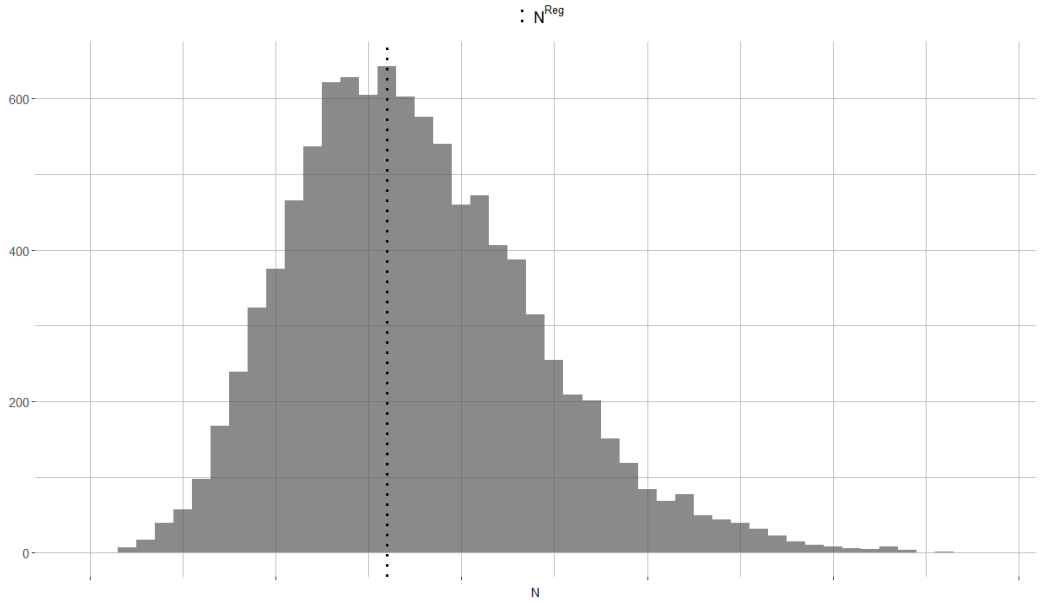


Figure A.3 10000 values generated according to the posterior distribution of N.

Bibliography

American Planning Association (2018). Land based classification standards. `https://www.planning.org/lbcs/`.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2015). *Hierarchical modelling and analysis of spatial data (2nd ed)*. CRC Press.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion), in V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference*, pp. 203–242. Holt, Reinhart and Winston.

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. Wiley.

Brown, J. and R. Churchill (2004). *Complex variables and applications (8th ed.)*. McGraw-Hill.

Calabrese, F., L. Ferrari, and V. D. Blondel (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys 47*, 25:1-25:20.

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

Cassel, C.-M., C.-E. Särndal, and J. Wretman (1977). *Foundations of Inference in Survey Sampling*. Wiley.

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). Wiley.

Deming, W. (1950). *Some theory of sampling*. Wiley.

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.

Devroye, L. (1986). *Non-uniform random variable generation*. Springer.

Doyle, J., P. Hung, R. Farrell, and S. Mcloone (2014). Population mobility dynamics estimated from mobile telephony data. *Journal of Urban Technology 21*, 109–132.

EPSG (2018). epsg.io – Coordinate Systems Worldwide. `https://epsg.io/28992`.

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`.

Eurostat (2014). Methodological manual for tourism statistics (v3.1). `http://ec.europa.eu/eurostat/documents/3859598/6454997/KS-GQ-14-013-EN-N.pdf`.

Eurostat, NIT, University of Tartu, Statistics Estonia, Positium, IFSTTAT, and Statistics Finland (2014). Feasibility study on the use of mobile poistioning data for tourism statistics. `http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies`.

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

Graham, R., D. Knuth, and O. Patashnik (1996). *Concrete Mathematics (2nd ed.)*. Addison-Wesley.

Grimmet, G. and D. Stirzaker (2004). *Probability and random processes* (3rd ed.). Oxford Science Publications.

Groves, R. (1989). *Survey errors and survey costs*. Wiley.

Hájek, J. (1981). *Sampling from a finite population*. Marcel Dekker Inc.

Hansen, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science 2*, 180–190.

Hansen, M., W. Hurwitz, and W. Madow (1966). *Sample survey: methods and theory* (7th ed.). Wiley.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Hedayat, A. and B. Sinha (1991). *Design and Inference in Finite Population Sampling*. Wiley.

ISO (2004). ISO 8601:2004. `https://www.iso.org/standard/40874.html`.

ISO (2007). ISO 19111:2007. `https://www.iso.org/standard/41126.html`.

Kruskal, W. and F. Mosteller (1979a). Representative sampling, i: Non-scientific literature. *International Statistical Review 47*, 13–24.

Kruskal, W. and F. Mosteller (1979b). Representative sampling, ii: scientific literature, excluding statistics. *International Statistical Review 47*, 111–127.

Kruskal, W. and F. Mosteller (1979c). Representative sampling, iii: the current statistical literature. *International Statistical Review 47*, 245–265.

Kruskal, W. and F. Mosteller (1980). Representative sampling, iv: The history of the concept in statistics. *International Statistical Review 48*, 169–195.

Lehtonen, R. and A. Veijanen (1998). Logistic generalized regression estimators. *Survey Methodology 24*, 51–55.

Lessler, J. and W. Kalsbeek (1992). *Nonsampling error in surveys*. Wiley.

Little, R. (2012). Calibrated bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics 28*, 309–334.

Manly, B. and J. e. Navarro-Alberto (2014). *Introduction to ecological sampling*. CRC Press.

Meersman, F. D., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, and H. Reuter (2016). Assessing the quality of mobile phone data as a source of statistics. *Q2016 Conference paper, June 2016*.

Nemhauser, G. and L. Wolsey (1999). *Integer and combinatorial optimization*. Addison-Wesley.

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

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*, 558–625.

Okabe, A., B. Boots, K. Sugihara, and S.-N. Chiu (2000). *Spatial tessellations: concepts and applications of Voronoi diagrams (2nd ed)*. Wiley.

Positium (2016). Technical documentation for required raw data from mobile network operators for official statistics. Technical report, Positium.

Positium (2017). Common plan for methodology and data processing of mobile phone data from mobile network operators for official statistics. Technical report, Positium.

Positium (2018). `https://www.positium.com/`.

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

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

Ricciato, F., P. Widhalm, F. Pantisano, and M. Craglia (2017). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing 35*, 65–82.

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.

Royle, J. and R. Dorazio (2014). *Hierarchical modeling and inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*, 99–119.

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

Seynaeve, G., C. Demunter, F. D. Meersman, Y. Baeyens, M. Debusschere, P. Dewitte, P. Lusyne, F. Reis, H. Reuter, and A. Wirthmann (2016). When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics. *14th Global Forum on Tourism Statistics (Venice, Italy, Nov. 2016)*.

Smith, T.M.F. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society A 139*, 183–204.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley.

Open Street Map Foundation. `https://www.openstreetmap.org`.

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

JAGS (2018). `http://mcmc-jags.sourceforge.net/`.

Stan (2018). Stan. `http://mc-stan.org/`.

Thompson, S. (2012). *Sampling*. Wiley.

UNECE (2013). Generic Statistical Business Process Model v5.0. https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

UNECE (2016). Generic Statistical Data Editing Models. https://statswiki.unece.org/display/VSH/GSDEMs.

Valliant, R., A. Dorfmann, and R. Royall (2000). *Finite population sampling and inference. A prediction approach*. Wiley.

Vanhoof, M., F. Reis, T. Ploetz, and Z. Smoreda (2018). Assessing the quality of home detection from mobile phone data for Official Statistics. *Journal of Official Statistics*. In press.

Wilysis (2018). Network Cell Info Lite app. https://play.google.com/store/apps/details?id=com.wilysis.cellinfolite&hl=en_419.

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 (2017). Deliverable 5.2. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.2.pdf.

Yates, F. (1965). *Sampling methods for censuses and surveys* (3rd ed.). Charles Griffins.

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