Salgado et al. *EPJ Data Science* (2021) 10:20
https://doi.org/10.1140/epjds/s13688-021-00275-w

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE

Open Access

An end-to-end statistical process with mobile network data for official statistics

David Salgado[1,2*], Luis Sanguiao[1], Bogdan Oancea[3], Sandra Barragán[1] and Marian Necula[4]

The views expressed in this working paper are those of the authors and do not necessarily reflect the views of their affiliating institutions.

*Correspondence:
david.salgado.fernandez@ine.es
[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Av. de Manoteras, 50-52, Madrid, Spain
[2]Dept. Statistics and Operations Research, Complutense University of Madrid, Plaza de las Ciencias, 3, Madrid, Spain
Full list of author information is available at the end of the article

Abstract

Mobile network data has been proven to provide a rich source of information in multiple statistical domains such as demography, tourism, urban planning, etc. However, the incorporation of this data source to the routinely production of official statistics is taking many efforts since a diversity of highly entangled issues (access, methodology, IT tools, quality, skills) must be solved beforehand. To do this, one-off studies with concrete data sets are not enough and a standard statistical production process must be put in place. We propose a concrete modular process structured into evolvable modules detaching the strongly technological layer underlying this data source from the necessary statistical analysis producing outputs of interest. This architecture follows the principles of the so-called ESS Reference Methodological Framework for Mobile Network Data. Each of these modules deals with a different aspect of this data source. We apply hidden Markov models for the geolocation of mobile devices, use a Bayesian approach on this model to disambiguate devices belonging to the same individual, compute aggregate numbers of individuals detected by a telecommunication network using probability theory, and model hierarchically the integration of auxiliary information from the telco market and official data to produce final estimates of the number of individuals across different territorial regions in the target population. A first simple illustrative proposal has been applied to synthetic data providing preliminary software tools and accuracy indicators monitoring the performance of the process. Currently, this exercise has been applied to the estimation of present population and origin-destination matrices. We present an illustrative example of the execution of these production modules comparing results with the simulated ground truth, thus assessing the performance of each production module.

Keywords: Mobile network data; Production framework; Official statistics; Statistical production process

1 Introduction

Mobile network data, i.e. digital data generated in a mobile telecommunication network by the interaction between a mobile station (device) and a base transceiver station (antenna, in loose terms) [1], constitutes a rich source of information for Social Science, in

Springer

general, and for Official Statistics, in particular. There already exist multiple excellent examples of one-off applications [2–15] (see supplementary material 1 for a more comprehensive list of references), but the production of official statistics in National Statistical Systems demands a fully-fledged production framework covering different aspects such as access conditions, methodological and quality frameworks, IT infrastructure (both hardware and software), statistical disclosure control, and identification of relevant indicators for a diversity of statistical domains in National and International Statistical Plans, mostly included as part of legal regulations. A number of illustrative case studies of mobile network data to the production of official statistics can already be found in the literature [16–26]. Moreover, efforts are under way to construct a production framework [27–29] with some recent examples of an end-to-end statistical production process [30]. The need for a process-oriented production system instead of a product-oriented or even domain-oriented system is well-known in Official Statistics, where important initiatives have been carried out in the last decade to avoid so-called stove pipe models driving National Statistical Offices (NSOs) to production in silos, models which reduce the cost-efficiency to the point of endangering the future feasibility of the production of official statistics [31].

There exist two important issues which raise immediate rightful concerns when using mobile network data for statistical purposes. These are (i) privacy and confidentiality of network subscribers and (ii) access conditions to data by NSOs. We shall not be dealing with these issues in the next sections, but we mention the general principles for the context in which our proposed process is to be considered. Privacy and confidentiality of any statistical information collected, processed, and disseminated by NSOs have been, are, and will be a priority for any kind of data source. Traditional survey data is indeed identified personal data and concerns about its protection are duly accounted for with a specific production phase known as statistical disclosure control [32, 33]. All kind of survey and administrative data about personal habits, causes of death, business revenues, VAT and personal taxes, etc. are collected, processed, and aggregated and official statistics are disseminated under a negligible risk of reidentification of statistical units, whatever their nature is. Not only is this commitment present with new digital data sources in general and mobile network data in particular, but is it also reinforced.

Regarding access, this is an intricately complex unsolved issue where many, many facets need to be considered simultaneously. Currently, there exist concrete agreements between some NSOs/research centres/universities and Mobile Network Operators (MNOs) for research on limited data sets, but the conditions for routinely production of official statistics are yet to be found. By and large, in our view, MNOs will need to become an active part of the official statistical production process and this brings novel challenges. We identify at least the following restrictions to be jointly satisfied to arrive at a feasible solution. Firstly, security, confidentiality, and privacy must be legally and technically assured during the whole process, involving the approval by the national Data Protection Authorities. In this sense, we underline the traditional role of NSOs in collecting and processing sensitive information. Currently, we consider that any kind of mobile network data processing must be undertaken in the original information systems of MNOs. However, notice that further research needs to be conducted. For example, there exists both theoretical and empirical evidence [34, 35] that privacy is not preserved even after aggregating data under certain conditions. Secondly, appropriate territorial and time breakdowns for target indicators and aggregates for the social good, potentially to be included in sectorial legal

regulations, must be identified so that valuable information for data-based policy making and decision taking can be produced and disseminated for free. Thus, the relevant role of statistical offices in society according to the Fundamental Principles of Official Statistics [36] would be strengthened. Thirdly, a new branch of economic activity is growing on the basis of digital data and data analytics [37]. This is usually substantiated in the so-called monetization of data generated by enterprises during their business activities. MNOs are not an exception and due to the technologically complex data ecosystem of telecommunication networks, investments are needed (mobile network data for statistical purposes do not exist, a preprocessing stage is needed). Thus, a trade-off between public and private interests must be found. In this line of thoughts, as we have expressed elsewhere [38], public-private partnerships arise as an optimal solution, in which win-win agreements are indeed feasible. The present methodological proposal, beyond the statistical contents included hereafter, provides also an insight on aspects to be taken into account when finding these agreements.

To our best knowledge, mobile network data can be used at least in three (complementary) ways, namely (i) focusing on geolocation of network events to analyse population counts, displacement patterns, and mobility-related phenomena in general (see most references above), (ii) focusing on the type of applications generating the Internet traffic from the devices (see e.g. [39]), and (iii) investigating interactions between devices to analyse different aspects of social networks [40]. In the following, we shall focus only on the geolocation of network events.

We make a proposal for an end-to-end statistical process going from the raw telco data generated at the mobile telecommunication networks to the final target population count estimates. The proposal follows the principles of functional modularity adapted to statistical production [41] focusing on input and output data as well as the throughput of each production step. The proposal so far focuses on a single-MNO scenario. The next sections describe each of the functional modules of the statistical process. In Sect. 2 we provide a description of the (synthetic) data used to illustrate the proposal. In Sect. 3 we describe the module to geolocate mobile devices. In Sect. 4 we propose a method to disambiguate devices carried by the same individual. In Sect. 5 we include general considerations to identify devices pertaining to the target population under analysis. In Sect. 6 we suggest a method to aggregate data from the device level to the territorial unit level. In Sect. 7 we propose to use hierarchical modelling to infer population counts in the target population from the population counts in the network, integrating at the same time auxiliary information. In Sect. 8 we integrate all modules in a production chain. Finally, in Sect. 9 we close with some conclusions and future prospects.

It is important to underline that the proposal is formulated with a priority on modularity and evolvability so that continuous improvements can be introduced adapting to concrete restrictions from actual production conditions. The statistical methods illustrating each module are not intended to be closed and definitive, but rather on the contrary to pave the way for more complex scenarios.

2 Data description

Our strategy to build a production framework revolves around the use of synthetic network event data. Our choice is motivated by the following reasons: (i) to have actual ground truth figures allowing us to conduct a thorough performance assessment of methods and parameters and a better understanding by comparison between actual population

counts and their estimates, (ii) to identify different concrete aspects of the problem by configuring different scenarios in order to propose specific elements in the methodology to deal with them, (iii) to avoid the issue about the access to real data (see above) and its consequences (lack of data, confidentiality and privacy risks, legal concerns,…), and (iv) to provide a body of technical knowledge to reach informed partnership agreements with MNOs (otherwise, how do we know what to agree upon?). Real data cannot provide these conditions for research.

In this line, we have developed a network event data simulator. The simulator is a highly modular software [42] implementing agent-based simulating scenarios with different elements configured by the user. The basic elements are:

- a geographical territory represented by a map;
- a population of individuals carrying 0, 1, or 2 mobile devices during their displacement;
- several choices of displacement patterns for the individuals;
- a telecommunication network configuration and connection mechanisms between devices and base transceiver stations (BTS) in terms of a radio wave propagation model;
- a reference grid for analysis.

The simulator works essentially by using a radio wave propagation model to simulate the connection mechanism between the BTSs and each mobile device during the displacement of each individual. The connection mechanism is an extreme simplification of the real world extracting the essential features for statistical analysis. The core output data consists of a time sequence of radio cell IDs and event codes (connection, disconnection, etc.) for each device along the duration of the simulation. Signalling data (i.e. passive data not depending on subscribers' behaviour) are simulated instead of Call Detail Records or any other active data generated by individuals (call, SMS, Internet connections, …). Notice, however, that the simulator is highly configurable and the user can adjust input parameters and impose data patterns or empirical distributions on the outputs to simulate different conditions according to his/her needs.

Before starting to develop an entirely new tool for synthetic data generation we made an inventory of the existing tools in this area, checking if there is one that can be used for our purposes. The *cdr-gen* project [43] is a very simple Call Detail Record (CDR) generator written in Java that allows the user to configure up to a certain extent the parameters of the calls (duration distribution, type of call, etc.) but has no support for defining the geographical coordinates of the mobile devices, the movement of the people carrying mobile devices or the parameters of the network. Another CDR generator [44] written in Scala allows users to generate CDR data with different models or with a mix of models. A simulation implies several steps: generate the cells, the mobile operator, the users, the social network of users and eventually generate the interaction between users. However, the capabilities to run complex simulations are not present, the cells of the mobile network are generated randomly with a fixed shape. There is no support to define our own maps and the (at least some) technical parameters of the BTS. NetSim [45] is a software that enables users to simulate a network comprising of devices and links, and study the behavior of this network. While this is a complex software that includes a user-friendly GUI and capabilities to simulate several real mobile network communications protocols, it is a commercial product with a limited version for academic institutions. An important drawback of

this simulation software comes from the fact that it is oriented on producing data needed for mobile telecommunication network optimizations and it does not provide the kind of data that matches our statistical needs. Another network simulation software that we've tested is OPNET Network Simulator [46]. Besides being a commercial software, it is also oriented on producing data needed for mobile telecommunication network optimization and it does no output the information that we need to produce population estimates. The traffic simulation packages SUMO [47] and MATSim [48] are more similar to our needs of modeling the population mobility but unfortunately they don't have any support for mobile devices and networks. Considering the minimal set of features that we defined for our synthetic data simulator we couldn't find any pre-existing software to entirely fulfil our needs, and we proceeded to develop our own simulation software.

For the time being, since our priority is the simulator as a whole, the different elements implemented so far are kept as simple as possible. Firstly, regarding the population of individuals, displacement patterns[1] are basically a sequence of stays (no movement) and random walks with/without a drift with two possible speeds (namely, walk and car speeds). The drift, the speeds, and the shares of individuals with 0, 1, and 2 devices are easily configured by the user. Only closed populations can be simulated so far, i.e. individuals cannot abandon or enter into the territory under analysis. Secondly, radio wave propagation models [49] are mathematical representations of the electromagnetic interaction between mobile stations and base transceiver stations in a telecommunication network which simplifies planning, configuration, and management avoiding numerical solutions of Maxwell's equations with real world complex boundary conditions. These models are used in the simulator to reproduce the BTS-device connections. We are using two very simple models for the connection mechanism. For omnidirectional radio cells:

- We model the so-called Received Signal Strength (RSS) for a device at a distance r from the radio cell as

$$\mathrm{RSS}(r) = 30 + 10 \cdot \log_{10}(P) - 10 \cdot \gamma \cdot \log_{10}(r), \tag{1}$$

  where P stands for the radio cell emission power (in Watts) and γ is the so-called path loss exponent (or attenuation factor). Notice that RSS is provided in dBm. Each device connects to the BTS producing the highest signal strength in each tile until the radio cell reaches its maximum capacity. Both the emission power and the path loss are selected as input parameters by the user.
- In agreement with Tennekes *et al.* [30], we further model a so-called Signal Dominance Measure (SDM) by making a logistic transformation on the RSS:

$$\mathrm{SDM}(r) = \frac{1}{1 + \exp(-S_{\mathrm{steep}} \cdot (\mathrm{RSS}(r) - S_{\mathrm{mid}}))}, \tag{2}$$

  where S_{steep} and S_{mid} are chosen according to characteristics of each radio cell. Each device connects to the BTS providing the highest signal dominance measure in each tile until the radio cell reaches its maximum capacity. Both S_{steep} and S_{mid} are selected as input parameters by the user, too.

[1] During the review process of the present work, Lévy flights have been also included as a displacement pattern.

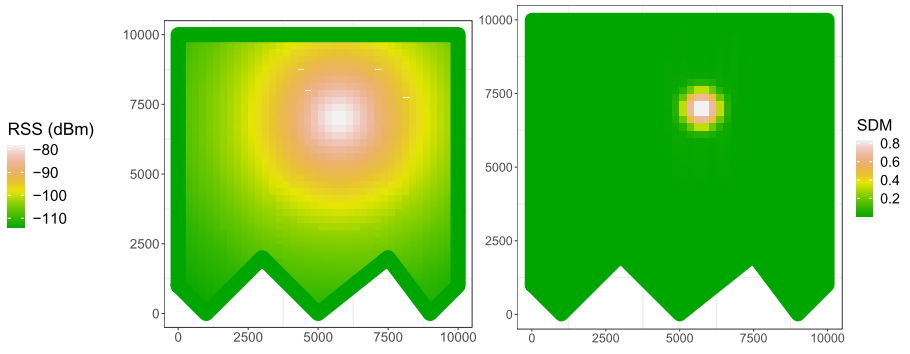

Figure 1 Received signal strength and signal dominance measure. Received signal strength and signal dominance measure for an omnidirectional radio cell according to models (1) and (2)

In both cases, minimal thresholds for both RSS and SDM are selected by the user below which no connection is possible. Coverage areas are indeed computed in this simple way. See Fig. 1 for an illustrative example of the RSS and SDM of a given radio cell.

For directional radio cells, more parameters are needed (see [30]). For simplicity, we shall use only omnidirectional radio cells in this work. The inclusion of directional radio cells would only increase the computational effort and parameter specifications by the user, not introducing conceptual novelties.

For the next sections to illustrate our proposed production model, we have configured a scenario over an irregular polygon with a bounding box of 10 km × 10 km, across which $N = 500$ individuals move according to a sequence of stays and random walks with a drift, 186 of them carrying at least one device (32 of them carrying two devices). We have configured 67 omnidirectional radio cells and 3 directional radio cells at the same site. See Fig. 2 and animated gif `individuals.gif` in [50]. Parameters are further specified in the supplementary material. To simulate the BTS-device connections, the RSS model has been used in the simulation.

3 Geolocation of mobile devices

3.1 Model specification and construction

The ultimate goal of the proposed set of modules is to provide common production steps valid for any statistical domain detaching the highly technological substratum of this data source from the statistical analysis producing different outputs and insights. This first module focuses on the geolocation information in the telecommunication network about mobile devices. There already exist multiple techniques to geolocate a mobile station in a radio telecommunication network [51–57], but they focus on providing a high-quality telecommunication service. Instead, we focus on statistical purposes and many of these computationally demanding techniques are not necessary. Our design is based on the following premises. First, following [28], the design should be as much modular as possible so that the geolocation information for statistical analyses is not directly affected by changes in the telecommunication technology. At the same time, the design should allow the module to evolve according to this technology. Second, we shall use data generated in the network and shall not access data generated in the mobile devices. Indeed, we shall use only the minimal set of information needed for the production of official statistics. Much research is needed to agree on this minimal data set depending on case studies and

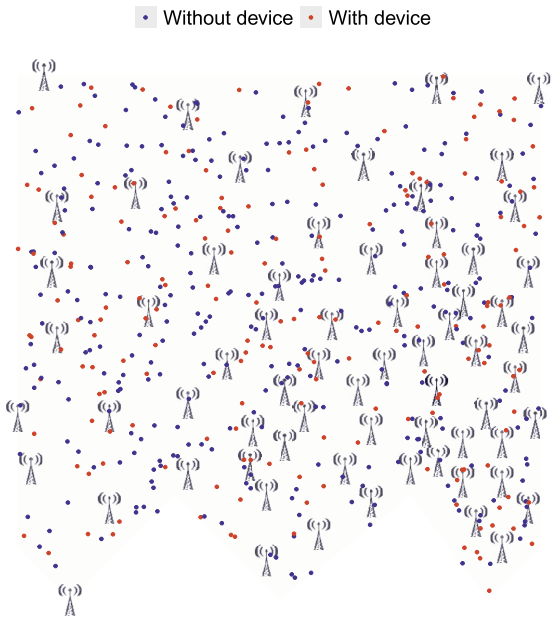

Figure 2 Network configuration. Positions of 70 BTS (67 omnidirectional radio cells –in grey– and 3 directional radio cells – in black), positions of individuals at time $t = 0$. (See complementary online material for an animated gif for the displacement of individuals [50]

simulation exercises. Basically, we focus on the digital trace left by mobile devices in the network and not on applications actively generating data for this purpose. Third, quality is a concern of first priority in the production of official statistics. In this sense, we shall account for the uncertainty underlying the whole production process so that estimates will be produced together with accuracy indicators. Fourth, the design of modules should allow us to integrate multiple data sources such as information from the telco market (penetration rates, market shares, etc.) and from Official Statistics (register-based residential population figures, land use, etc.).

Let us illustrate these premises with a concrete example. Let us think of the evolution from 3G technology to 4G technology. The modularity will be introduced by using a reference grid dividing the geographical territory of analysis into tiles and providing the probability for each device to be geolocated at each tile. Data abstraction is implemented just through the statistical model providing these location probabilities: we get location probabilities independently of the underlying technology. Indeed, when this technology evolves (from 3G to 4G), the statistical model computing the probabilities may be made more sophisticated including more variables or more accurate data, but at the end we still have location probabilities. Available data can be just the radio cell IDs of each connection or can be completed using other variables such as Timing Advance, Angle of Arrival, etc. Furthermore, we can naturally account for uncertainty in the geolocation information since we have probability distributions. Indeed, the use of probability models will allow us to integrate in a natural way information from auxiliary data sources.

Now, we formalise our approach. We begin by introducing the input data. We shall denote by $\mathbf{E}_d(t)$ the set of network event variables regarding mobile device d at time instant t. These may be the radio cell ID, the Timing Advance (TA), the Angle of Arrival (AoA), … or any network variable reflecting the digital trace of mobile device d at time t. Notice that

these are telco variables which will certainly evolve and change according to the telecommunication technology. Also, notice that these contain sensitive information about each device (hence about each individual) and thus must not leave the information systems of MNOs (in-situ processing). NSOs do not need access to these variables, only to the design of their processing. Next, we shall denote by $\boldsymbol{\theta}^{\text{net}}$ the parameters for the radiowave propagation model such as the emission power, the path loss exponent, etc. (see models (1) and (2) above). Although these parameters do not contain sensitive information about the subscribers, they reveal important technological information in the competitive telecommunication market. NSOs do not need access to these variables either, but the models must be jointly agreed with MNOs. Finally, we shall denote by $\mathbf{I}^{\text{aux}}$ any auxiliary information about the geographical territory such as the land use or transport networks or any other external data source such as a population register. This information is indeed public, but it may also incorporate data at the micro level produced (and not disseminated) by NSOs.

The displacement of devices across the geographical territory bears an evident dynamical ingredient in which we have access to a set of observed variables (network variables $\mathbf{E}_d(t)$) and a set of unobserved variables (location at each tile i, which we shall denote by $T_{dt} = i, i = 1, \ldots, N_T$). A natural mathematical description of this situation can be provided using hidden Markov models (HMMs) [58, 59], in which we model the time sequence of hidden (unobserved) variables $\mathbf{S}_{dt}$ for each device d at each time instant t and a time sequence of observed variables $\mathbf{O}_{dt}$, which in our case will be the network variables $\mathbf{O}_{dt} = \mathbf{E}_{dt}$. For simplicity, we shall assume that the state variables $\mathbf{S}_{dt}$ reduce to the tile location T_{dt} (see left panel of Fig. 3).

Now, we need two models:

- A transition model, providing details about the evolution (displacement) of the devices:

$$\mathbb{P}\left(T_{dt} = j | T_{dt-1} = i, \mathbf{I}^{\text{aux}}\right) \equiv a_{ij}. \tag{3}$$

- An emission model, providing details about the generation of network variables:

$$\mathbb{P}\left(\mathbf{E}_{dt} = \mathbf{E}_k | T_{dt} = i, \mathbf{I}^{\text{aux}}\right) \equiv b_i(\mathbf{E}_k) \tag{4}$$

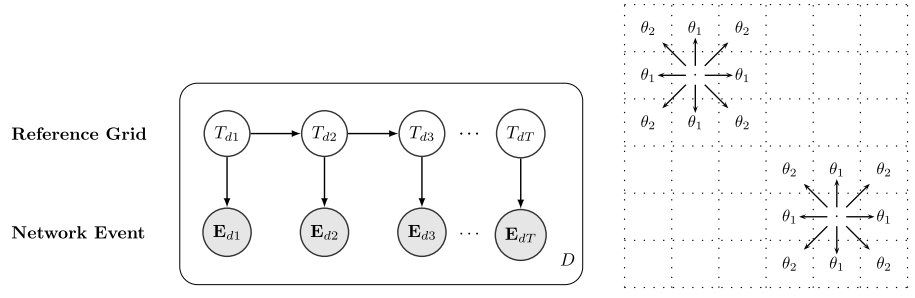

Figure 3 Graphical representation of a hidden Markov model. (Left) Observed variables are network variables $\mathbf{E}_{dt}$ and unobserved variables are tile locations T_{dt}. (Right) Transition probabilities in the reference grid for adjacent cells

For the transition model we make a fairly generic proposal not imposing a displacement pattern on the devices. We propose to choose the time regime in such a way as to have a one-tile-long displacement at most at each time instant t. Transition probabilities θ_1 and θ_2 between tiles are estimated maximising the likelihood for each device d (see right panel of Fig. 3).

To detach the technological and statistical layers we propose to substantiate the emission model (4) as a radio wave propagation model independent of the transition model so that $b_i(\mathbf{E}_k)$ is computed in terms of models (1) or (2) taking the centre of the tile as the reference point for the distance r. Notice that the emission model involves the network configuration parameters $\boldsymbol{\theta}^{\text{net}}$ (emission power, path loss exponent, S_{mid}, S_{steep} in our simple case). Notice diverse relevant points. Firstly, should we have richer raw telco data to consider more complex radio propagation models, we could immediately improve the accuracy with a more sophisticated computation of the emission probabilities. In case of lacking data for these models, we could resort to geometrical considerations as with the Voronoi tessellation. The ideal recommendation is to work together with MNOs to identify the more feasible data set for the computation of these likelihoods. Ultimately, this will also depend on the chosen final accuracy in our estimates. Secondly, a cautious reader may rapidly suggest that the emission probabilities can also be modelled in terms of unknown parameters to be estimated later on. In theory, this is always possible as in many other applications of HMMs. However, in our case we suggest to deal with the emission probabilities independently as a separate (sub)module in the whole process allowing us to detach the more technological stages directly dependent on raw telco data from the more statistical upper layers involving population count estimation. In this way, the joint work by MNOs and NSOs around the sensitive telco data is focused on this step paving the way for the functional modularity of the statistical process thus providing a concrete proposal for the implementation of the ESS RMF. Thirdly, the computational cost of the emission probabilities is fixed in time. If N_A denotes the number of radio cells in the geographical territory under analysis and the grid size is N_T, at most we need to compute $N_T \times N_A$ emission probabilities to conform the matrix $B = [b_{ik}]$, $i = 1, \ldots, N_T$, $k = 1, \ldots, N_A$. This is done once and for all t (assuming time homogeneity). Fourthly, notice that having the numerical values of the emission probabilities will allow us to simplify the computation of the likelihood for the HMMs reducing its parameter dependency only to the transition model. Finally, if missing values are to be used according to the time padding procedure described in the supplementary material (which guarantees the maximum one-tile distance restriction), *for numerical convenience later on* the corresponding emission probabilities can be conveniently set to 1, i.e. $b_{i0} = \mathbb{P}(E_{t_n} = \cdot | T_{t_n} = i, \mathbf{I}^{\text{aux}}) = 1$. This will greatly facilitate the expression of the HMM likelihood and its further optimization. Remind that this probability is not real and completely meaningless.

Lastly, the initial state (prior) distribution $\pi_i \equiv \mathbb{P}(T_{d0} = i | \mathbf{I}^{\text{aux}})$ is provided by the statistician. Currently, we consider either a noninformative uniform distribution ($\pi_i \propto 1$) or a so-called network distribution (based on the network configuration, e.g. $\pi_i \propto \text{RSS}_i$).

Once a model is fitted for each device, we can use the forward-backward algorithm [59] to compute the (posterior) location probabilities $\gamma_{dti} \equiv \mathbb{P}(T_{dt} = i | \mathbf{E}_{d1:T}, \mathbf{I}^{\text{aux}})$, i.e. the location probability at each tile i and each time instant t conditional on all the network and event information available for device d (see Fig. 4 for the location probabilities at time $t = 0$ and animated gifs `postLocLayer*.gif` in [50]). Also, we compute the (poste-

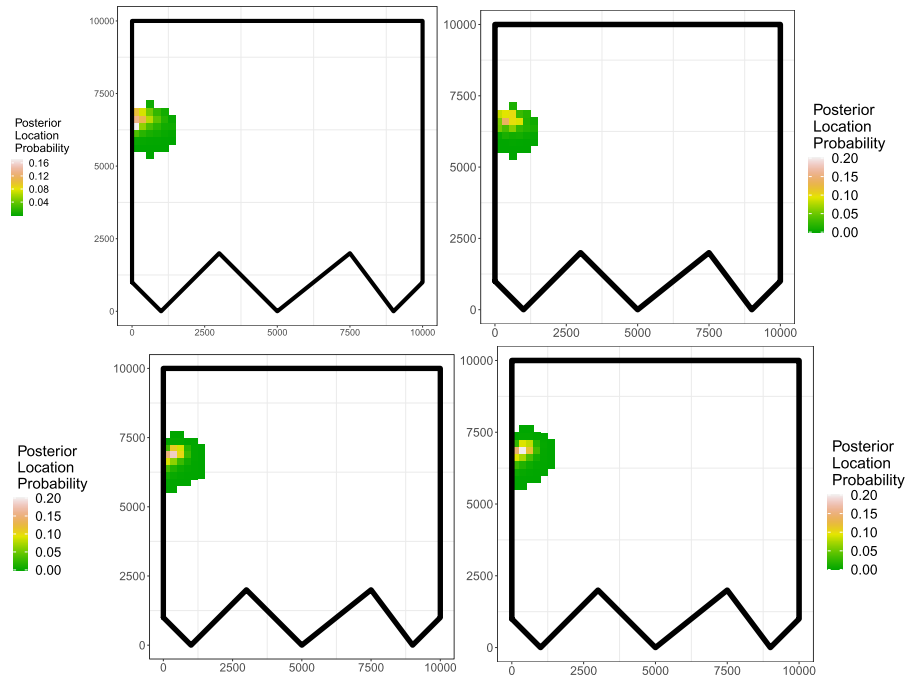

Figure 4 Posterior location probabilities. Posterior location probabilities at time $t = 0$ for a given device under emission models RSS with uniform prior (top left), RSS with network prior (top right), SDM with uniform prior (bottom left), and SDM with network prior (bottom right)

rior) joint location probabilities $\gamma_{dt,ij} \equiv \mathbb{P}(T_{dt} = i, T_{dt-1} = j|\mathbf{E}_{d1:T}, \mathbf{I}^{\text{aux}})$. These probabilities γ_{dti} and $\gamma_{dt,ij}$ constitute the output data for this module. Mathematical details of the whole model construction are included in the supplementary material.

3.2 Model evaluation

To evaluate the performance of these geolocation models we shall mimic the usual approach in Official Statistics to focus on the mean squared error as the most relevant figure of merit for accuracy, concentrating on their bias and variance components. In this line of thought, we shall introduce the following definitions:

1. The *center of location probability* $\mathbf{cp}_{dt}$ of device d at time t defined as

$$\mathbf{cp}_{dt} = \sum_{i=1}^{N_T} \gamma_{dti} \begin{pmatrix} x_i^{(c)} \\ y_i^{(c)} \end{pmatrix}, \tag{5}$$

where $x_i^{(c)}, y_i^{(c)}$ stand for the x and y coordinates of the centroid of tile i. This can be understood as an estimation of the position of the device according to the posterior mean. Notice that this quantity plays a similar role to a first-order spatial moment for the distribution γ_{dti}. Then, we can view the Euclidean distance between the true position $\mathbf{r}_{dt}^*$ and the center of location probability $\mathbf{cp}_{dt}$ of a device d at time t as a bias-equivalent indicator of the geolocation estimation procedure:

$$\mathrm{b}_{dt} = \left\| \mathbf{cp}_{dt} - \mathbf{r}_{dt}^* \right\|. \tag{6}$$

2. The *radius of location probability dispersion* rd_{dt} of device d at time t with respect to position $\mathbf{r}^*_{dt} = (x^*_{dt} y^*_{dt})^T$ defined as

$$\mathrm{rd}_{dt}\left(\mathbf{r}^*_{dt}\right) = \sqrt{\sum_{i=1}^{N_T} \gamma_{dti}\left[\left(x_i^{(c)} - x^*_{dt}\right)^2 + \left(y_i^{(c)} - y^*_{dt}\right)^2\right]}, \tag{7}$$

where (x^*_{dt}, y^*_{dt}) stands for the reference x and y coordinates of the device d at time t. This can be understood as a root mean squared dispersion with respect to a reference position. Notice that this quantity plays a similar role to a standard spatial deviation for the distribution γ_{dti} when the reference position is taken as the center of location probability:

$$\mathrm{rmsd}_{dt} = \mathrm{rd}_{dt}(\mathbf{cp}_{dt}). \tag{8}$$

Notice that we can also generalize these definitions by using alternative distance functions instead of the Euclidean distance such as the Manhattan distance or similar. Obviously, these figures of merit are not exhaustive and we can propose more (e.g. to measure the kurtosis, concentration, etc.). Having the set of probability distributions γ_{dti} and the true position values many choices arise.

In Figs. 5 and 6 we represent the distributions of b_{dt} and rmsd_{dt} for the population of devices in our simulated scenario. The advantage of using a simulator providing a ground truth is that we may draw relevant conclusions. Firstly, the RSS emission model seems to provide more accurate estimates in terms of the distance to the true position of the devices, but the SDM emission model with the uniform prior provides less disperse spatial distributions. Since the connection type (see Table 1 in the supplementary material) is `strength`, i.e. the BTS-device connection mechanism establishes according to the RSS, the RSS emission model is trivially closer to this true connection mechanism, providing best geolocation estimates. Furthermore, according to Fig. 1, the SDM emission model is more localized (this is the effect of the logistic transformation), thus the root mean squared dispersion is lower. Secondly, the radio wave propagation model plays a central role in the emission model and thus in the geolocation procedure. This underlies the importance of the joint MNO-NSO collaboration in the design stage. The RSS emission model is too simplistic for real life conditions (e.g. due to the load balancing of the network) and the SDM emission model needs an accurate estimation of the parameters S_{mid} and S_{steep}. Thirdly, the use of a dynamical approach with an HMM allows us to compute location probabilities even for those time instants in which no network event is recorded. Lastly, there exist time instants where a radio cell oscillation phenomenon is detected because the mobile device moves in the frontier of two neighboring coverage areas. In the HMM approach, contrary to intuition, this leads to an accurate geolocation estimate since we are having more information (from two radio cells) than otherwise. Thus, with the dynamical approach we gain in accuracy.

4 Device duplicity

The target populations of statistical analyses of network mobile data are populations of human individuals (present population, domestic tourists, commuters, etc.). It is well-known that a non-negligible fraction of mobile subscribers carries more than one device.

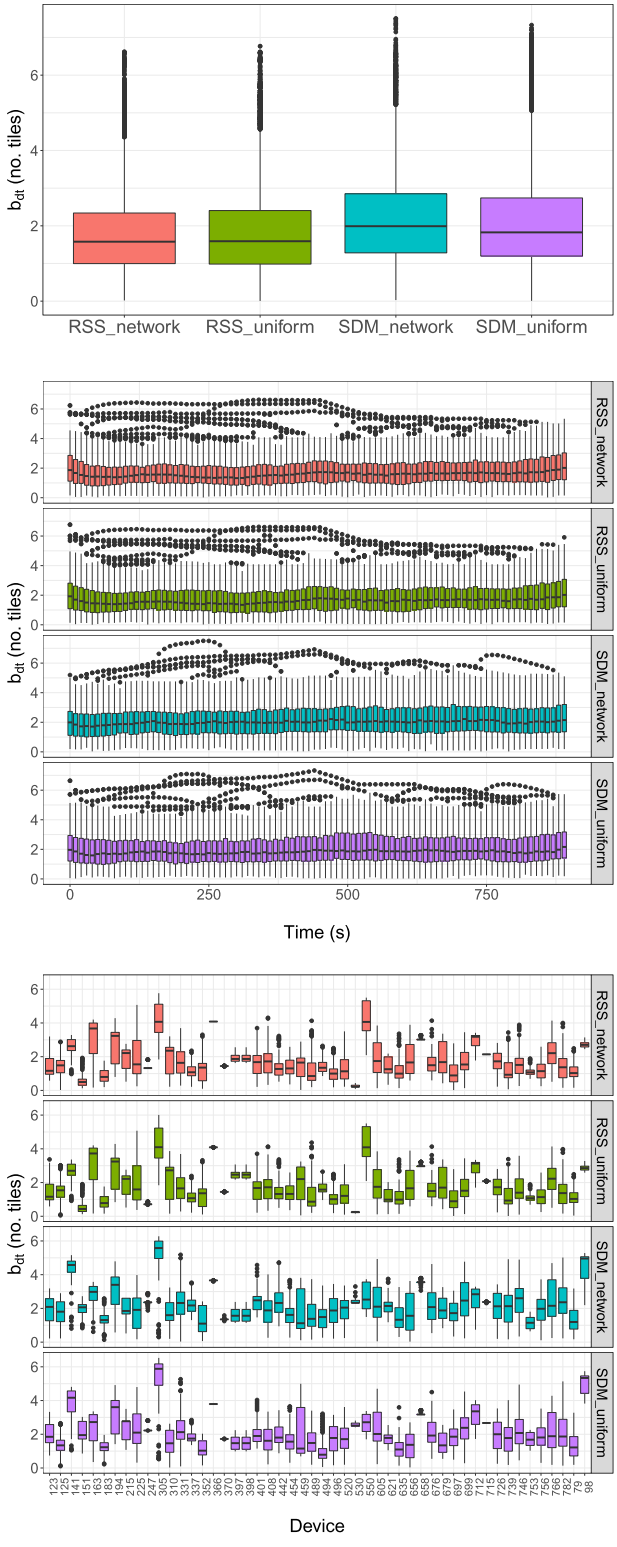

Figure 5 Distance between centers of location probabilities and true positions. (Top) Distribution of distance $\mathbf{b}_{dt}$ for all times t and all devices d for models $\mathbf{RSS}$ and $\mathbf{SDM}$ with uniform and network priors. (Middle) Time evolution of distributions of distance $\mathbf{b}_{dt}$ for all devices d for the same models. (Bottom) Distribution per device of distance $\mathbf{b}_{dt}$ for all times t for the same models (only 50 randomly selected are shown)

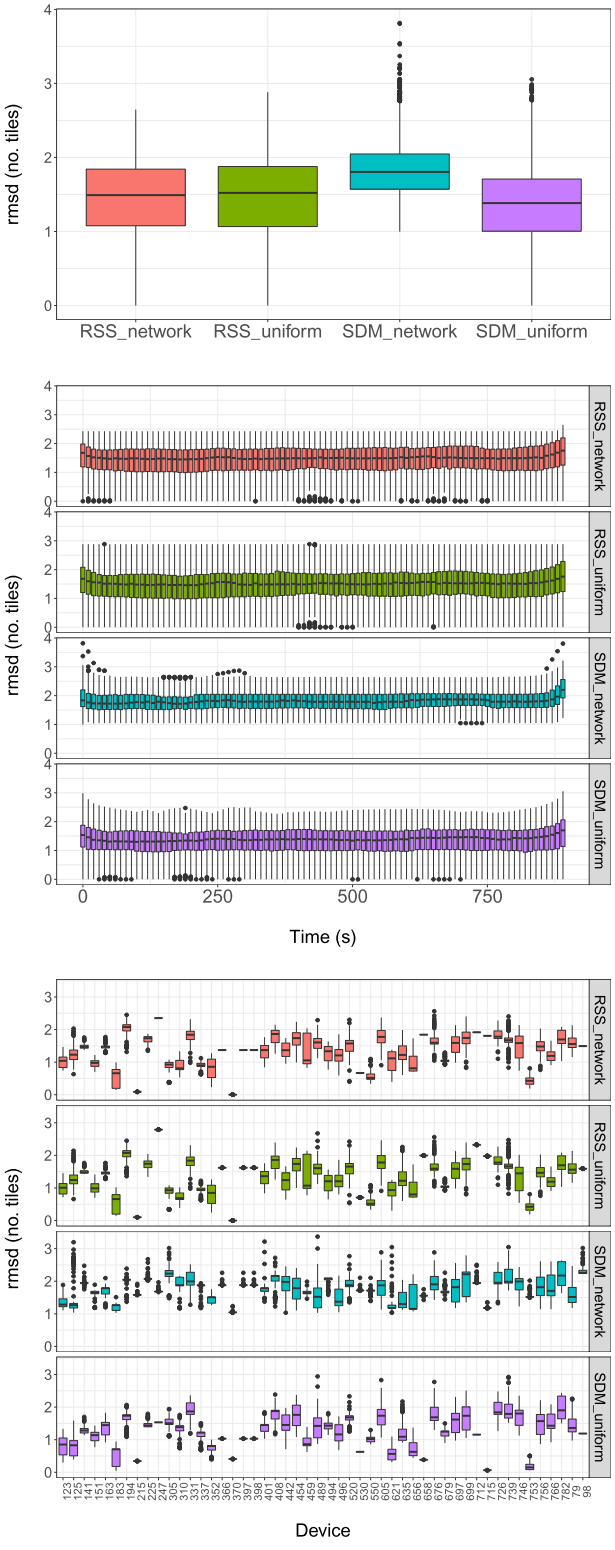

Figure 6 Root mean squared dispersions of location probabilities. (Top) Distribution of root mean squared dispersions $\mathbf{rmsd}_{dt}$ for models **RSS** and **SDM** with uniform and network priors. (Middle) Time evolution of distributions of distance $\mathbf{rmsd}_{dt}$ for the same models. (Bottom) Distribution per device of root mean squared dispersion $\mathbf{rmsd}_{dt}$ for all times t for the same models (only 50 randomly selected are shown)

We shall call this *device multiplicity*. The goal of this module will be to compute a device-multiplicity probability $p_d^{(n)}$ for each mobile device d, i.e. the probability that a device d is carried by an individual carrying n devices. The input data for this module will be the same input data as for the geolocation module, since we will make use of the same HMM.

4.1 Computation of multiplicity probabilities

For illustrative purposes we shall make the working assumption that an individual carries at most two devices. The generalization to more devices is just a matter of computational complexity of this same approach. We shall follow a Bayesian hypothesis testing approach. For each device d we shall consider the disjoint set of hypotheses $\{H_{dd'}\}_{d'=1,\dots,D}$ meaning that the devices d and d' are carried by the same individual. When $d = d'$ this reduces to mobile device d being the only mobile device carried by its corresponding individual. We focus on computing

$$p_d^{(1)} = \mathbb{P}\big(H_{dd}|\mathbf{E}_{d1:T}, \mathbf{I}^{\text{aux}}\big), \tag{9}$$

where we are using the same notation as in Sect. 3. Since the entire event set Ω_d for device d can be decomposed as $\Omega_d = \bigcup_{d'=1}^{D} H_{dd'}$, we can make use of Bayes' theorem to write:

$$\begin{aligned} p_d^{(1)} &= \frac{\mathbb{P}(\mathbf{E}_{d1:T}|H_{dd}, \mathbf{I}^{\text{aux}}) \cdot \mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}})}{\mathbb{P}(\mathbf{E}_{d1:T}|H_{dd}, \mathbf{I}^{\text{aux}}) \cdot \mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}}) + \sum_{d' \neq d} \mathbb{P}(\mathbf{E}_{d1:T}, \mathbf{E}_{d'1:T}|H_{dd'}, \mathbf{I}^{\text{aux}}) \cdot \mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})} \\ &= \frac{1}{1 + \sum_{d' \neq d} \alpha_{dd'} \cdot \exp(\ell_{dd'} - \ell_d)}, \end{aligned} \tag{10}$$

where we have defined the prior probability ratios $\alpha_{dd'} = \frac{\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})}{\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}})}$ and the integrated log-likelihoods $\ell_d = \mathbb{P}(\mathbf{E}_{d1:T}|H_{dd}, \mathbf{I}^{\text{aux}})$ for a single device d and $\ell_{dd'} = \mathbb{P}(\mathbf{E}_{d1:T}, \mathbf{E}_{d'1:T}|H_{dd'}, \mathbf{I}^{\text{aux}})$ for two devices d and d'. These quantities are computed as follows. Firstly, the integrated log-likelihood ℓ_d for a single device d corresponds to the HMM model introduced above. Secondly, the integrated log-likelihood $\ell_{dd'}$ for two devices d and d' is computed according to the HMM duplicity model represented by the graphical model in Fig. 7. Computation is conducted in a similar way as before with the noticeable difference in the emission model: emission probabilities are computed as the product of the original single-device emission probabilities for d and d' (see supplementary material for details).

For the specification of priors we reason as follows. The key ingredient is the auxiliary information $\mathbf{I}^{\text{aux}}$. For example, if some auxiliary information at the device level is available (for instance from the Customer Relationship Management database) showing that

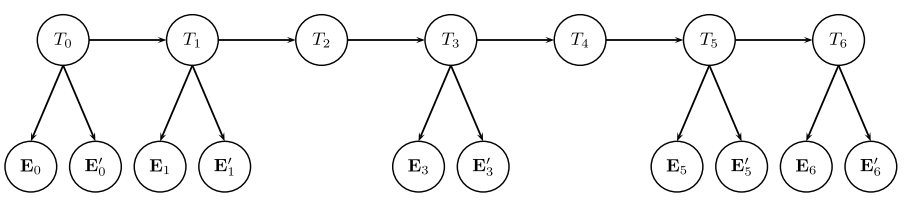

Figure 7 HMM duplicity model. Graphical model for an individual carrying two devices, thus generating pair of events $\mathbf{E}_{d1:T}$ and $\mathbf{E}_{d'1:T}$

devices d and any other d' reside in far away locations, then naturally $\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}}) \approx 0$ so that $p_d^{(1)} \approx 1$, as expected.

If no individual prior information is used, we can reason as follows. Firstly, let λ_d denote the prior odds ratio $\lambda_d = \frac{\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}})}{1-\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}})}$, which expresses how much more probable is that an individual carries a priori only one device d than another device together with d. This quantity may be fixed using auxiliary information from an external source (e.g. the CRM database or an external survey). Secondly, since no auxiliary information is used, a priori any other device d' can be the second device, so that $\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})$ is constant for any other device $d' \neq d$. Since $\Omega_d = \bigcup_{d'=1}^{N_D} H_{dd'}$, then $\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}}) + (N_D - 1) \cdot \mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}}) = 1$ for any other device d'. We arrive at

$$\begin{aligned}
\mathbb{P}\big(H_{dd}|\mathbf{I}^{\text{aux}}\big) &= \frac{\lambda_d}{1+\lambda_d},\\
\mathbb{P}\big(H_{dd'}|\mathbf{I}^{\text{aux}}\big) &= \frac{1}{(1+\lambda_d)\cdot(N_D-1)},\\
\alpha_{dd'} &= \frac{1}{\lambda_d \cdot (N_d - 1)},\\
p_d^{(1)} &= \frac{1}{1+\frac{\exp(-\ell_d)}{\lambda_d\cdot(N_d-1)}\sum_{d'\neq d}\exp(\ell_{dd'})}.
\end{aligned} \tag{11}$$

A natural choice for λ_d when there are more devices N_D than individuals N^{net} in the network is given by

$$\lambda_d = \frac{1-\frac{2\times(N_D-N^{\text{net,ext}})}{\binom{N_D}{2}}}{\frac{2\times(N_D-N^{\text{net,ext}})}{\binom{N_D}{2}}},$$

where $N^{\text{net,ext}}$ is an estimate of N^{net} from an external source (CRM database, etc.). If an external estimate $\hat{r}_2$ of the fraction of individuals r_2 in the network carrying two devices is available, then we can choose

$$\lambda_d = \frac{1-\hat{r}_2}{\hat{r}_2}.$$

If we can provide local estimates (because devices are assigned to delimited regions), then we do not need to consider the whole set of mobile devices and we can set

$$\lambda_d = \frac{N_D^{\text{loc}}}{N_D}\lambda_d^{\text{loc}},$$

where the same reasoning as above applies to λ_d^{loc} at a local scale.

4.2 Results on simulated data

We have applied this approach to our simulated data set with $N = 500$ individuals in the target population, $N^{\text{net}} = 186$ individuals detected by the network (subscribers), and $N_D = 218$ mobile devices. Obviously, there exist individuals carrying two devices. We apply the formalism above to provide duplicity probabilities $p_d^{(2)} = 1 - p_d^{(1)}$ for each device d. We set the value $\lambda_d^{(1)} = \frac{0.85}{0.15}$ assuming faithful external information (the result is robust

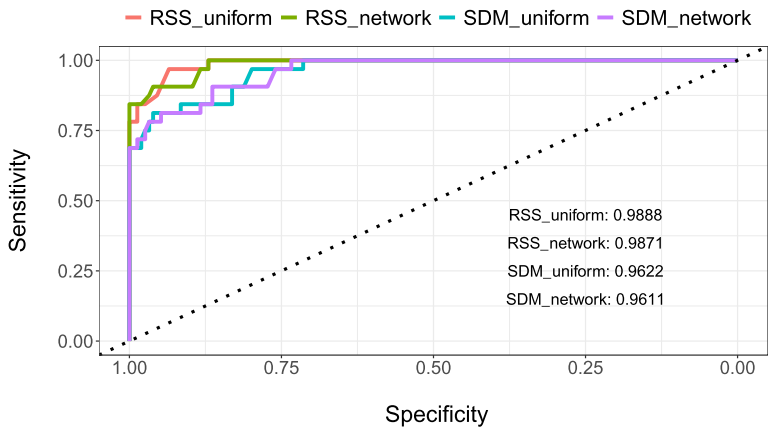

Figure 8 ROC curves for the duplicity probabilities $p_d^{(2)}$. ROC curves for the duplicity probabilities $p_d^{(2)} = 1 - p_d^{(1)}$ computed according to eq. (9) for HMMs with two priors and two emission models

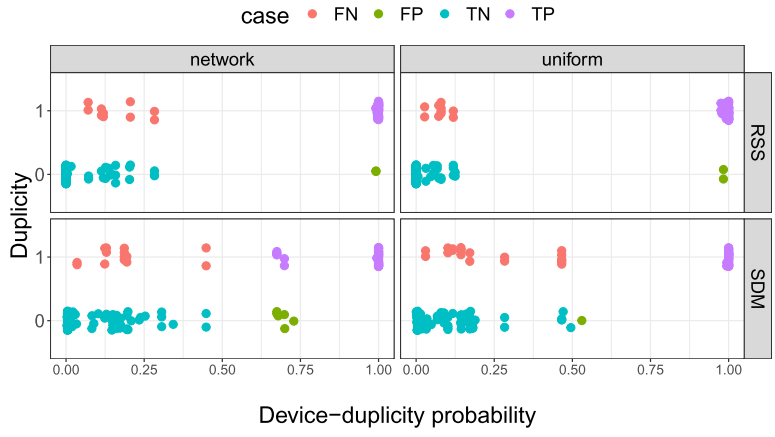

Figure 9 Cases for the duplicity probabilities $p_d^{(2)}$. True/false positive/negative cases for the duplicity probabilities $p_d^{(2)} = 1 - p_d^{(1)}$ computed according to eq. (9) for four different HMMs with two priors and two emission models

enough around this value – see supplementary material for details). The duplicity probabilities are computed in four scenarios combining two different emission models (RSS and SDM) with two different prior location probabilities (uniform and network). We compare the results with the (synthetic) ground truth to assess the performance. In Fig. 8 we represent the ROC curves for the duplicity probabilities for the four models, together with their corresponding area under the curve (AUC). In Fig. 9 we represent the different cases (true/false positive/negative) in each HMM model.

Taking into account that the connection mechanism in this simulation is based on the RSS and that the initial true positions are chosen at random by the simulator (not based on the network configuration), we conclude that the larger the mismatch between the connection mechanism (the reality) and the emission model (the chosen model), the poorer the performance of the classification of devices, as one may expect. The SDM choice for the emission model departs from the actual handover mechanism and we observe in Fig. 9 that duplicity probabilities show lower quality. This is also observed with the priors in the

same figure: the uniform choice is more appropriate to this simulated scenario than the network choice. This shows the importance of the collaboration between MNOs and NSIs in incorporating the network configuration into the emission model and the choice of location priors using as much auxiliary information as possible (home and other anchor points estimation).

For these results we also observe that false negative cases are generated by those pairs of devices having exactly the same pairwise degenerate sequence of network events in which only one BTS connects to each pair of devices. The algorithm fails to detect them as devices carried by the same individual. This is explained by the HMM itself, since the transition matrix is the diagonal matrix and no transition is indeed allowed. In this case the duplicity is much less probable than the single device per individual. A complementary test is needed when a connection to only one BTS is detected, which in turn will be less probable as the time period of analysis is longer.

For the case of false positive cases, we observe that these arise from quasi-identical sequences of network events, which is an expected behaviour. With longer time periods of observation, these cases will presumably come to be negligible.

5 Statistical filtering

This module is devoted to the identification of the target population in the mobile network data set and derived data sets (posterior location probabilities, for example). In practical terms this amounts to identifying domestic tourists, inbound tourists, commuters, etc. in our data sets. We refer to this as *statistical filtering*, where we use the term *statistical* to distinguish this filtering exercise from the preprocessing steps in which, e.g., machine-to-machine data are previously filtered out. Notice that the latter rests mostly on technological issues and definitions, whereas the former is a clearly statistical analytical exercise.

As in the whole approach proposed in this work, we shall be focusing on geolocation data, i.e., on movement data discarding interaction information (e.g. calls among subscribers) or Internet traffic (e.g. usage of mobile apps). In a fully-fledged production environment in real conditions, the ideal scenario would be to use as much information as possible. Thus, we shall concentrate on analyses upon the geolocation data, i.e. upon the network event data and location probabilities derived thereof.

Regretfully, given the problems in accessing real mobile network data, and the current status of development of the network event data simulator, the contents of this module are not so far developed as the preceding ones. The current displacement patterns for individuals (hence also for mobile devices) in the data simulator are restricted to random walks, random walks with drift, both with intermixing periods of stops (stays, i.e. no displacement at all) for the whole population. In this sense, we lack synthetic data to test concrete proposals, not as with the geolocation of data. We would need more complex and realistic individual displacement patterns and elements (Lévy flights, home/work locations, usual environments, etc.). For this reason, we will limit ourselves to provide more generic guidelines to be implemented in the future both on real data and on synthetic data after a further development of the network event data simulator.

5.1 General approach

Our proposed approach for the statistical filtering of target populations is strongly based on the geolocation outputs obtained from the preceding process modules. Different aspects are to be taken into account. As before, the target mobile network data is assumed to

be basically some form of signalling data so that time frequency and spatial resolution are high enough as to allow us to analyse movement data in a meaningful way. In this sense, for example, CDR data only provides up to a few records per user in an arbitrary day which makes virtually impossible any rigorous data-based reasoning in this line. Next, the use of hidden Markov models, as described in Sect. 3, implicitly incorporates a time interpolation which will be very valuable for this statistical filtering exercise. In this way we avoid the issues arising from noncontinuous traces approaches (see e.g. [60] for home location algorithms). However, a wider analysis is needed to find the optimal time scope. The spatial resolution issue is dealt with by using the reference grid introduced in Sect. 3. This releases the analyst from spatial techniques such as Voronoi tessellation, which introduces too much noise for our purposes. Nonetheless, the uncertainty measures computed from the underlying probabilistic approach for geolocation must be taken into account to deal with precision issues in different regions (e.g. high-density populated vs. low-density populated). The algorithms to be developed to statistically filter the target population will be mainly based on quantitative measures of movement data. In particular, from the HMMs fitted to the data (especially the location probabilities) we shall derive a probability-based trajectory per device which will be the basis for these algorithms.

Once a trajectory is assigned to each device, different indicators and measures of movement shall be computed upon which we shall apply algorithms to determine usual environment, home/work location, second home location, leisure activity times and locations, etc. A problematic aspect with this new data source is that traditional statistical definitions will need some revision or refinement. For example, in the home detection problem, which is an intermediate problem in the identification of target populations, census data (or similar official data) are commonly used to calibrate or validate estimates. The notion of home obtained from traditional sources is mainly an administrative concept arising from the use of administrative registers. In this way, e.g., a University student may be registered in her family home whereas she spends nine months in a college. What definition of home should then be used? This has introduced subtleties like the distinction between residential and present population in official statistics. In this line of thought, an important input for target population identification algorithms is the establishment of a clear-cut definition for each statistical concept involved, so that the algorithms are designed to cover these definitions. A critical issue in the development of this kind of algorithms is the validation procedure. On the one hand, the use of the simulator, once more complex and realistic displacement patterns have been introduced, will offer us in the future a validation against the simulated ground truth. On the other hand, with real data two main problems need to be tackled, namely (i) the use of pseudoanonymised real data will prevent us to link mobile device records with official registers, so only indirect aggregated validation procedures can be envisaged (thus inviting the ecological fallacy to permeate the whole analysis), and (ii) the representativity of the tested sample of devices to validate the algorithm for the whole population needs to be rigorously assessed.

In the supplementary material we provide a generic view of quantitative measures of movement data, together with some concrete illustrative examples, upon the probability-based trajectories assigned to the geolocated data (location probabilities) obtained from the application of an HMM. For our illustrative example focused on present population counts and origin-destination matrices estimation, we shall skip this step, since the whole population will be our target population.

6 Aggregation of individuals detected by a network

This module focuses on providing a probability distribution for the number of individuals detected by a mobile telecommunication network. This module will take the posterior location probabilities and the multiplicity probabilities as input data. After introducing some general remarks, we shall provide a method to build the target probability distribution, which will then be adapted to provide also the probability distribution of individuals displacing between territorial units at each time instant.

6.1 General remarks

Firstly, the aggregate information is on the number of *detected individuals*, not on the number of devices. This is a very important difference with virtually any other approach found in the literature (see e.g. [6, 10]). We take advantage of the preceding modules working at the device level to study in particular the device multiplicity per individual. This has strong implications regarding agreements between NSOs and MNOs to access and use their mobile network data for statistical purposes.

As we can easily see, working with the number of devices instead of the number of individuals poses severe identifiability problems requiring more auxiliary information. Let us consider an extremely simplified illustrative example. Let us consider a population U_1 of 5 individuals with 2 devices each one and a population U_2 of 10 individuals with 1 device each one. Suppose that in order to make our inference statement about the number N of individuals in the population we build a statistical model relating N and the number of devices $N^{(\mathrm{dev})}$, that is, basically we have a probability distribution $\mathbb{P}_N(N^{(\mathrm{dev})})$ for the number $N^{(\mathrm{dev})}$ of devices dependent on the number of individuals, from which we shall infer N. In this situation we have $\mathbb{P}_{N^{(1)}} = \mathbb{P}_{N^{(2)}}$ even when $N^{(1)} \neq N^{(2)}$. There is no statistical model whatsoever capable of distinguishing between U_1 and U_2 (see Definition 5.2 in [61] for unidentifiable parameters in a probability distribution). To cope with the duplicity of devices using an aggregated number of devices we would need further auxiliary information, which furthermore must be provided at the right territorial and time scale.

Secondly, we shall use again the language of probability in order to carry forward the uncertainty already present in the preceding stages all along the end-to-end process. In another words, if the geolocation of network events is conducted with certain degree of uncertainty (due to the nature itself of the process - see Sect. 3) and if the duplicity of a given device (carried by an individual with another device) is also probabilistic in nature (see Sect. 4), then a priori it is impossible to provide a certain number of individuals[2] in a given territorial unit. For this reason, we shall focus on the probability distribution of the number of individuals detected by the network and shall avoid producing a point estimation. Notice that having a probability distribution amounts to having all statistical information about a random phenomenon and you can choose a point estimation (e.g. the mean, the mode or the median of the distribution) together with an uncertainty measure (coefficient of variation, credible intervals, etc.).

Thirdly, the problem is essentially multivariate and we must provide information for a set of territorial units. Thus, the probability distribution must be a multivariate distribution. Notice that this is not equivalent to providing a collection of marginal distributions

[2]Notice that this same argument is valid for the number of devices.

over each territorial unit. Obviously, there will be a correlation structure, the most elementary expression of which is that individuals detected in a given territorial unit cannot be detected in another region, so that the final distribution needs to incorporate this restriction in its construction.

Finally, the process of construction of the final multivariate distribution for the number of detected individuals must make as few modelling assumptions as possible, if any. In case an assumption is made (and this should be accomplished in any use of statistical models for the production of official statistics, in our view), it should be made as explicit as possible and openly communicated and justified. In this line of thought, we shall strongly base the aggregation procedure on the results of preceding modules avoiding any extra hypothesis. Basically, our starting assumptions for the geolocation and the duplicity detection will be carried forward as far as possible without introducing new modelling assumptions of any kind.

6.2 Probability distribution of the number of detected individuals

To implement the principles outlined above, we shall slightly change the notation. Firstly we define the vectors $\mathbf{e}_i^{(1)} = \mathbf{e}_i$ and $\mathbf{e}_i^{(2)} = \frac{1}{2}\mathbf{e}_i$, where $\mathbf{e}_i$ is the ith canonical unit vector in $\mathbb{R}^{N_T}$ (with N_T the number of tiles in the reference grid). These definitions are set up under the working assumption of individuals carrying at most 2 devices in agreement with the proposal devised in Sect. 4. Should we consider a more general situation, the generalization is obvious, although more computationally demanding.

Next, we define the random variable $\mathbf{T}_{dt} \in \{\mathbf{e}_i^{(1)}, \mathbf{e}_i^{(2)}\}_{i=1,\dots,N_T}$ with probability mass function $\mathbb{P}(\mathbf{T}_{dt}|\mathbf{E}_{1:D}, \mathbf{I}^{\text{aux}})$ given by

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(1)}|\mathbf{E}_{1:D}, \mathbf{I}^{\text{aux}}\right) = \gamma_{dti} \cdot p_d^{(1)}, \tag{12a}$$

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(2)}|\mathbf{E}_{1:D}, \mathbf{I}^{\text{aux}}\right) = \gamma_{dti} \cdot p_d^{(2)}, \tag{12b}$$

where $p_d^{(1)}$ and $p_d^{(2)}$ ($p_d^{(1)} + p_d^{(2)} = 1$) are the device duplicity probabilities introduced in Sect. 4. Notice that this is a categorical or multinoulli random variable. Finally, we define the multivariate random variable $\mathbf{N}_t^{\text{net}}$ providing the number of individuals $[\mathbf{N}_t^{\text{net}}]_i = N_{ti}^{\text{net}}$ detected by the network at each tile $i = 1, \dots, N_T$ at time instant t:

$$\mathbf{N}_t^{\text{net}} = \sum_{d=1}^{D} \mathbf{T}_{dt}. \tag{13}$$

The sum spans over the number of devices filtered as members of the target population according to Sect. 5. If we are analysing, say, domestic tourism, D will amount to the number of devices in the network classified with a domestic tourism pattern according to the algorithms designed and applied in the preceding module. For illustrative examples, since we have not developed the statistical filtering module yet, we shall concentrate on present population.

The random variable $\mathbf{N}_t^{\text{net}}$ is, by construction, a Poisson multinomial random variable. The properties and software implementation of this distribution are not trivial (see e.g. [62]) and we shall use Monte Carlo simulation methods by convolution to generate random variates according to this distribution.

The reasoning behind this proposal can be easily explained with a simplified illustrative example. Let us consider an extremely simple scenario with 5 devices and 5 individuals (thus, none of them carrying two devices), and 9 tiles (a 3×3 reference grid). Let us consider that the location probabilities $\gamma_{dti} = \gamma_{ti}$ are the same for all devices d at each time instant and each tile. In these conditions $p_d^{(1)} = 1$ and $p_d^{(2)} = 0$ for all d. Let us focus on the univariate (marginal) problem of finding the distribution of the number of devices/individuals in a given tile i. If each device d has probability γ_{ti} of detection at tile i, then the number of devices/individuals at tile i will be given by a binomial variable Binomial$(5, \gamma_{ti})$. If the probabilities were not equal, then the number of devices/individuals would be given by a Poisson binomial random variable Poisson – Binomial$(5; \gamma_{1ti}, \gamma_{2ti}, \gamma_{3ti}, \gamma_{4ti}, \gamma_{5ti})$, which naturally generalizes the binomial distribution. If we focus on the whole multidimensional problem, then instead of having binomial and Poisson-binomial distributions, we must deal with multinomial and Poisson-multinomial variables. Finally, if $p_d^{(2)} \neq 0$ for all d, we must avoid double-counting, hence the factor $\frac{1}{2}$ in the definition of $\mathbf{e}_i^{(2)}$.

Notice that the only assumption made so far (apart from the trivial question of the maximum number of 2 devices carried by an individual) is the independence for two devices to be detected at any pair of tiles i and j. This independence assumption allows to claim that the number of detected individuals distributes as a Poisson-multinomial variable, understood as a sum of independent multinoulli variables with different parameters. There is no extra assumption in this derivation. The validation of this assumption is subtle, since ultimately it will depend on the correlation between the displacement patterns of individuals in the population. If the tile size is chosen small enough, we claim that the assumption holds fairly well and it is not a strong condition imposed on our derivations. On the other hand, if the tiles are too large (think of an extreme case about a reference grid being composed of whole provinces as tiles), we should expect correlations in the detection of individuals: those living in the same province will have very large correlation and those living in different provinces will show nearly null correlation. Thus, the size of the tiles imposes some limitation to the validity of the independence assumption. Even the transport network in a territory will certainly influence these correlations. Currently, we cannot analyse quantitatively the relationship between the size of the tiles and the independence assumption with the network data simulator because we need both realistic simulated individual displacement patterns and simulated correlated trajectories (probably connected to the sharing of usual environments, home/work locations, etc.).

The issue about the size of the tile also makes us consider the computation of the distribution of the number of detected individuals at a coarser territorial degree. Let us consider a coarser territorial breakdown composed of combination of tiles called, say, regions. We shall denote them as $\bar{T}_r = \bigcup_{i \in \mathcal{I}_r} T_i$, where $\mathcal{I}_r$ denotes the set of tile indices composing region r. If the independence assumption still holds (because the size of the region is still small enough), then we can reproduce the whole derivation above just by defining the location probability $\bar{\gamma}_{dtr}$ at region r as

$$\bar{\gamma}_{dtr} = \sum_{i \in \mathcal{I}_r} \gamma_{dti}. \tag{14}$$

The subsequent elaboration to build the final Poisson-multinomial-distributed number of detected individuals is completely similar. Notice again that there exists a limitation in

the sum of device-level distributions put by the size of the underlying region breakdown. The random vector $\bar{\mathbf{N}}_t^{\text{net}}$ of individuals per region in terms of the deduplicated location $\bar{\mathbf{T}}_{dt}$ per region would be also expressed as a sum:

$$\bar{\mathbf{N}}_t^{\text{net}} = \sum_{d=1}^{D} \bar{\mathbf{T}}_{dt}. \tag{15}$$

Notice that this decomposition allows us to write straightforwardly the mean vector and the covariance matrix for $\bar{\mathbf{N}}_t^{\text{net}}$. Define the deduplicated location probabilities per region as $\bar{\gamma}_{dtr}^{\text{dedup}} \equiv (1 - \frac{p_d^{(2)}}{2}) \cdot \bar{\gamma}_{dtr}$ for all regions $r = 1, \ldots, R$. Then

$$\mathbb{E}\big[\bar{\mathbf{N}}_t^{\text{net}}\big] = \sum_{d=1}^{D} \sum_{r=1}^{R} \bar{\gamma}_{dtr}^{\text{dedup}} \mathbf{e}_r, \tag{16}$$

$$\mathbb{V}\big[\bar{\mathbf{N}}_t^{\text{net}}\big] = \sum_{d=1}^{D} \sum_{r=1}^{R} \bar{\gamma}_{dtr}^{\text{dedup}} \cdot \big(1 - \bar{\gamma}_{dtr}^{\text{dedup}}\big) E_{rr}. \tag{17}$$

6.3 Probability distribution for the number of detected individuals moving between territorial units

The construction of the probability distribution for the number of individuals $\bar{\mathbf{N}}_t^{\text{net}}$ detected by the network can be easily generalized to the number of individuals $\bar{N}_{t,\cdot\cdot}^{\text{net}}$ detected by the network moving between territorial units. We begin by defining matrices $E_{rs}^{(1)} = E_{rs}$ and $E_{rs}^{(2)} = \frac{1}{2} E_{rs}$, where E_{rs} are the Weyl matrices of dimension $R \times R$. Next, we define the matrix random variable $E_{dt} \in \{E_{rs}^{(1)}, E_{rs}^{(2)}\}_{r,s=1\ldots,R}$ with probability mass function given by

$$\mathbb{P}\big(E_{dt} = E_{rs}^{(1)}\big) = \bar{\gamma}_{dt,sr} \cdot p_d^{(1)}, \tag{18a}$$

$$\mathbb{P}\big(E_{dt} = E_{rs}^{(2)}\big) = \bar{\gamma}_{dt,sr} \cdot p_d^{(2)}, \tag{18b}$$

where $\bar{\gamma_{dt,sr}}$ stands for the joint location probabilities computed in the geolocation module aggregated to the regions $r, s = 1, \ldots, R$. Notice that, although matrix-valued, this is still a categorical or multinoulli random variable. Then, we can define the origin-destination matrix between regions of individuals detected by the network by

$$\bar{\mathbf{N}}_t^{\text{net}} = \sum_{d=1}^{D} E_{dt}, \tag{19}$$

which, as before, distributes according to a Poisson-multinomial distribution. Again, we shall use Monte Carlo techniques to deal with it. If we define the deduplicated joint location probabilities $\bar{\gamma}_{dt,sr}^{\text{dedup}} = (1 - \frac{p_d^{(2)}}{2}) \cdot \bar{\gamma}_{dt,sr}$, then the mean origin-destination matrix is given by

$$\mathbb{E}\big[\bar{\mathbf{N}}_t^{\text{net}}\big] = \sum_{d=1}^{D} \sum_{r,s=1}^{R} \bar{\gamma}_{dt,sr}^{\text{dedup}} \cdot E_{rs}. \tag{20}$$

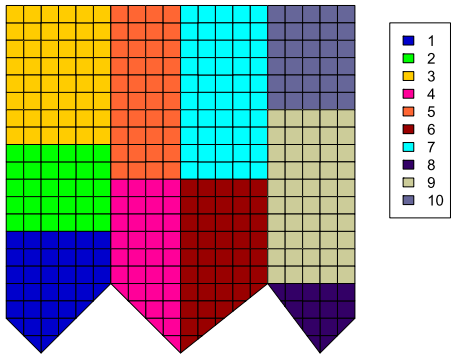

Figure 10 Regions as aggregated territorial units of analysis. Regions are obtained by aggregation of tiles of the reference grid

6.4 An example with simulated data

Let us illustrate this approach with an example generated with the mobile network event simulator. We consider again the toy scenario with a population of 186 subscribers with 218 mobile devices in a territory with a bounding box of 10 km $\times$ 10 km divided into 10 regions as in Fig. 10. The simulator provides the true position of each individual at each time instant as well as the correspondence between individuals and devices so that we can make a comparison with the (synthetic) ground truth.

The posterior distributions of the number of individuals $\bar{N}_t^{\text{net}}$ per region detected by the network is computed with Monte Carlo techniques and the results are represented in Fig. 11. Once we have posterior distributions we can also compute credible intervals for each region and each time instant (see Fig. 12). Although we can observe a good degree of accuracy, there exists a non-negligible number of regions and time instants in which the intervals do not cover the true values. A deeper analysis to unravel the roles of the geolocation and the duplicity probability computation is needed and is beyond the scope of this paper (false negative cases for duplicity has not been corrected, the HMM state definition does not include velocity, and regions and coverage areas have no correlation at all, thus all being very simplistic – see Sect. 9).

We can also construct origin-destination matrices for the number of individuals detected by the network and compare with true values provided by the simulator. Indeed, according to the proposed methodology we can even compute their credible intervals (see Fig. 13).

These probabilities, together with the device duplicity probabilities and auxiliary information from official data and the telco market, will be the input data for the last module on inference.

7 Inference

The final module focuses on the computation of the probability distribution for the number of individuals in the target population conditioned on the number of individuals detected by the network and some auxiliary information. Our first observation is that this auxiliary information is absolutely necessary to provide a meaningful inference on the target population due to similar identifiability reasons as those mentioned in Sect. 6.1 to introduce the deduplication module. This auxiliary information will be basically telco market information in the form of penetration rates (ratio of number of devices to number of individuals in the target population) and register-based population data. This information will provide the necessary link between the number of individuals at the network level

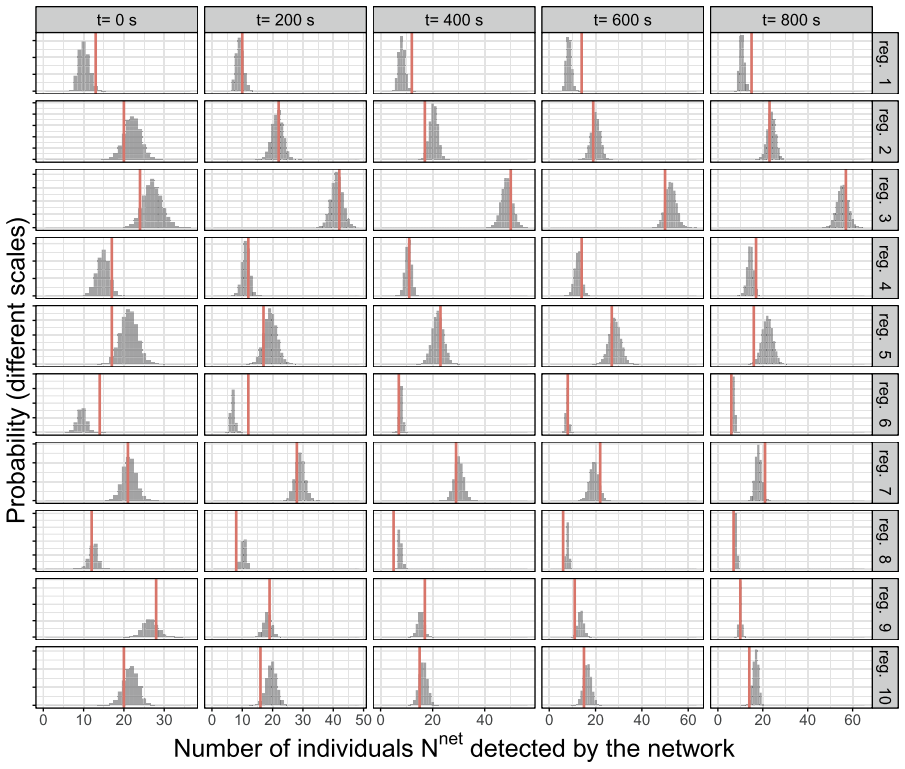

Figure 11 Posterior distributions for the number of individuals $\bar{N}_t^{\mathbf{net}}$ per region. Posterior distributions for the number of individuals $\bar{N}_t^{\mathbf{net}}$ per region detected by the network compared to true values (in red) computed according to eq. (15) using the RSS geolocation model with uniform prior. Only a sample of time instants is shown for visibility's sake

and at the target population level. This combination of data sources is indeed desirable not only to produce better and more accurate estimates but also to provide more coherent information among diverse data sources. However, notice that this data integration must avoid imposing findings from one data source on the other data source thus precluding new insights about the target population.

In more concrete terms, register-based population figures offer information about society from a concrete demographic perspective (residential population) with a given degree of spatial and time breakdown. Mobile network data, however, provides the opportunity to reach unprecedented spatial and time scales as well as a complementary view on the population (present population). The integration of sources, in our view, must be careful with these differences bringing similarities and contrasts at the same time into the statistical analysis. In this line of thought, we propose to use hierarchical models (i) to produce probability distributions, (ii) to integrate data sources, and (iii) to account for the uncertainty and the differences of concepts and scales.

We propose a two-staged modelling exercise. Firstly, we assume that there exists an initial time instant t_0 in which both the register-based target population and the actual population can be assimilated in terms of their physical location. We can assume, e.g., that at 6:00am all devices stay physically at the residential homes declared in the population register. This assumption will trigger the first stage in which we compute a probability distribution for the number of individuals $\mathbf{N}_{t_0}$ of the target population in all regions in terms

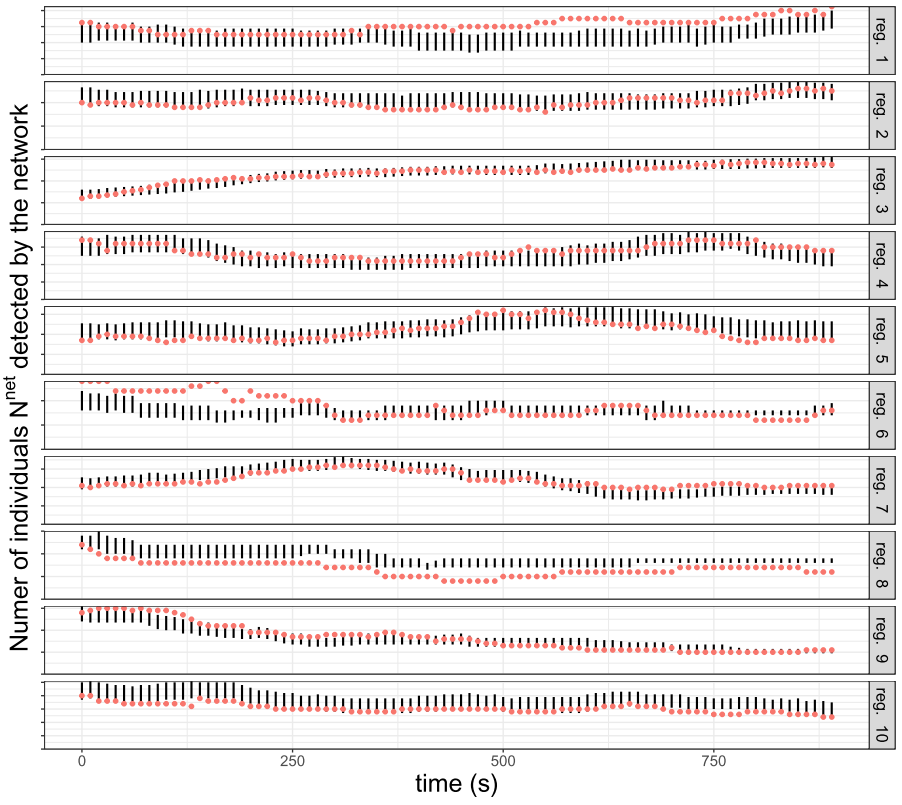

Figure 12 Credible intervals for the number of individuals $\bar{N}_t^{\mathbf{net}}$ per region. Credible intervals (prob = 95%) for the number of individuals $\bar{N}_t^{\mathbf{net}}$ per region detected by the network compared to true values (in red) computed according to the distribution in eq. (15) using the RSS geolocation model with uniform prior

of the number of individuals $\mathbf{N}_0^{\text{net}}$ detected by the network and the auxiliary information. Secondly, we assume that individuals displace over the geographical territory independently of the MNO, i.e. subscribers of MNO 1 will show a displacement pattern similar to those of MNO 2. This assumption will trigger the second stage in which we provide a probability distribution for the number of individuals $\mathbf{N}_t$ for later times $t > t_0$.

Regarding the origin-destination matrix, we can use the same assumptions to infer the number of individuals moving from one region to another at time instant t, also providing credible intervals as an accuracy indicator.

7.1 Present population at the initial time t_0

For ease of notation we shall drop the time index in this section. The auxiliary information is provided by the penetration rates P_r^{net} of the MNO and the register-based population N_r^{reg} at each region r. We shall combine N_r^{net}, P_r, and N_r^{reg} to produce the probability distribution for $\mathbf{N} = (N_1, \ldots, N_R)^T$. We follow the approach used in the species abundance problem in Ecology [63]. This approach clearly distinguishes between the state and the observation process. The state process is the underlying dynamical process of the population and the observation process is the procedure by which we get information about the location and timestamp of each individual in the target population. The different available auxiliary information will be integrated using different levels in the hierarchy of the statistical model.

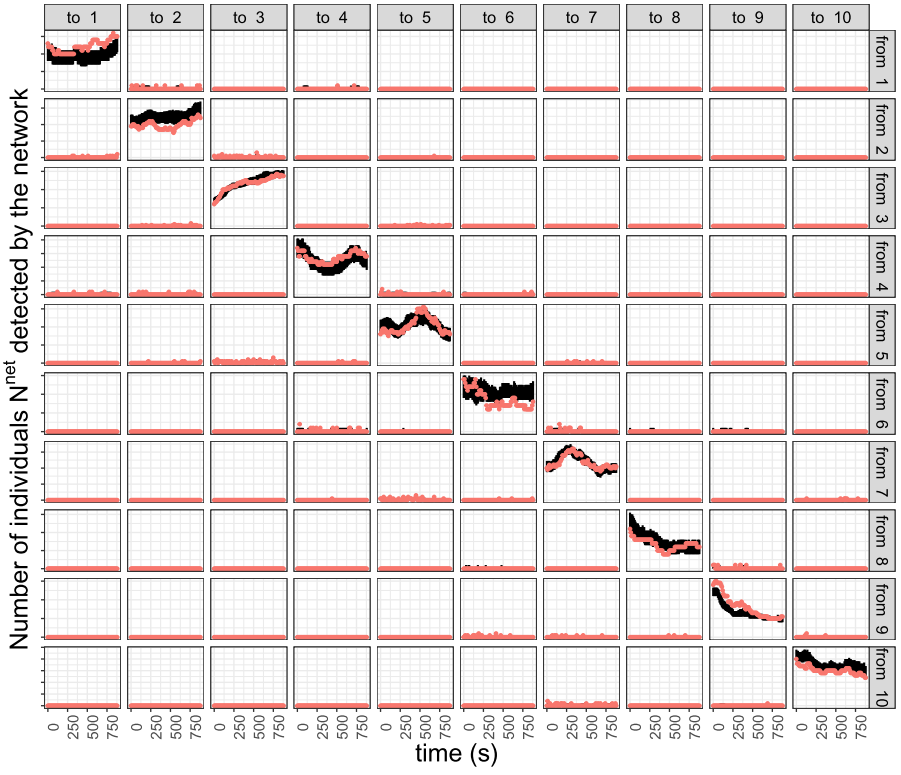

Figure 13 Credible intervals for the OD matrix of the number of individuals N_t^{net} per region. Credible intervals (prob = 95%) for the OD matrix of number of individuals $\tilde{N}_t^{\text{net}}$ per region detected by the network compared to true values (in red) computed according to the distribution in eq. (19) using the RSS geolocation model with uniform prior

The first level makes use of the detection probability p_r of individuals detected by the telecommunication network in each region r. We model

$$N_r^{\text{net}} \simeq \text{Binomial}(N_r, p_r). \tag{21}$$

Model (21) makes the only assumption that the probability of detection p_r for all individuals in region r is the same. This probability of detection amounts basically to the probability of an individual of being a subscriber of the given mobile telecommunication network. This assumption will be further discussed below. As a first approximation, we may think of p_r as a probability related to the penetration rate P_r of the MNO in region r.

As an overview of the hierarchy of models, we shall firstly consider only the observation process, i.e. no population dynamics (state process) is modelled. In the hierarchy, we shall be introducing deeper degrees of uncertainty on the detection probabilities p_r. Finally, we shall introduce also the state process modelling N_r.

At the first level, we shall consider the detection probability p_r as an external parameter taken e.g. from the national telecommunication regulator (not really the case). The posterior probability distribution for N_r in terms of N_r^{net} and p_r will be given by

$$\mathbb{P}\left(N_r | N_r^{\text{net}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}(N_r - N_r^{\text{net}}; 1 - p_r, N_r^{\text{net}} + 1) & \text{if } N_r \geq N_r^{\text{net}}, \end{cases}$$

where $\text{negbin}(k;p,r) \equiv \binom{k+r-1}{k} p^k (1-p)^r$ denotes the probability mass function of a negative binomial random variable of values $k \geq 0$ with parameters p and r. Once we have a distribution, we can provide a point estimator, a posterior variance, a posterior coefficient of variation, a credible interval, and as many indicators as possible computed from the distribution. For example, if we use the MAP criterion (the posterior mode) or the posterior mean we can provide as point estimators

$$\widehat{N}_r^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{(1-p_r)\cdot N_r^{\text{net}}}{p_r} \right\rfloor, \tag{22a}$$

$$\widehat{N}_r^{\text{mean}} = N_r^{\text{net}} + \frac{(1-p_r)\cdot (N_r^{\text{net}}+1)}{p_r}. \tag{22b}$$

Let us now introduce the second level focused on the uncertainty in the detection probability p_r. A priori, we can think of a detection probability p_{kr} per individual k in the target population and try to device some model to estimate p_{kr} in terms of auxiliary information (e.g. sociodemographic variables, income, etc.). We would need subscription information related to these variables for the whole target population, which is unattainable. Instead, we may consider that the detection probability p_{kr} shows a common part for all individuals in region r plus some additional unknown terms, i.e. something like $p_{kr} = p_r + \text{noise}$. At a first stage, we propose to implement this idea by modeling $p_r \simeq \text{Beta}(\alpha_r, \beta_r)$ and choosing the hyperparameters α_r and β_r according to the penetration rates P_r^{net} and the register-based population figures N_r^{reg}.

Notice that the penetration rate is also subjected to the problem of device duplicities (individuals having two or more devices). To deduplicate, we make use of the duplicity probabilities $p_d^{(i)}$ computed in Sect. 4 under the same assumptions (at most two devices per individual) and of the posterior location probabilities $\bar{\gamma}_{dr}$ in region r for each device d. Notice that we have also dropped out the time subscript for ease of notation, since we are currently focusing on the initial time t_0. We define

$$\Omega_r^{(1)} = \frac{\sum_{d=1}^{D} \bar{\gamma}_{dr}\cdot p_d^{(1)}}{\sum_{d=1}^{D} \bar{\gamma}_{dr}}, \tag{23a}$$

$$\Omega_r^{(2)} = \frac{\sum_{d=1}^{D} \bar{\gamma}_{dr}\cdot p_d^{(2)}}{\sum_{d=1}^{D} \bar{\gamma}_{dr}}. \tag{23b}$$

The deduplicated penetration rates are defined as

$$\tilde{P}_r^{\text{net}} = \left(\Omega_r^{(1)} + \frac{\Omega_r^{(2)}}{2}\right)\cdot P_r^{\text{net}}. \tag{23c}$$

To get a feeling on this definition, let us consider a very simple situation. Let us consider $N_r^{(1)} = 10$ individuals in region r with 1 device each one, $N_r^{(2)} = 3$ individuals in region r with 2 devices each one, and $N_r^{(0)} = 2$ individuals in region r with no device. Let us assume that we can measure the penetration rate with certainty, so that $P_r^{\text{rm}} = \frac{16}{15}$. The devices are assumed to be neatly detected by the HMM (i.e. $\tilde{\gamma}_{dr} = 1 - O(\epsilon)$) and duplicities are also inferred correctly ($p_d^{(2)} = O(\epsilon)$ for $d^{(1)}$ and $p_d^{(2)} = 1 - O(\epsilon)$ for $d^{(2)}$). Then $\Omega_r^{(1)} = \frac{10}{16} + O(\epsilon)$ and $\Omega_r^{(2)} = \frac{6}{16} + O(\epsilon)$. The deduplicated penetration rate will then be $\tilde{P}_r^{\text{net}} = \frac{13}{15} + O(\epsilon)$, which

can be straightforwardly understood as a detection probability for an individual in this network in region r.

Let us now denote by N_r^{reg} the population of region r according to an external population register. Then, we fix

$$\alpha_r + \beta_r = N_r^{\text{reg}}, \tag{24a}$$

$$\frac{\alpha_r}{\alpha_r + \beta_r} = \tilde{P}_r^{\text{net}}, \tag{24b}$$

which immediately implies that

$$\alpha_r = \tilde{P}_r^{\text{net}} \cdot N_r^{\text{reg}}, \tag{25a}$$

$$\beta_r = \left(1 - \tilde{P}_r^{\text{net}}\right) \cdot N_r^{\text{reg}}. \tag{25b}$$

There are several assumptions in this choice. Firstly, on average, we assume that detection takes place with probability $\tilde{P}_r^{\text{net}}$. We find this assumption reasonable. Another alternative choice would be to use the mode of the beta distribution instead of the mean. Secondly, detection is undertaken over the register-based population. We assume some coherence between the official population count and the network population count. A cautious reader may object that we do not need a network-based estimate if we already have official data at the same time instant. We can make several comments in this regard:

- As stated above, a degree of coherence between official estimates by combining data sources to conduct more accurate estimates is desirable. By using register-based population counts in the hierarchy of models, we are indeed combining both data sources. In this combination notice, however, that the register-based population is taken as an external input in our model. There exist alternative procedures in which all data sources are combined at an equal footing [64, 65]. We deliberately use the register-based population as an external source and do not intend to re-estimate it by combination with mobile network data.
- Register-based populations and network-based populations show clearly different time scales. The coherence we demand will be forced only at the given initial time t_0 after which the dynamics of the network will provide the time scale of the network-based population counts without further reference to the register-based population.

Thirdly, the penetration rates P_r^{net} and the official population counts N_r^{reg} come without error. Should this not be attainable or realistic, we would need to introduce a new hierarchy level to account for this uncertainty (see below). Lastly, the deduplicated penetration rates are computed as a deterministic procedure (using a mean point estimation), i.e. the deduplicated penetration rates are also subjected to uncertainty, thus we should also introduce another hierarchy level to account for this uncertainty.

Then, we can readily compute the posterior distribution for N_r:

$$\mathbb{P}\left(N_r | N_r^{\text{net}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{betaNegBin}(N_r - N_r^{\text{net}}; N_r^{\text{net}} + 1, \alpha_r - 1, \beta_r) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases} \tag{26}$$

It is a displaced beta negative binomial distribution

$$\text{betaNegBin}(k; s, \alpha, \beta) \equiv \frac{\Gamma(k+s)}{k!\Gamma(s)} \frac{\text{B}(\alpha+s, \beta+k)}{\text{B}(\alpha, \beta)}$$

with support in $N_r \geq N_r^{\text{net}}$ and parameters $s = N_r^{\text{net}} + 1$, $\alpha = \alpha_r - 1$ and $\beta = \beta_r$. Again, we can provide point estimates as well as posterior variances, credible intervals, etc. Under the MAP and the mean criterion we have

$$\widehat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{(1-\tilde{P}_r^{\text{net}}) \cdot N_r^{\text{net}}}{\tilde{P}_r^{\text{net}}} - \frac{N_r^{\text{net}}}{N_r^{\text{reg}} \cdot \tilde{P}_r^{\text{reg}}} \right\rfloor,$$

$$\widehat{N}^{\text{mean}} = N_r^{\text{net}} + \frac{(N_r^{\text{net}}+1) \cdot (1-\tilde{P}_r^{\text{net}}) \cdot N_r^{\text{reg}}}{\tilde{P}_r^{\text{reg}} \cdot N_r^{\text{reg}} - 1}.$$

The uncertainty is accounted for by computing the posterior variance, the posterior coefficient of variation, or credible intervals in the usual way. Notice that when $\alpha_r, \beta_r \gg 1$ (i.e., when $\min(\tilde{P}_r^{\text{net}}, 1-\tilde{P}_r^{\text{net}}) \cdot N_r^{\text{reg}} \gg 1$) the beta negative binomial distribution (26) reduces to the negative binomial distribution

$$\mathbb{P}\big(N_r | N_r^{\text{net}}\big) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r+\beta_r-1}, N_r^{\text{net}}+1) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases} \tag{28}$$

Note also that $\frac{\beta_r}{\alpha_r+\beta_r-1} \approx 1 - \tilde{P}_r^{\text{net}}$ so that in this case we do not need the register-based population (this is similar to dropping out the finite population correction factor in sampling theory for large populations). In this case, under the MAP and the mean criterion for this distribution we have

$$\widehat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{(1-\tilde{P}_r^{\text{net}})}{\tilde{P}_r^{\text{net}}} \cdot N_r^{\text{net}} \right\rfloor,$$

$$\widehat{N}^{\text{mean}} = N_r^{\text{net}} + \frac{(1-\tilde{P}_r^{\text{net}})}{\tilde{P}_r^{\text{net}}} \cdot \big(N_r^{\text{net}}+1\big).$$

Let us now introduce a further level of uncertainty by modelling also the hyperparameters (α_r, β_r) so that the relationship between these parameters and the external data sources (penetration rates and register-based population counts) is also uncertain. We can go all the way down the hierarchy, assume a cross-cutting relationship between parameters and some hyperparameters and postulate

$$N_r^{\text{net}} \simeq \text{Bin}(N_r, p_r), \quad \text{for all } r = 1, \ldots, R, \tag{29a}$$

$$p_r \simeq \text{Beta}(\alpha_r, \beta_r), \quad \text{for all } r = 1, \ldots, R, \tag{29b}$$

$$\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r+\beta_r}\right), \alpha_r + \beta_r\right)$$
$$\simeq \text{N}\big(\mu_{\gamma r}\big(\gamma_0, \gamma_1; \tilde{P}_r^{\text{net}}\big), \tau_\gamma^2\big) \times \text{Gamma}\left(1+\xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1, \ldots, R, \tag{29c}$$

$$\big(\log \gamma_0, \gamma_1, \tau_\gamma^2, \xi\big) \simeq \text{f}_\gamma\big(\log \gamma_0, \gamma_1, \tau_\gamma^2\big) \times \text{f}_\xi(\xi), \tag{29d}$$

where we have denoted $\mu_{\gamma r}(\gamma_0, \gamma_1; \bar{P}_r^{\text{net}}) \equiv \log(\gamma_0[\frac{\bar{P}_r^{\text{net}}}{1-\bar{P}_r^{\text{net}}}]^{\gamma_1})$ and f_γ and f_ξ stand for prior distributions.

The interpretation of this hierarchy is simple. It is just a beta-binomial model in which the beta parameters α_r, β_r are correlated with the deduplicated penetration rates. This correlation is expressed through a linear regression model upon their logits with common regression parameters across the regions, both the coefficients and the uncertainty degree. On average, the detection probabilities p_r will be the deduplicated penetration rates with uncertainty accounted for by hyperparameters γ_0, γ_1, τ_γ^2. For large population cells, the hyperparameter ξ drops out so that finally the register-based population counts N_r^{reg} play no role in the model.

Under the specifications (29a)–(29d), after some tedious computations, we can show that the multivariate distribution for the number of individuals $\mathbf{N}$ in the target population conditional on the number of individuals $\mathbf{N}^{\text{net}}$ detected by the network is given by a continuous mixture:

$$\mathbb{P}\big(\mathbf{N}|\mathbf{N}^{\text{net}}\big) \propto \int_{\mathbb{R}^R} d^R\mathbf{y}\,\omega_{\text{obs}}\big(\mathbf{y};\bar{\mathbf{P}}^{\text{net}}\big)\prod_{r=1}^{R}\frac{\text{negbin}(N_r - N_r^{\text{net}}; 1-p(y_r), N_r^{\text{net}}+1)}{p(y_r)}, \tag{30}$$

where

- $\text{negbin}(k;p,r)$ stands for the probability mass function of the negative binomial distribution for variable k and parameters p and r;
- $p(y_r) \equiv \frac{e^{y_r}}{1+e^{y_r}}$;
- $\omega_{\text{obs}}(\mathbf{y};\mathbf{P}^{\text{net}}) = \int_{\Omega_\beta} \mathrm{d}\log\gamma_0\,\mathrm{d}\gamma_1\,\mathrm{d}\tau_\gamma^2 \mathrm{f}_\gamma(\log\gamma_0,\gamma_1,\tau_\gamma^2)\mathrm{n}(\mathbf{y};\boldsymbol{\mu}_\gamma(\gamma_0,\gamma_1;\bar{\mathbf{P}}^{\text{net}}),\boldsymbol{\Sigma}_\gamma)$ where
  - $\mathrm{n}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ stands for the probability density function of the multivariate normal distribution for variable $\mathbf{x}$ and mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
  - $\mu_{\gamma r}(\gamma_0,\gamma_1;\bar{P}_r^{\text{net}}) = \log(\gamma_0[\frac{\bar{P}_r^{\text{net}}}{1-\bar{P}_r^{\text{net}}}]^{\gamma_1})$.
  - $\boldsymbol{\Sigma}_\gamma = \tau_\gamma^2 \mathbb{I}_{R\times R}$.

In this derivation, again the assumption $\alpha_r, \beta_r \gg 1$ is taken for granted. In rigour, we should have included $\mathbf{P}^{\text{net}}$ as conditioning random variables together with $\mathbf{N}^{\text{net}}$, but we have opted to keep the notation as simple as possible. To have an expression which can be computed we need to further specify the prior f_γ. As a first example, let us consider $\gamma_0 = \gamma_1 = 1$ and $\tau_\gamma^2 \to 0^+$. This amounts to having certainty about the values of α_r and β_r, as above, so that $\omega_{\text{obs}}(\mathbf{y};\bar{\mathbf{P}}^{\text{net}}) = \prod_{r=1}^R \delta(y_y - \log\bar{P}_r^{\text{net}})$, where $\delta(\cdot)$ stands for the Dirac delta function. Upon normalization expression (30) reduces to

$$\mathbb{P}\big(\mathbf{N}|\mathbf{N}^{\text{net}}\big) = \prod_{r=1}^{R}\text{negbin}\big(N_r - N_r^{\text{net}}; 1-\tilde{P}_r^{\text{net}}, N_r^{\text{net}}+1\big). \tag{31}$$

The marginal distribution for region r reduces to (28), which was also obtained above through a direct reasoning.

Finally, we can also introduce the state process. The system is a human population and we can make a common modelling hypothesis to represent the number of individuals N_r in region r of the target population as a Poisson-distributed random variable in terms of the population density, i.e.

$$N_r \simeq \text{Poisson}(A_r\sigma_r), \tag{32}$$

where σ_r stands for the population density of region r and A_r denotes the area of region r. We choose to model N_r in terms of the population density to make an auxiliary usage of some results already found in the literature [6]. Similarly to the observation process, we introduce the following hierarchy:

$$N_r^{\text{net}} \simeq \text{Bin}(N_r, p_r), \quad \text{for all } r = 1, \ldots, R, \tag{33a}$$

$$N_r \simeq \text{Poisson}(A_r \sigma_r), \quad \text{for all } r = 1, \ldots, R, \tag{33b}$$

$$p_r \simeq \text{Beta}(\alpha_r, \beta_r), \quad \text{for all } r = 1, \ldots, R, \tag{33c}$$

$$\sigma_r \simeq \text{Gamma}(1 + \zeta_r, \theta_r), \quad \text{for all } r = 1, \ldots, R, \tag{33d}$$

where the hyperparameters will express the uncertainty about the register-based population and the detection probability. The values for α_r and β_r are taken from (25a)–(25b). Regarding the hyperparameters θ_r and ζ_r, notice that the modes of the gamma distributions are at $\tau_r = \zeta_r \cdot \theta_r$ and the variances are given by $\mathbb{V}(\tau_r) = (\zeta_r + 1) \cdot \theta_r^2$. We shall parameterise these gamma distributions in terms of the register-based population densities σ_r^{reg} as

$$\begin{aligned} \zeta_r \cdot \theta_r &= \sigma_r^{\text{reg}} + \Delta\sigma_r, \\ \sqrt{(\zeta_r + 1) \cdot \theta_r^2} &= \epsilon_r \cdot \sigma_r^{\text{reg}}, \end{aligned}$$

where ϵ_r can be viewed as the coefficient of variation for σ_r^{reg} and $\Delta\sigma_r$ can be interpreted as the bias for σ_r^{reg}. This parametrization implies that

$$\begin{aligned} \theta_r(\Delta\sigma_r, \epsilon_r) &= \frac{\sigma_r^{\text{reg}}}{2}\left(1 + \frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}\right)\left[\sqrt{1 + \left(\frac{2\epsilon_r}{1 + \frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}}\right)^2} - 1\right], \\ \zeta_r(\Delta\sigma_r, \epsilon_r) &= \frac{2}{\sqrt{1 + (\frac{2\epsilon_r}{1 + \frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}})^2} - 1}. \end{aligned} \tag{34}$$

Under assumptions (33a)–(33d) and assuming $\alpha_r, \beta_r \gg 1$, as above, we get

$$\mathbf{P}\big(\mathbf{N} | \mathbf{N}^{\text{net}}\big) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r + \beta_r} \cdot Q(\theta_r), N_r^{\text{net}} + 1 + \zeta_r\right), \tag{35}$$

where $Q(\theta_r) \equiv \frac{A_r\theta_r}{1 + A_r\theta_r}$. The interpretation of this hierarchy is also simple. It is just a Poisson-gamma model in which the gamma parameters have been chosen so that we account for the uncertainty in the register-based population figures N_r^{reg}. Usual point estimators are easily derived from (35):

$$\begin{aligned} \widehat{N}_r^{\text{MAP}} &= N_r^{\text{net}} + \left\lfloor \frac{(1 - \tilde{P}_r^{\text{net}}) \cdot Q(\theta_r)}{1 - (1 - \tilde{P}_r^{\text{net}}) \cdot Q(\theta_r)} \big(N_r^{\text{net}} + \zeta_r\big) \right\rfloor, \\ \widehat{N}_r^{\text{mean}} &= N_r^{\text{net}} + \frac{(1 - \tilde{P}_r^{\text{net}}) \cdot Q(\theta_r)}{1 - (1 - \tilde{P}_r^{\text{net}}) \cdot Q(\theta_r)} \cdot \big(N_r^{\text{net}} + 1 + \zeta_r\big). \end{aligned}$$

Accuracy indicators such as posterior variance or credible intervals are computed from the distribution (35) as usual. Expression (35) contains the uncertainty of both the observation and the state processes. In the limiting case $\epsilon_r^+ \to 0$ and $\Delta\sigma_r \to 0$, i.e. having certainty about the state process, and with equations (25a)–(25b), we have the Poisson limit of the negative binomial distribution so that

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{poisson}\left(N_r - N_r^{\text{net}}; \left(1 - \bar{P}_r^{\text{net}}\right) \cdot A_r \sigma_r^{\text{reg}}\right). \tag{36}$$

The MAP estimator is trivially $\hat{N}^{\text{MAP}} = N_r^{\text{net}} + \lfloor (1 - \bar{P}_r) A_r \sigma_r^{\text{reg}} \rfloor$ and the mean estimator is trivially $\hat{N}^{\text{mean}} = N_r^{\text{net}} + (1 - \bar{P}_r) A_r \sigma_r^{\text{reg}}$, both of which can be readily read as the sum of the individuals detected by the network and the individuals not detected by the network accounted for by the population register.

On the contrary, when $\epsilon_r \to \infty$ (i.e. having no information at all about the state process), we have $Q(\theta_r) = 1$ and $\zeta_r = 0$ so that

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; 1 - \bar{P}_r, N_r^{\text{net}} + 1\right), \tag{37}$$

which is the same expression as (31), as expected, since having no information about the state process is equivalent to having only the observation process. Notice that we can also introduce more levels in the hierarchy regarding the state process:

$$N_r^{\text{net}} \simeq \text{Binomial}(N_r, p_r), \quad \text{for all } r = 1, \ldots, R, \tag{38a}$$

$$N_r \simeq \text{Poisson}(A_r \sigma_r), \quad \text{for all } r = 1, \ldots, R, \tag{38b}$$

$$p_r \simeq \text{Beta}(\alpha_r, \beta_r), \quad \text{for all } r = 1, \ldots, R, \tag{38c}$$

$$\sigma_r \simeq \text{Gamma}\left(\zeta + 1, \frac{e^{\theta_r}}{\zeta}\right), \quad \text{for all } r = 1, \ldots, R, \tag{38d}$$

$$\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r + \beta_r}\right), \alpha_r + \beta_r\right)$$

$$\simeq \text{N}\left(\mu_{\gamma r}\left(\gamma_0, \gamma_1; \bar{P}_r^{\text{net}}\right), \tau_\gamma^2\right) \times \text{Gamma}\left(1 + \xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1, \ldots, R, \tag{38e}$$

$$\theta_r \simeq \text{N}\left(\mu_{\delta r}\left(\delta_0, \delta_1; \sigma_r^{\text{reg}}\right), \tau_\delta^2\right), \quad \text{for all } r = 1, \ldots, R, \tag{38f}$$

$$\left(\log \gamma_0, \gamma_1, \tau_\gamma^2, \xi\right) \simeq \text{f}_\gamma\left(\log \gamma_0, \gamma_1, \tau_\gamma^2\right) \times \text{f}_\xi(\xi) \tag{38g}$$

$$\left(\log \delta_0, \delta_1, \delta_\delta^2, \zeta\right) \simeq \text{f}_\delta\left(\log \delta_0, \delta_1, \delta_\delta^2\right) \times \text{f}_\zeta(\zeta), \tag{38h}$$

where we have denoted $\mu_{\delta r}(\delta_0, \delta_1; \sigma_r^{\text{reg}}) \equiv \log(\delta_0 [\sigma_r^{\text{reg}}]^{\delta_1})$ and $f_\gamma, f_\xi, f_\delta, f_\zeta$ stand for prior distributions.

The interpretation of this hierarchy is also simple. It is just a combined beta-binomial and Poisson-gamma model in which the gamma parameters have been chosen so that the mode is at $\exp(\theta_r)$ with an uncertainty degree provided by ζ. Notice that the smaller ζ, the more degree of uncertainty about the value of θ_r. The mode is correlated with the register-based population density σ_r^{net} through a linear regression.

Under the specifications (38a)–(38h), again after some tedious computation, we can show that the multivariate distribution for the number of individuals $\mathbf{N}$ in the target population conditional on the number of individuals $\mathbf{N}^{\text{net}}$ detected by the network is given by

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) \propto \int_{\mathbb{R}^R} d^R\mathbf{y}\,\omega_{\text{obs}}\left(\mathbf{y};\bar{\mathbf{P}}^{\text{net}}\right)\prod_{r=1}^{R}\frac{\text{negbin}(N_r - N_r^{\text{net}};1-p(y_r),N_r^{\text{net}}+1,)}{p(y_r)}$$
$$\times \int_{\mathbb{R}^R} d^R\mathbf{z}\,\omega_{\text{state}}\left(\mathbf{z};\boldsymbol{\sigma}^{\text{reg}}\right)\prod_{r=1}^{R}\text{negbin}\left(N_r;q\left(\frac{A_r e^{z_r}}{\zeta}\right),1+\zeta\right), \tag{39}$$

where

- $\text{negbin}(k;p,r)$ stands for the probability mass function of the negative binomial distribution for variable k and parameters p and r;
- $p(y_r) \equiv \frac{e^{y_r}}{1+e^{y_r}}$;
- $\omega_{\text{obs}}(\mathbf{y};\mathbf{P}^{\text{net}}) = \int_{\Omega_\gamma} \mathrm{d}\log\gamma_0\,\mathrm{d}\gamma_1\,\mathrm{d}\tau_\gamma^2 \mathrm{f}_\gamma(\log\gamma_0,\gamma_1,\tau_\gamma^2)\mathrm{n}(\mathbf{y};\boldsymbol{\mu}_\gamma(\gamma_0,\gamma_1;\bar{\mathbf{P}}^{\text{net}}),\boldsymbol{\Sigma}_\gamma)$ where
  - $\mathrm{n}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ stands for the probability density function of the multivariate normal distribution for variable $\mathbf{x}$ and mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.
  - $\mu_{\gamma r}(\gamma_0,\gamma_1;\bar{P}_r^{\text{net}}) = \log(\gamma_0[\frac{\bar{P}_r^{\text{net}}}{1-\bar{P}_r^{\text{net}}}]^{\gamma_1})$.
  - $\boldsymbol{\Sigma}_\gamma = \tau_\gamma^2\mathbb{I}_{R\times R}$;
- $q(\frac{A_r e^{z_r}}{\zeta}) \equiv \frac{\frac{A_r e^{z_r}}{\zeta}}{1+\frac{A_r e^{z_r}}{\zeta}}$;
- $\omega_{\text{state}}(\mathbf{z};\boldsymbol{\sigma}^{\text{reg}}) = \int_{\Omega_{\delta,\zeta}} \mathrm{d}\log\delta_0\,\mathrm{d}\delta_1\,\mathrm{d}\delta_\delta^2\,\mathrm{d}\zeta\,\mathrm{f}_\delta(\log\delta_0,\delta_1,\delta_\delta^2)\times\mathrm{f}_\zeta(\zeta)n(\mathbf{z};\boldsymbol{\mu}_\delta(\delta_0,\delta_1;\boldsymbol{\sigma}^{\text{net}}),\boldsymbol{\Sigma}_\delta)$ with
  - $\mathrm{n}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ stands for the probability density function of the multivariate normal distribution for variable $\mathbf{x}$ and mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.
  - $\mu_{\delta r}(\delta_0,\delta_1;\sigma_r^{\text{reg}}) = \log(\delta_0[\sigma_r^{\text{reg}}]^{\delta_1})$.
  - $\boldsymbol{\Sigma}_\delta = \tau_\delta^2\mathbb{I}_{R\times R}$.

Notice how this expression reveals both factors arising from the observation and the state processes, respectively. When $\gamma_0,\gamma_1,\delta_0,\delta_1 \to 1$, $\zeta \to \zeta^*$, and $\tau_\gamma^2,\tau_\delta^2 \to 0^+$ (i.e. when having fully accurate information about the parameters α_r, β_r and θ_r), we have $\omega_\gamma(\mathbf{y}) = \delta(\mathbf{y}-\boldsymbol{\mu}_\gamma)$ and $\omega_\delta(\mathbf{z}) = \delta(\mathbf{z}-\boldsymbol{\mu}_\delta)$ so that after normalization equation (39) reduces to

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R}\text{negbin}\left(N_r - N_r^{\text{net}};(1-\bar{P}_r)\cdot Q_r\left(\zeta^*\right),N_r^{\text{net}}+\zeta^*+1\right), \tag{40}$$

where we have denoted $Q_r(\zeta) \equiv q(\frac{A_r\sigma_r^{\text{reg}}}{\zeta})$, which is indeed again equation (35).

7.2 Present population at times $t > t_0$

Now, we propose to produce probability distributions for the number of individuals N_{tr} in the target population for times $t > t_0$ at region r. Currently, we consider only *closed* populations, i.e. neither individuals nor devices enter into or leave the territory under analysis along the whole time period. This important restriction is posed to introduce progressively the different methods in order to get a thorough assessment of every single aspect of the procedure. It will have to be lifted in future work (e.g. considering sink and source tiles in the reference grid).

Our reasoning tries to introduce as less assumptions as possible. Thus, we begin by considering a balance equation. Let us denote by $N_{t,rs}$ the number of individuals moving from region s to region r in the time interval $(t-1,t)$. Then, we can write

$$\begin{aligned} N_{tr} &= N_{t-1r} + \sum_{\substack{r_t=1 \\ r_t \neq r}}^{N_T} N_{t,rr_t} - \sum_{\substack{r_t=1 \\ r_t \neq r}}^{N_r} N_{t,r_t r} \\ &= \sum_{r_t=1}^{N_T} \tau_{t,rr_t} \cdot N_{t-1r_t}, \end{aligned} \tag{41}$$

where we have defined $\tau_{t,rs} = \frac{N_{t,rs}}{N_{t-1s}}$ (0 if $N_{t-1s} = 0$). Notice that $\tau_{t,rs}$ can be interpreted as an aggregate transition probability from region s to region r at time interval $(t-1,t)$ in the target population.

We make the assumption that individuals detected by the network move across regions in the same way as individuals in the target population. Thus, we can use

$$\tau^{\text{net}}_{t,rs} \equiv \frac{N^{\text{net}}_{t,rs}}{N^{\text{net}}_{t-1s}}$$

to model $\tau_{t,rs}$. In particular, as our first choice we shall postulate $\tau_{t,rs} = \tau^{\text{net}}_{t,rs}$. The probability distributions of N^{net}_{st-1} and $[\mathbf{N}^{\text{net}}_t]_{sr} = N^{\text{net}}_{t,rs}$ were indeed already computed in the aggregation module (Sect. 6).

Finally, we mention two points. On the one hand, random variables N_{rt} are defined recursively in the time index t, so that once we have computed the probability distribution at time t_0, then we can use (41) to compute the probability distribution at later times $t > t_0$. On the other hand, Monte Carlo techniques should be again used to build these probability distributions. Once we have probability distributions, we can make point estimations and compute accuracy indicators as above (posterior variance, posterior coefficient of variation, credible intervals).

7.3 Origin-destination matrices

The inference of the origin-destination matrices for the target population is more delicate than the present population because auxiliary information from population registers do not contain this kind of information. Therefore, the statistical models proposed above for the present population estimation cannot be applied. As a first important conclusion we point out that, in our view, National Statistical Plans should start considering what kind of auxiliary information is needed to make a more accurate use of Mobile Network Data and new digital data, in general.

We can provide a simple argument extending the above model to produce credible intervals for the origin-destination matrices. If N_{tr} and $\tau_{t,rs}$ denote the number of individuals of the target population at time t in region r and the aggregate transition probability from region s to region r at the time interval $(t-1,t)$, then we can simply define $N_{t,rs} = N_{t-1s} \times \tau_{t,rs}$ and trivially build the origin-destination matrix for each time interval $(t-1,t)$. Under the same general assumption as before, if individuals are to move across the geographical territory independently of their mobile network operator (or even not being a sub-

scriber or carrying two devices), we can postulate as a first simple choice $\tau_{t,rs} = \tau^{\text{net}}_{t,rs}$, as before.

7.4 An example with simulated data

Let us again illustrate this approach with the same example generated with the mobile network event simulator. We consider once more the toy scenario with a population of 500 individuals and 186 subscribers with 218 mobile devices in a territory with a bounding box of 10 km $\times$ 10 km divided into 10 regions as in Fig. 10. The simulator provides the true position of each individual at each time instant so that we can make a comparison with the (synthetic) ground truth.

For the time being, we shall only provide results for the posterior distributions (26), (28), and (35), leaving the full hierarchies for future work. Taking advantage of the simulated ground truth we shall provide results taking as prior information different ranges of N^{net} and N^{reg} to better appreciate how errors in the input data affect the final estimates. Firstly, we shall consider values $N^{\text{net}} = (1 + \text{rb}^{\text{net}}) \cdot N^{\text{net0}}$, so that we can investigate the effect of the bias in the input number of individuals detected by the network with respect to their true values N^{net0}. Secondly, similarly, we shall consider values $N^{\text{reg}} = (1 + \text{rb}^{\text{reg}}) \cdot N^{\text{reg0}}$, so that we can investigate the effect of the bias in the input number of individuals according to the population register with respect to their true values N^{reg0}. Finally, for the model with the process (35), we shall also consider the range of values for the coefficient of variation of N^{reg} given by $\text{cv}^{\text{Nnet}} = 0.01, 0.05, 0.10, 0.15, 0.20$. In all cases we shall only use the RSS geolocation model with uniform prior.

In Figs. 14, 15, and 16 we represent the credible intervals for the initial number of individuals for different values of rb^{net} and rb^{reg}. In the case with the process model we have focused on the largest coefficient of variation $\text{cv}^{\text{Nnet}} = 0.2$.

We observe that the uncertainty grows as the bias of the number of individuals according to the population register also grows in the positive direction (overestimation). We can also observe that the uncertainty grows in the same fashion with respect to the bias in the number of individuals detected by the network. The sensitivity in the case of the model with the state process (35) is also evident, thus inviting not to model the state process. That is, if the state process (number of individuals in the target population) is not modelled more robustly, then errors in the register-based population figures will translated into the estimates based on mobile network data. In our view, this is a clear example of how prior hypotheses on the generating model for target variables are dangerous in Official Statistics (historically favouring design-based inference over the model-based approach). Now models need to be used, robustness becomes a priority. Finally, we also see an overestimation effect (intervals displacing upwards) as the biases grow. Further analysis is needed, but in general the computed credible intervals cover the true values fairly accurately.

For the present population at later times and the origin-destination matrices we will see directly in the next section how to integrate all modules to produce final estimates from the initial input data from the telecommunication network.

8 Integration of production modules

Once every module is designed and implemented, we must integrate them all into a production chain. The basic idea is to concatenate them into a sequence so that the output

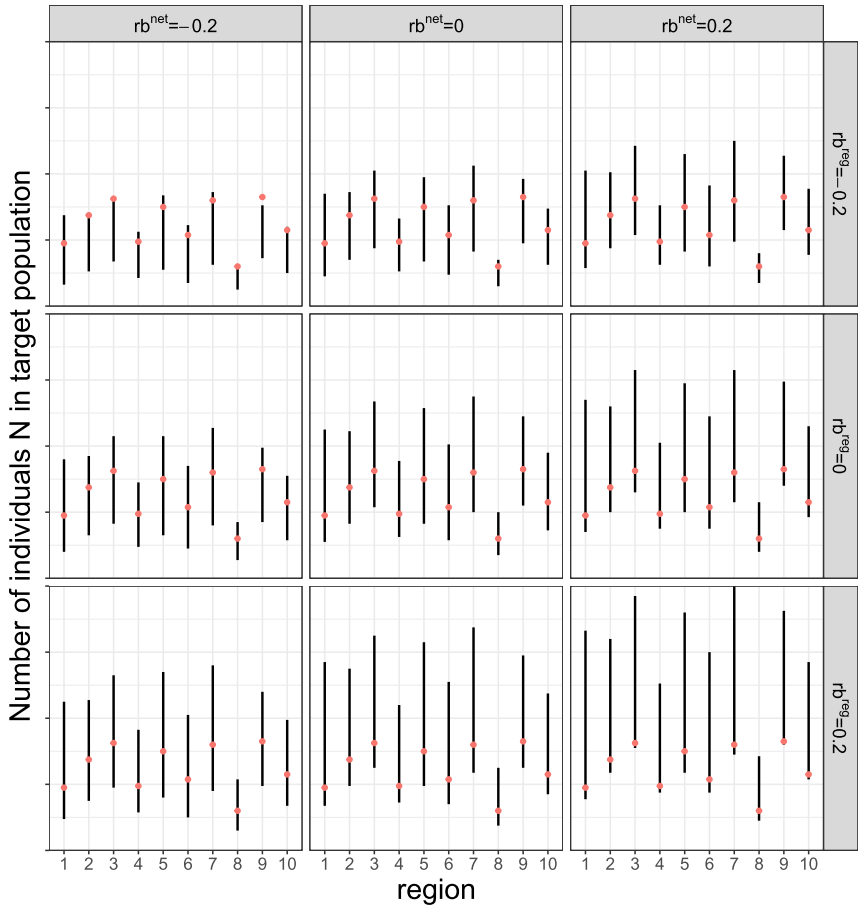

Figure 14 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob = 95%) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the distribution in eq. (26) with the RSS geolocation model and uniform prior and different values of $\mathbf{rb}^{\text{net}}$ and $\mathbf{rb}^{\text{reg}}$

data from each module is the input data for the next. Mathematically, for the present population use case this can be expressed as

$$\mathbb{P}\big(\mathbf{N}_t|\mathbf{E}_{0:T},\mathbf{N}^{\text{reg}},\mathbf{P}^{\text{net}}\big)=\sum_{N_{tr}^{\text{net}}\geq 0}\mathbb{P}\big(\mathbf{N}_t|\mathbf{N}_t^{\text{net}},\mathbf{N}^{\text{reg}},\mathbf{P}^{\text{net}}\big)\mathbb{P}\big(\mathbf{N}_t^{\text{net}}|\mathbf{E}_{0:T}\big). \tag{42}$$

We have computed the credible intervals for the number of individuals in the target population at each time instant t. To carry out the computation we need to specify the geolocation model (together with the HMM prior), the number of individuals according to the population register and the penetration rates. In Figs. 17, 18, and 19 we represent the initial set of credible intervals with the RSS model with uniform prior for different values for the relative bias and the coefficient of variation for the population register figures and the three inference models above (see [50] for an animated gif with the time sequences of credible intervals). Notice that the probability distribution for the number N_{tr}^{net} of individuals detected by the network is computed from the aggregation module.

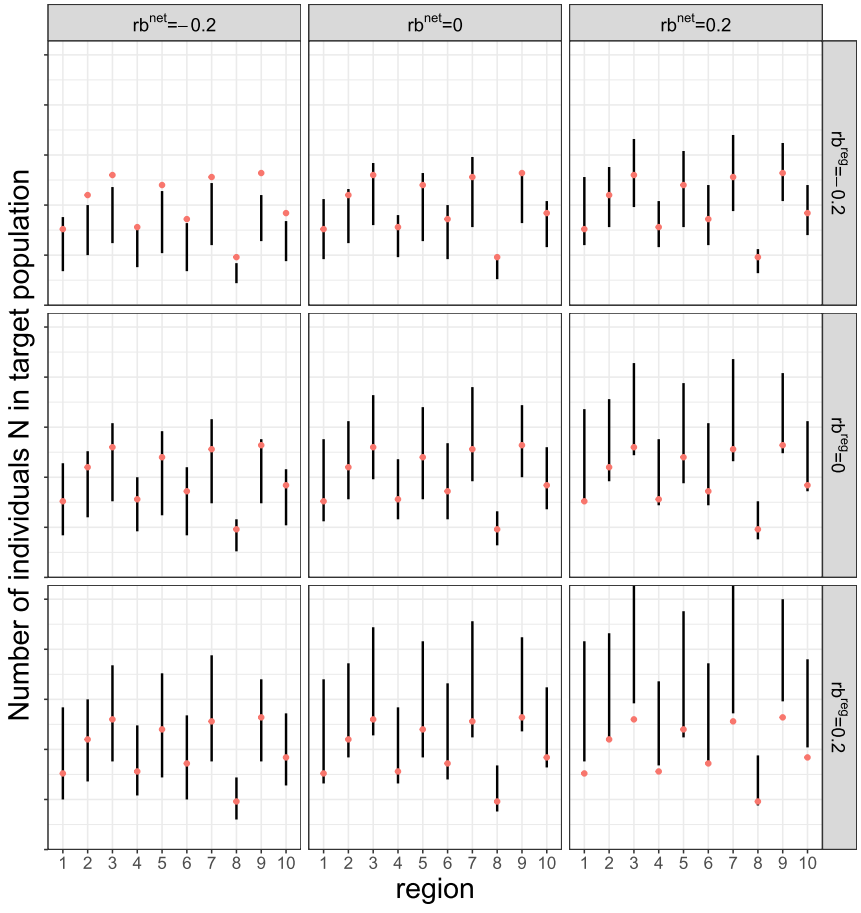

Figure 15 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob = 95%) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the distribution in eq. (28) with the RSS geolocation model and uniform prior and different values of $\mathbf{rb}^{\mathbf{net}}$ and $\mathbf{rb}^{\mathbf{reg}}$

For the origin-destination matrices at times $t > 0$ we apply this same procedure following the methodology described in the preceding section, with the distribution for N_{tr}^{net} and $N_{t,rs}^{\text{net}}$ again computed from the aggregation module. The sequence of origin-destination matrices with the same choices as above is represented in Fig. 20 for $\text{cv}^{\text{reg}} = 0.01$ and $\text{rb}^{\text{reg}} = 0$ and in Fig. 21 for $\text{cv}^{\text{reg}} = 0.20$ and $\text{rb}^{\text{reg}} = 0.20$ for the beta negative binomial inference model (see [50] for the same representation for the negative binomial and negative binomial state process models).

The combination of choices is multiple so that the whole process can be adapted to the complex nature of reality. For our simple scenario we have focused on how to build this modular process. Notice that more sophisticated models can be built in each module, but the whole structure remains the same.

9 Discussion, conclusions and future prospects

The design of a production framework based on mobile network data for Official Statistics is not new. Already Eurostat [66] compiled a body of documents assessing the feasibility of using this data source for producing official tourism statistics promoting research on

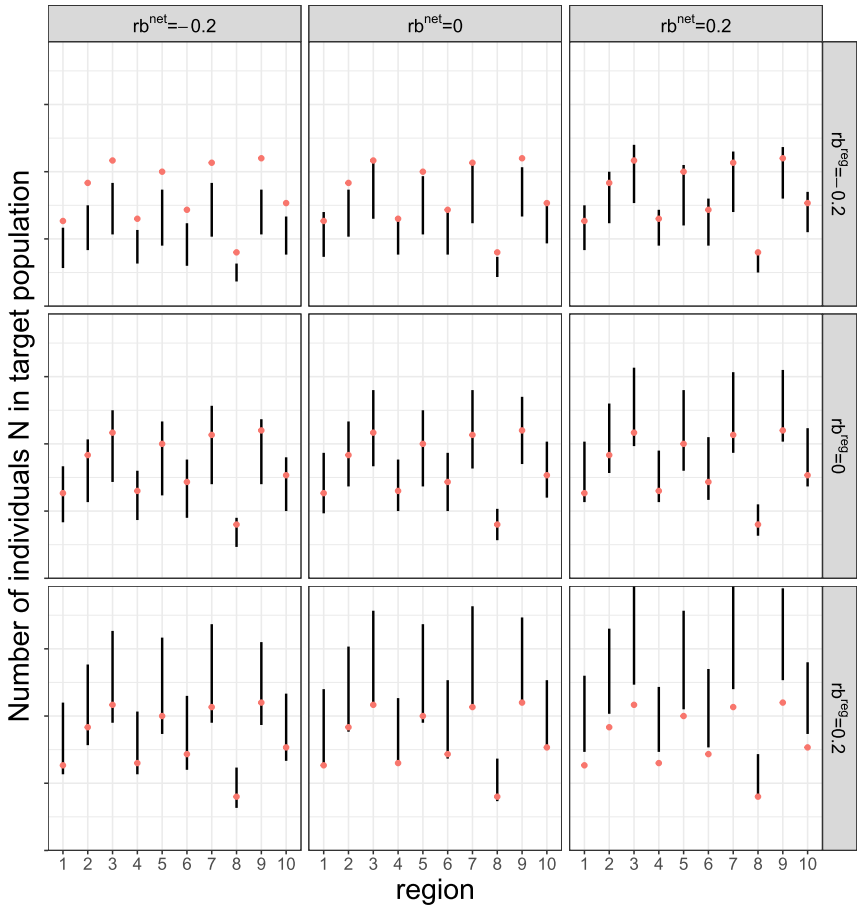

Figure 16 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob = 95%) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the distribution in eq. (35) with the RSS geolocation model and uniform prior and different values of $\mathbf{rb}^{\mathbf{net}}$ and $\mathbf{rb}^{\mathbf{reg}}$ ($\mathbf{cv}^{\mathbf{reg}} = 0.20$)

Big Data through the so-called VIP project ESSnet on Big Data [67]. The United Nations Global Working Group for Big Data [68] published a first version of a handbook to incorporate this data source into the production of official statistics. The need for a methodological framework for multiple statistical domains was clearly identified and proposed for the ESS [28], giving rise to the so-called ESS Reference Methodological Framework for Mobile Network Data. More recently, Saluveer et al. [69] have published their proposal based on empirical experience to produce tourism statistics for the Estonian Central Bank. All these works clearly convince about the value and the feasibility of using these data to gain in relevance, timeliness, and opportunity. In the European context, the international project ESSnet on Big Data [70] has recently taken the first steps towards the detailed construction of this methodological framework dealing with the many aspects covered in preceding sections. The static approach for the geolocation of mobile devices proposed by Tennekes et al. [30] for the present population count estimation is the origin of the current HMM models, which supersedes those by including transition models. A wider scope has been recently proposed and tested with real data by Ricciato et al. [29]. This is again a static approach (no transition model used) and "produce simple point estimates, while

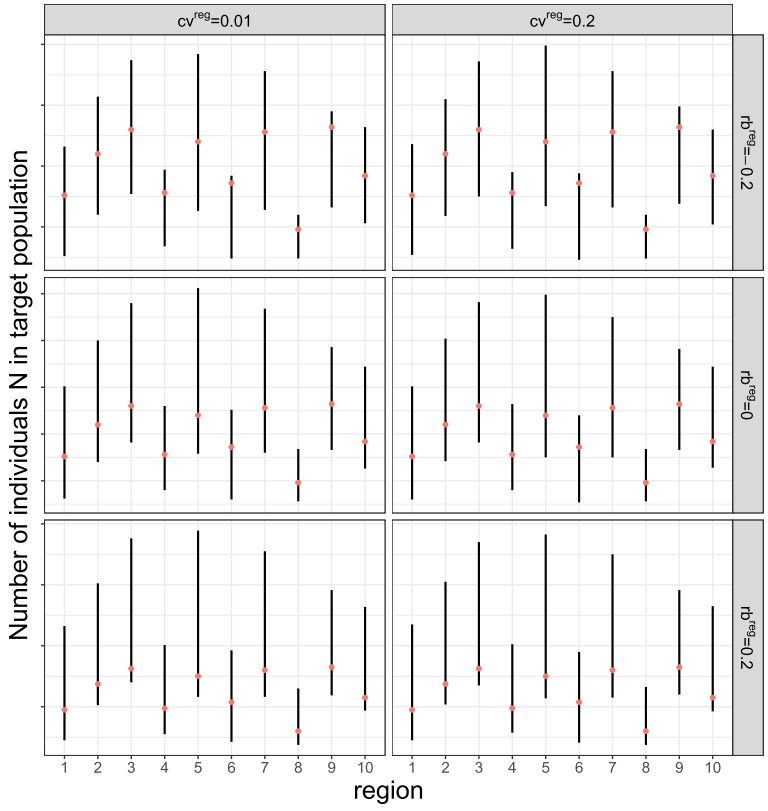

Figure 17 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob = 95%) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the whole module integration using the RSS geolocation model with uniform prior and the beta negative binomial inference model. See animated gif `N_CI_bnb_t.gif` in [50] for the full time sequence of credible intervals and different values of $\mathbf{rb}^{\mathbf{reg}}$ and $\mathbf{cv}^{\mathbf{reg}}$

it would be desirable to develop estimation procedures that deliver also some measure of uncertainty" [29, page 9].

Our process aims at providing a solution for this using HMMs to construct spatio-temporal interpolations and posterior distributions to assess uncertainty by incorporating auxiliary information in a natural way. We do not stick to a concrete HMM but rather on the contrary we propose a generic framework open to different instantiations of HMMs (different state definitions, different emission models, different transition models, etc.). A priori we cannot identify the most favourable specifications for these models and thorough research with different scenarios in the simulator is recommended. Notice that the HMMs already contain the static approach (by setting the transition probability matrix to diagonal) and that even for the simple model proposed above, different strategies can be followed (estimating the parameters for each device independently, for the whole set of devices simultaneously, for a selection of devices, updating parameters at every given time interval, …). The framework is versatile enough to adapt to many different circumstances.

In this line of thought, Ricciato et al. [29] conducted a thorough application of four estimation procedures upon real data and further get numerical results based on synthetic data reaching at least one prominent conclusion for our proposal: "stronger atten-

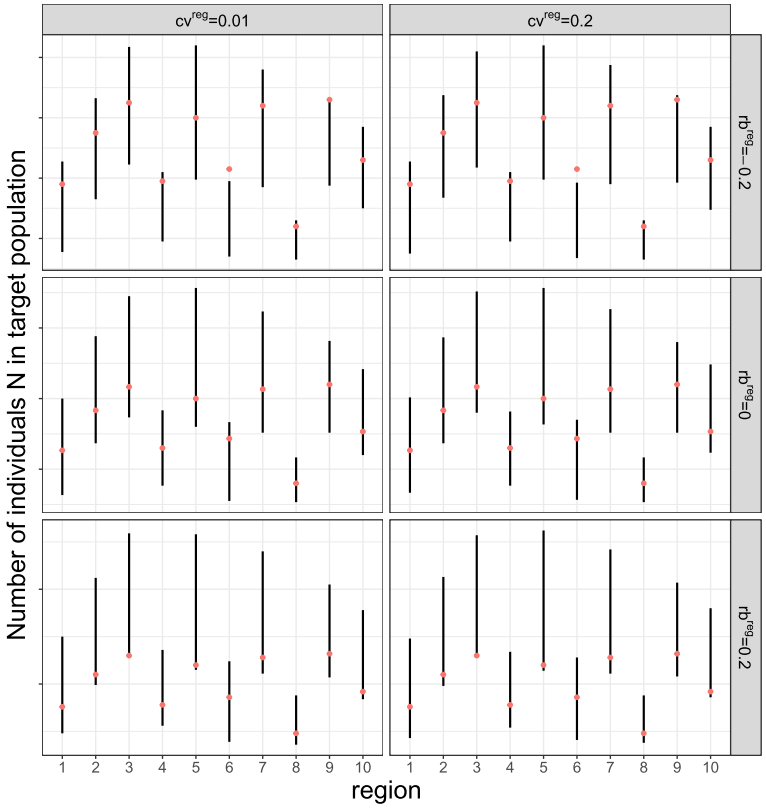

Figure 18 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob = 95%) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the whole module integration using the RSS geolocation model with uniform prior and the negative binomial inference model and different values of $\mathbf{rb}^{\mathbf{reg}}$ and $\mathbf{cv}^{\mathbf{reg}}$

tion should be put by the research community to the issue of *methodological sensitivity*". This is intimately connected to the use of synthetic data incorporating a ground truth to test the methods and to the identification of procedures to assess the goodness of fit of the models, especially those incorporating geospatial considerations (see [29] for details). Regarding *methodological sensitivity*, we notice that our simulator allows the researcher to build different scenarios to investigate the sensitivity of final estimates with regard to multiple parameters (tile size, time range, number of BTS, population size, etc.). Indeed, our simulation exercise in preceding sections have shown how different emission models can be used to analyse the results at different stages of the whole process (geolocation, deduplication, aggregation, and inference – the latter two not shown in the text for brevity's sake). A systematic and thorough sensitivity analysis should be conducted. In our view, these geospatial considerations are fundamental and should be extended beyond the goodness of fit to the inference module as well [71].

To generate the illustrative example included above, apart from the network data event simulator [42], we have developed independent prototyping R packages for each module. Package `destim` for geolocation [72]. Package `deduplication` for deduplicating devices [73]. Package `aggregation` to get the probability distributions of the aggregate number of individuals detected by the network [74]. Package `inference` to get the probability distributions of the aggregate number of individuals in the target pop-

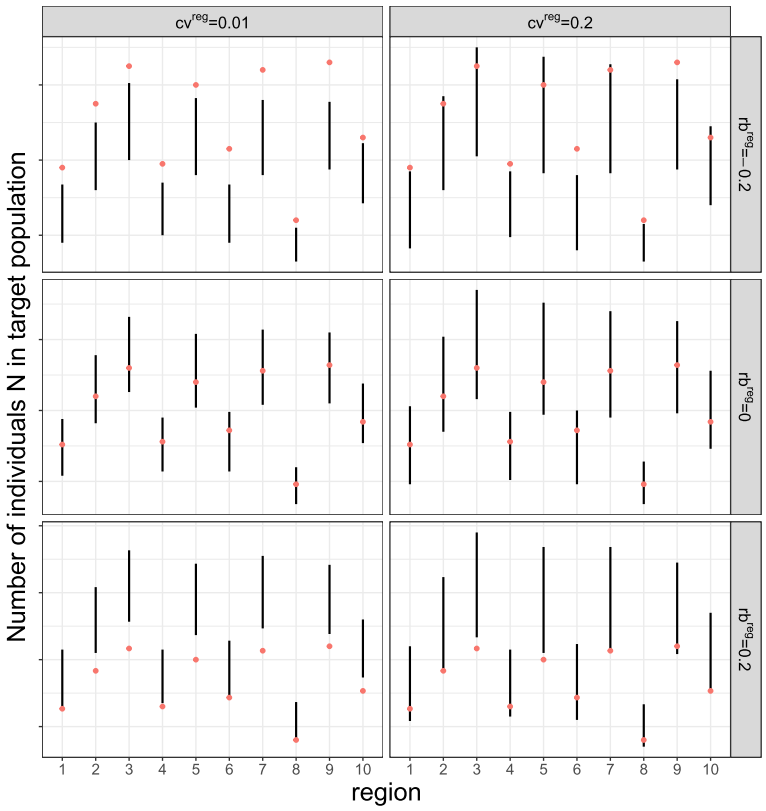

Figure 19 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob = 95%) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the whole module integration using the RSS geolocation model with uniform prior and the negative binomial state process inference model and different values of $\mathbf{rb}^{\mathbf{reg}}$ and $\mathbf{cv}^{\mathbf{reg}}$

ulation [75]. All these packages, although in a prototyping stage, already allow us to apply the methodological proposals above using synthetic data from the simulator or any other real data set with similar contents. Parallelization programming techniques have been applied in preparation for the scalability needed in more realistic scenarios (see supplementary material for some preliminary evaluation of computational complexity).

The whole methodology for the use of mobile network data in official statistical production needs further research and testing. For example, the consideration of a multi-MNO scenario will bring new issues that need to be tackled (data integration, use of privacy-preserving techniques,…). But also the preceding modules should be further developed. The geolocation module uses HMMs, which provide a versatile framework to seek more accurate geolocation either using more complex radio wave propagation models for the emission model and using more complex definitions of the HMM state to account for the transition pattern across the territory. The use of continuous geolocation brings another avenue of research to be further explored beyond the use of a reference grid. The deduplication module can be made more sophisticated accordingly, i.e. in parallel to the geolocation module. The generalization for deduplication of an arbitrary number of devices carried by the same individual needs to be done. The whole statistical filtering needs to be developed with a further stage of the network event data simulator and real data. An

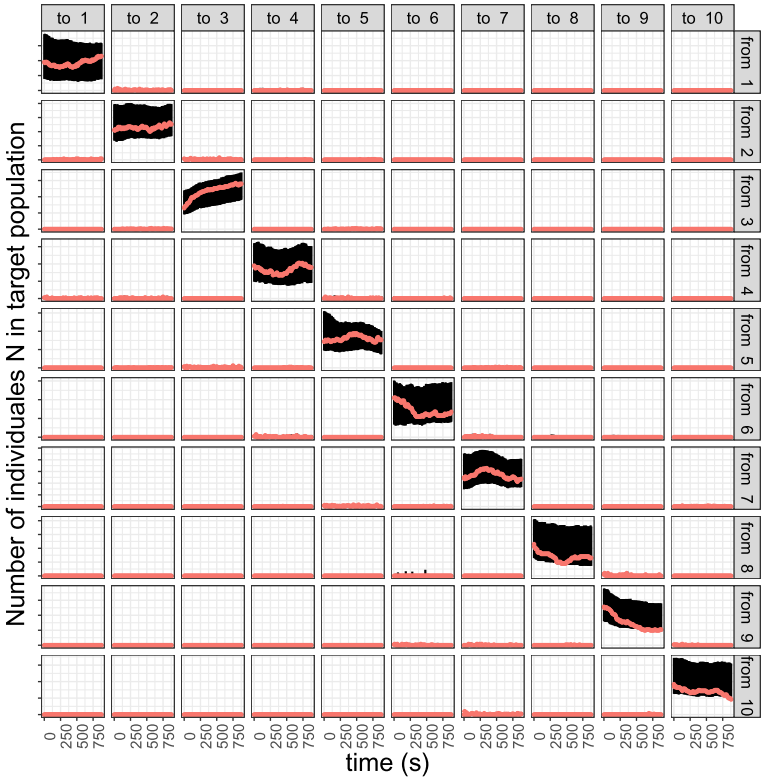

Figure 20 Credible intervals for the OD matrix. Credible intervals (prob = 95%) for OD matrix compared to true values (in red) computed according to the whole module integration with the RSS geolocation model with uniform prior and the beta negative binomial with $\mathbf{rb}^{\mathbf{reg}} = 0$ and $\mathbf{cv}^{\mathbf{reg}} = 0.01$

important new ingredient regarding the identification of devices comprised by the target population is the potential random nature of the number D of devices in our proposal. This would introduce a new level in the hierarchy in which D will be a new integer-valued random variable. The aggregation module should be made more general by comprising any number of deduplicated devices. The inference module deals with the estimation in each region r separately. This should be superseded by a truly multivariate treatment (e.g. using a Dirichlet-multinomial model). Also, spatial correlations should also be considered in the modelling exercise.

In our view, Official Statistics should avoid past errors and struggle for a process-oriented approach to production. Concentrating on statistical domains with an abuse of one-off use cases will bring the risk of growing silos again in the production. In our view, the construction of this process-oriented statistical process with mobile network data should be made in partnerships with MNOs in an international collaborating context clearly identifying those critical elements in the methodology (which data to access and how to process them). The process must be end-to-end so that the whole methodology of the production of official statistics can be openly disseminated. Any form of private-public partnership must involve national and international Data Protection Agencies as a key element in privacy and confidentiality preservation.

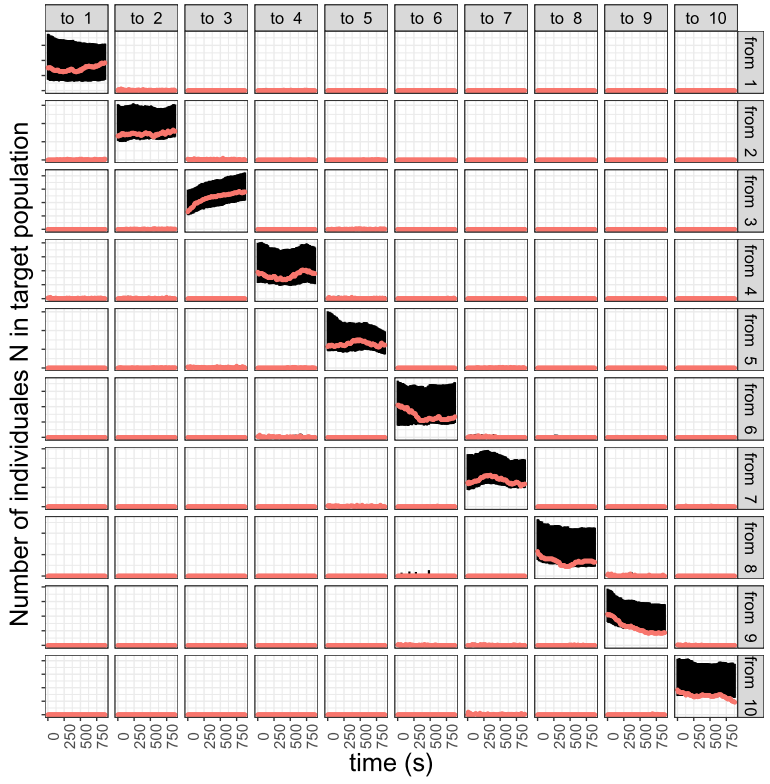

Figure 21 Credible intervals for the OD matrix. Credible intervals (prob = 95%) for OD matrix compared to true values (in red) computed according to the whole module integration with the RSS geolocation model with uniform prior and the beta negative binomial with $\mathbf{rb}^{\mathbf{reg}} = 0.20$ and $\mathbf{cv}^{\mathbf{reg}} = 0.20$

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-021-00275-w.

Additional file 1. The pdf file entitled "Supplementary material for 'An end-to-end statistical process with mobile network data for Official Statistics'" contains extra details about the computation carried out in the main text. Source code for these computations can be visited in the URL specified in the declaration section. (PDF 1.3 MB)

Acknowledgements
The authors acknowledge M.Á. Martínez-Vidal, S. Lorenzo, M. Suárez-Castillo, R. Radini, T. Tuoto, M. Offermans, M. Tennekes, S. Hadam, and F. Ricciato for invaluable insights and debates. We also thank three anonymous referees for their comments to gain readibility.

Funding
This work is part of ongoing projects at Statistics Spain (INE) and Statistics Romania (INS) in joint collaboration with the European Statistical System under Grant Agreement Number 847375-2018-NL-BIGDATA (ESSnet on Big Data II).

Availability of data and materials
Data, scripts, and source code are freely available at https://figshare.com/articles/dataset/_/12861095.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors have contributed equally. All authors read and approved the final manuscript.

Author details
[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Av. de Manoteras, 50-52, Madrid, Spain. [2]Dept. Statistics and Operations Research, Complutense University of Madrid, Plaza de las Ciencias, 3, Madrid,

Spain. [3]Dept. Business Administration, University of Bucharest, 90 Panduri Street, Bucharest, Romania. [4]Dept. Innovative Tools in Official Statistics, Statistics Romania (INS), 16 Libertatii Bvd, Bucharest, Romania.

Publisher's Note

Received: 2 September 2020 Accepted: 19 April 2021 Published online: 29 April 2021

References

1. Miao G, Zander J, Sung W, Slimane SB (2016) Fundamentals of mobile data networks. Cambridge University Press, Cambridge
2. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958
3. Ahas R, Silm S, Järv O, Saluveer E, Tiru M (2010) Using mobile positioning data to model locations meaningful to users of mobile phones. J Urban Technol 17(1):3–27. https://doi.org/10.1080/10630731003597306
4. Phithakkitnukoon S, Smoreda Z, Olivier P (2012) Socio-geography of human mobility: a study using longitudinal mobile phone data. PLoS ONE 7(6):39253. https://doi.org/10.1371/journal.pone.0039253
5. Calabrese F, Diao M, Lorenzo GD, Ferreira J, Ratti C (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp Res, Part C, Emerg Technol 26:301–313. https://doi.org/10.1016/j.trc.2012.09.009
6. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888–15893. https://doi.org/10.1073/pnas.1408439111
7. Louail T, Lenormand M, Ros OGC, Picornell M, Herranz R, Frias-Martinez E, Ramasco JJ, Barthelemy M (2014) From mobile phone data to the spatial structure of cities. Sci Rep 4(1):5276. https://doi.org/10.1038/srep05276
8. Iqbal MS, Choudhury CF, Wang P, González MC (2014) Development of origin–destination matrices using mobile phone call data. Transp Res, Part C, Emerg Technol 40:63–74. https://doi.org/10.1016/j.trc.2014.01.002
9. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4(1):10. https://doi.org/10.1140/epjds/s13688-015-0046-0
10. Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2015) High resolution population estimates from telecommunications data. EPJ Data Sci 4(1):4. https://doi.org/10.1140/epjds/s13688-015-0040-6
11. Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F (2016) An analytical framework to nowcast well-being using mobile phone data. Int J Data Sci Anal 2(1–2):75–92. https://doi.org/10.1007/s41060-016-0013-2
12. Raun J, Ahas R, Tiru M (2016) Measuring tourism destinations using mobile tracking data. Tour Manag 57:202–212. https://doi.org/10.1016/j.tourman.2016.06.006
13. Ricciato F, Widhalm P, Pantisano F, Craglia M (2017) Beyond the "single-operator, CDR-only" paradigm: an interoperable framework for mobile phone network data analyses and population density estimation. Pervasive Mob Comput 35:65–82. https://doi.org/10.1016/j.pmcj.2016.04.009
14. Graells-Garrido E, Caro D, Parra D (2018) Inferring modes of transportation using mobile phone data. EPJ Data Sci 7(1):49. https://doi.org/10.1140/epjds/s13688-018-0177-1
15. Wang Z, He SY, Leung Y (2018) Applying mobile phone data to travel behaviour research: a literature review. Travel Behav Soc 11:141–155. https://doi.org/10.1016/j.tbs.2017.02.005
16. Debusschere M, Sonck J, Skaliotis M (2016) Official statistics and mobile network operator partner up in Belgium. In: OECD statistics newsletter, pp 11–14
17. Williams S (2016) Statistical uses for mobile phone data: literature review. Technical report, Office for National Statistics
18. Nurmi O (2016) Improving the accuracy of outbound tourism statistics with mobile positioning data. In: 15th global forum on tourism statistics, Cusco, Peru
19. Izquierdo-Valverde M, Mascuñano JP, Velasco-Gimeno M (2016) Same-day visitors crossing borders a big and data approach using traffic control. In: 14th global forum on tourism statistics, Venice, Italy
20. Dattilo B, Radini R, Sabato M (2016) How many SIM in your luggage? A strategy to make mobile phone data usable in tourism statistics. In: 14th global forum on tourism statistics
21. Senaeve G, Demunter C (2016) When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics. In: 14th global forum on tourism statistics, Venice, Italy
22. Meersman FD, Seynaeve G, Debusschere M, Lusyne P, Dewitte P, Baeyens Y, Wirthmann A, Demunter C, Reis F, Reuter HI (2016) Assessing the quality and of mobile and phone data as a source of statistics. In: European conference on quality in official statistics (Q2016), Madrid
23. Reis F, Seynaeve G, Wirthmann A, de Meersman F, Debusschere M (2017) Land use classification based on present population daily profiles from a big data source. https://www.conference-service.com/NTTS2017/documents/agenda/data/abstracts/abstract_172.html
24. Sakarovitch B, de Bellefon M-P, Givord P, Vanhoof M (2019) Estimating the residential population from mobile phone data, an initial exploration. Écon. Stat. 505d:109–132. https://doi.org/10.24187/ecostat.2018.505d.1968
25. Galiana L, Sakarovitch B, Smoreda Z (2018) Understanding socio-spatial segregation in French cities with mobile phone data. DGINS18
26. Lestari TK, Esko S, Sarpono SE, Rufiadi R (2018) Indonesia's experience of using signaling mobile positioning data for official tourism statistics. In: 15th world forum on tourism statistics, Cusco, Peru. http://www.15th-tourism-stats-forum.com/papers.html
27. UN (2017) Handbook on the use of mobile phone data for official and statistics
28. Ricciato F (2018) Towards a reference methodological framework for processing MNO data for official statistics. In: 15th world forum on tourism statistics

29. Ricciato F, Lanzieri G, Wirthmann A, Seynaeve G (2020) Towards a methodological framework for estimating present population density from mobile network operator data. Pervasive Mob Comput 68:101263. https://doi.org/10.1016/j.pmcj.2020.101263
30. Tennekes M, Gootzen YAPM, Shah SH (2020) A Bayesian approach to location estimation of mobile devices from mobile network operator data. Resreport, Statistics Netherlands (CBS). https://www.cbs.nl/-/media/_pdf/2020/22/cbds_working_paper_location_estimation.pdf
31. UNECE (2011) Strategic vision of the high-level group for strategic developments in business architecture in statistics. UNECE (Ed.), 59th Plennay session of conference of European statisticians, item 4. High-level group for the modernisation of official statistics. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/1.e.pdf
32. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES (2012) Statistical disclosure control. Wiley, Chichester. https://doi.org/10.1002/9781118348239
33. Templ M (2017) Statistical disclosure control for microdata. Springer, Berlin. https://doi.org/10.1007/978-3-319-50272-4
34. de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3(1):1376. https://doi.org/10.1038/srep01376
35. Xu F, Tu Z, Li Y, Zhang P, Fu X, Jin D (2017) Trajectory recovery from ash. In: Proceedings of the 26th international conference on world wide web. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3038912.3052620
36. UNECE (1992) Fundamental principles of official statistics. Technical report, United Nations. https://www.unece.org/stats/fps.html
37. European Commission (2020) Shaping Europe's digital future. https://ec.europa.eu/digital-single-market/en
38. Salgado D, Oancea B (2020) On new data sources for the production of official statistics. Statistics Spain (INE) working paper. 01/2020. https://www.ine.es/ss/Satellite?blobcol=urldata&blobheader=application/pdf&blobheadername1=Content-Disposition&blobheadervalue1=attachment;filename=art_doctr012020.pdf&blobkey=urldata&blobtable=MungoBlobs&blobwhere=603/210/art_doctr012020.pdf&ssbinary=true
39. Ucar I, Gramaglia M, Fiore M, Smoreda Z, Moro E (2019) Netflix or youtube? Regional income patterns of mobile service consumption. In: NetMob 2019, Oxford, UK
40. Barabási A-L (2016) Network science. Cambridge University Press, Cambridge. http://networksciencebook.com/
41. Salgado D, Esteban ME, Novás M, Saldaña S, Sanguiao L (2018) Data organisation and process design based on functional modularity for a standard production process. J Off Stat 34(4):811–833. https://doi.org/10.2478/jos-2018-0041
42. Oancea B, Necula M, Sanguiao L, Salgado D, Barragán S (2019) A simulator for network event data. Technical report, Statistics Romania (INS) and Statistics Spain (INE). https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPI_Deliverable_I2_Data_Simulator_-_A_simulator_for_network_event_data.pdf_en
43. Bordin MV (2017) A call detail record (CDR) generator. https://github.com/mayconbordin/cdr-gen
44. Real Impact Analysis (2014). CDR generator. https://github.com/RealImpactAnalytics/cdr-generator
45. Tetcos (2019) NetSim user manual. https://www.tetcos.com/downloads/v12/NetSim_User_Manual.pdf
46. Zhen L, Hongji Y (2012) Unlocking the power of OPNET modeler. Cambridge University Press, New York
47. Krajzewicz D, Erdmann J, Behrisch M, Bieker L (2012) Recent development and applications of sumo - simulation of urban mobility. J Adv Syst Meas 5(3&4):128–138
48. Horni A, Nagel K, Axhausen KW (2016) The multi-agent transport simulation MATSim. Ubiquity Press, London
49. Shabbir N, Sadiq MT, Kashif H, Ullah R (2011) Comparison of radio propagation models for long term evolution (LTE) network. Int J Next-Gener Netw 3(3):27–41. https://doi.org/10.5121/ijngn.2011.3303
50. Salgado D, Sanguiao L, Oancea B, Barragán S, Necula M (2020) Collection of data sets and scripts for "An end-to-end statistical process with mobile network data for official statistics". https://figshare.com/articles/dataset/_/12861095
51. Caffery JJ, Stuber GL (1998) Overview of radiolocation in CDMA cellular systems. IEEE Commun Mag 36(4):38–45. https://doi.org/10.1109/35.667411
52. Dye M, Baylin F (2001) Mobile positioning. Mobile Lifestreams Ltd, London
53. Gustafsson F, Gunnarsson F (2005) Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements. IEEE Signal Process Mag 22(4):41–53. https://doi.org/10.1109/msp.2005.1458284
54. Gezici S (2007) A survey on wireless position estimation. Wirel Pers Commun 44(3):263–282. https://doi.org/10.1007/s11277-007-9375-z
55. Mohammadi M, Molaei E, Naserasadi A (2011) A survey on location based services and positioning techniques. Int J Comput Appl 24(5):1–5. https://doi.org/10.5120/2946-3928
56. Liu D, Sheng B, Hou F, Rao W, Liu H (2014) From wireless positioning to mobile positioning: an overview of recent advances. IEEE Syst J 8(4):1249–1259. https://doi.org/10.1109/jsyst.2013.2295136
57. Mahyuddin MFM, Isa A, Zin M, Afifah Maheran AH, Manap Z, Ismail MK (2017) Overview of positioning techniques for LTE technology. J Telecommun Electron Comput Eng 9(2–13):43–50
58. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286. https://doi.org/10.1109/5.18626
59. Bishop CM (2006) Pattern recognition and machine learning. Springer, Cambridge
60. Vanhoof M, Reis F, Ploetz T, Smoreda Z (2018) Assessing the quality of home detection from mobile phone data for official statistics. J Off Stat 34(4):935–960. https://doi.org/10.2478/jos-2018-0046
61. Lehmann EL, Casella G (2003) Theory of point estimation. Springer, New York
62. Daskalakis C, Kamath G, Tzamos C (2015) On the structure, covering, and learning of Poisson multinomial distributions. https://doi.org/10.1109/FOCS.2015.77
63. Royle AJ, Dorazio RM (2009) Hierarchical modelling and inference in ecology. Elsevier, New York
64. Bryant JR, Graham PJ (2013) Bayesian demographic accounts: subnational population estimation using multiple data sources. Bayesian Anal 8(3):591–622. https://doi.org/10.1214/13-ba820
65. Bryant JR, Graham P (2015) A Bayesian approach to population estimation with administrative data. J Off Stat 31(3):475–487. https://doi.org/10.1515/jos-2015-0028

66. Eurostat (2014) Feasibility study on the use of mobile positioning data for tourism statistics?: consolidated report. Publications Office. https://doi.org/10.2785/55051
67. ESS (2016) ESSnet on Big Data. https://ec.europa.eu/eurostat/cros/content/essnet-big-data-0_en
68. UN GWG (2017) Handbook on the use of mobile phone data for official statistics. Technical report, United Nations. https://unstats.un.org/bigdata/blog/2019/mpd-task-team.cshtml
69. Saluveer E, Raun J, Tiru M, Altin L, Kroon J, Snitsarenko T, Aasa A, Silm S (2020) Methodological framework for producing national tourism statistics from mobile positioning data. Ann Tour Res 81:102895. https://doi.org/10.1016/j.annals.2020.102895
70. WPI (2018) Work package I on mobile network data. Project ESSnet on Big Data II. https://ec.europa.eu/eurostat/cros/content/WPI_Mobile_networks_data_en
71. Banerjee S, Carlin BC, Gelfand AE (2015) Hierarchical modelling and analysis for spatial data, 2nd edn. CRC Press, Boca Raton
72. Sanguiao L, Barragán S, Salgado D (2020) destim: an R package for mobile devices position estimation. R package version 0.1.0. https://github.com/Luis-Sanguiao/destim
73. Oancea B, Barragán S, Salgado D (2020) deduplication: an R package for deduplicating mobile device counts into population individual counts. R package version 0.1.0. https://github.com/bogdanoancea/deduplication
74. Oancea B, Barragán S, Salgado D (2020) aggregation: an R package to produce probability distributions of aggregate number of mobile devices. R package version 0.1.0. https://github.org/bogdanoancea/aggregation
75. Oancea B, Barragán S, Salgado D (2020) inference: R package for computing the probability distribution of the number of individuals in the target population. R package version 0.1.0. https://github.com/bogdanoancea/inference