RESEARCH

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†]

[*]Correspondence: david.salgado.fernandez@ine.es
[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Paseo de la Castellana, 183, Madrid, Spain
Full list of author information is available at the end of the article
[†]The views expressed in this working paper are those of the authors and do not necessarily reflect the views of their affiliating institutions.

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) [1], constitutes a rich source of information for Social Science, in general, and for Official Statistics, in particular. There already exist multiple excellent examples of one-off applications [2–15] (see supplementary material 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, 28] with some recent examples of an end-to-end statistical production process [29]. 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 [30].

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 [31, 32]. 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 [33, 34] 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 [35] would be strengthened. Thirdly, a new branch of economic activity is growing on the basis of digital data and data analytics [36]. 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 [37], 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. 38], and (iii) investigating interactions between devices to analyse different aspects of social networks [39]. 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 [40] focusing on input and output data as well as the throughput of each production step. The next sections describe each of the functional modules of the statistical process. In section 2 we provide a description of the (synthetic) data used to illustrate the proposal. In section 3 we describe the module to geolocate mobile devices. In section 4 we propose a method to disambiguate devices carried by the same individual. In section 5 we include general considerations to identify devices pertaining to the target population under analysis. In section 6 we suggest a method to aggregate data from the device level to the territorial unit level. In section 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 section 8 we integrate all modules in a production chain. Finally, in section 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 [41] 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;
- a telecommunication network configuration in terms of a radiowave propagation model;
- a reference grid for analysis.

The simulator works essentially by using a radio wave propagation model to simulate the handover mechanism between the antennas 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 antenna 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, ...).

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 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, radiowave propagation models [42] 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. We are using two very simple models for the connection mechanism. For omnidirectional antennas:

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

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

  where P stands for the antenna 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 antenna producing the highest signal strength in each tile until the antenna 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.* [29], we further model a so-called Signal Dominance Measure (SDM) by making a logistic transformation on the RSS:

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

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

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 figure 1 for an illustrative example of the RSS and SDM of a given antenna.

For directional antennas, more parameters are needed (see [29]). For simplicity, we shall use only omnidirectional antennas in this work.

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 $\times$ 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 70 omnidirectional antennas. See figure 2 and animated gif `individuals.gif` in [43]. Parameters are further specified in the supplementary material.

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 [44–50], 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 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 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) [51, 52], 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 figure 3). Now, we need two models:

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

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

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

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

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 figure 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 NSIs 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 antennas 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. $b_{i0} = \mathbb{P}\left(E_{t_n} = \cdot \middle| T_{t_n} = i, \mathbf{I}^{\text{aux}}\right) = 1$. 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 $\pi_i \equiv \mathbb{P}\left(T_{d0} = i \middle| \mathbf{I}^{\text{aux}}\right)$ 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 [52] to compute the (posterior) location probabilities $\gamma_{dti} \equiv \mathbb{P}\left(T_{dt} = i \middle| \mathbf{E}_{d1:T}, \mathbf{I}^{\text{aux}}\right)$, 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 figure 4 for the location probabilities at time $t = 0$ and animated gifs `postLocLayer*.gif` in [43]). Also, we compute the (posterior) joint location probabilities $\gamma_{dt,ij} \equiv \mathbb{P}\left(T_{dt} = i, T_{dt-1} = j \middle| \mathbf{E}_{d1:T}, \mathbf{I}^{\text{aux}}\right)$. 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 mimick 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:

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

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

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

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:

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

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 figures 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 model seems to provide more accurate estimates in terms of the distance to the true position of the devices, but the SDM 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 handover mechanism follows the RSS

model, the emission model is trivially closer to this true handover mechanism, providing best geolocation estimates. Furthermore, according to figure 1, the SDM model is more localized (this is the effect of the logistic transformation), thus the root mean squared dispersion is lower. Secondly, the radiowave 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 model is too simplistic for real life conditions (e.g. due to the load balancing of the network) and the SDM 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 an antenna 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 antennas) 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. 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 $\{H_{dd'}\}_{d'=1,\ldots,D}$ 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

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

where we are using the same notation as in section 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:

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

where we have defined the prior probability ratios $\alpha_{dd'} = \frac{\mathbb{P}\left(H_{dd'}\middle|\mathbf{I}^{\text{aux}}\right)}{\mathbb{P}\left(H_{dd}\middle|\mathbf{I}^{\text{aux}}\right)}$ and the integrated log-likelihoods $\ell_d = \mathbb{P}\left(\mathbf{E}_{d1:T}\middle|H_{dd},\mathbf{I}^{\text{aux}}\right)$ for a single device d and $\ell_{dd'} = \mathbb{P}\left(\mathbf{E}_{d1:T},\mathbf{E}_{d'1:T}\middle|H_{dd'},\mathbf{I}^{\text{aux}}\right)$ 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 figure 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 devices d and any other d' reside in far away locations, then naturally $\mathbb{P}\left(H_{dd'}|\mathbf{I}^{\text{aux}}\right)\approx 0$ 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 $\mathbb{P}\left(H_{dd'}|\mathbf{I}^{\text{aux}}\right)$ is constant for any other device $d'\neq d$. Since $\Omega_d = \bigcup_{d'=1}^{N_D} H_{dd'}$, then $\mathbb{P}\left(H_{dd}|\mathbf{I}^{\text{aux}}\right)+(N_D-1)\cdot\mathbb{P}\left(H_{dd'}|\mathbf{I}^{\text{aux}}\right)=1$ for any other device d'. We arrive at

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

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 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 figure 8 we represent the ROC curves for the duplicity probabilities for the four models, together with their corresponding area under the curve (AUC). In figure 9 we represent the different cases (true/false positive/negative) in each model.

Taking into account that the handover 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 handover 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 figure 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.

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 antenna 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 antenna 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 and 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 section 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. 53, 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 section 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 next subsection we will 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. Thus, the starting point will be the construction of this probability-based trajectory for each device.

5.2 Probability-based paths

In our model introduced in section 3 the state of the HMM was defined in terms of the tile where the device is positioned. Thus, in this case the concept of space-time trajectory follows immediately as the time sequence of states, in which we shall use the coordinates of each tile to build the so-called *path* $\{(x_{dt_0}, y_{dt_0}), (x_{dt_1}, y_{dt_1}), \ldots, (x_{dt_N}, y_{dt_N})\}$, where at each time instant t_i the spatial coordinates x_{dt_i} and y_{dt_i} for device d are specified. In more complex definitions of states, another procedure should lead us to deduce the path from the adopted concept of HMM state. If auxiliary information for each tile is available, instead of the geographical centroid of each tile, another "statistical"centroid can be used (e.g. using land use information and/or official population density figures). It is obvious that the smaller the tiles, the more precise the estimation procedures.

Given an HMM, it is well-known that at least two different methods can be approached to build a sequence of states, i.e. a trajectory in our case. We can compute either the most probable sequence of states or the sequence of most probable states. In mathematical terms, the former is the sequence

$$T^*_{dt_0:t_N} = \text{argmax}_{T_{dt_0:t_N}} \mathbb{P}\left(T_{dt_0:t_N} \middle| \mathbf{E}_{dt_0:t_N}, \mathbf{I}^{\text{aux}}\right), \tag{12}$$

which can be computed by means of the Viterbi algorithm [see e.g. 54]. The second method is indeed given by

$$T^*_{dt_0:t_N} = \left(\text{argmax}_{T_{dt_0}} \gamma_{dt_0}, \text{argmax}_{T_{dt_1}} \gamma_{dt_1}, \ldots, \text{argmax}_{T_{dt_N}} \gamma_{dt_N}\right), \tag{13}$$

where $\gamma_{dt_j} = \mathbb{P}\left(T_{dt_j} \middle| \mathbf{E}_{dt_0:t_N}, \mathbf{I}^{\text{aux}}\right)$ are the posterior location (state) probabilities.

We choose the maximal posterior marginal (MPM) trajectory because it is more robust and because unimodal probabilities are expected so that differences will not be large [54].

5.3 Quantitative measures of movement data

Once a path is assigned to each device we can compute different indicators as well as joint measures. Following [55] (see also multiple references therein) we distinguish the following groups of measures:

- Time geography.- This represents a framework for investigating constraints such as maximum travel speed on movement in both the spatial and temporal dimensions. These constraints can be capability constraints (limiting movement possibilities because of biological/physical abilities), coupling constraints (specific locations a device must visit thus limiting movement possibilities), and authority constraints (specific locations a device cannot visit thus also limiting movement possibilities).
- Path descriptors.- These represent measurements of path characteristics such as velocity, acceleration, turning angles. By and large, they can be characterised based on space, time, and space-time aspects.
- Path similarity indices.- These are routinely used to quantify the level of similarity between two paths. Diverse options exist in the literature, some already taking into account that paths are sequences of stays and displacements [see e.g. 55].
- Pattern and cluster methods.- These seek to identify spatial–temporal patterns from the whole set of paths. These are mainly used to focus on the territory rather than on individual patterns. They also consider diverse aspects on space, time, and space-time features.
- Individual–group dynamics.- This set of measures compile methods focusing on individual device displacement within the context of a larger group of devices (e.g. a tourist within a larger group of tourists in the same trip).
- Spatial field methods.- These are based on the representation of paths as space or space-time fields. Different advanced statistical methods can be applied such as kernel density estimation or spatial statistics.
- Spatial range methods.- These are focused on measuring the area containing the device displacement, such as net displacement and other distance metrics.

We include an illustrative example with a set of indicators. We shall compute them on the simulated scenario with 218 devices in a territory with an irregular polygon shape and a bounding box of 10km × 10km. The indicators are strongly inspired on those used in animal trajectory analysis [see 56, and references therein].

1 Number of coordinates (`nCoord`).- This is the observed number of coordinates along the path, thus coincidental with the time extension of the HMM.

2 Path length (`length`).- This is the total length of the path, i.e.

$$\text{Length} = \sum_{t=1}^{T} \ell_t,$$

where $\ell_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$.

3 Path distance (`distance`).- This is the net distance between the initial and final fixes (points) in the path, i.e.

$$\text{distance} = \sqrt{(x_T - x_0)^2 + (y_T - y_0)^2}.$$

4 Path duration (`duration`).- This is the total duration of the path, i.e.

$$\text{duration} = t_N - t_0.$$

5 Mean velocity (`meanVelocity`).- This is the global mean velocity of the device along the path, i.e.

$$\text{meanVelocity} = \frac{1}{\text{duration}}(x_T - x_0, y_T - y_0).$$

Notice that it is a vector, thus we compute both the x- and y- dimensions.

6 Radius of gyration (`Rg`).- This is the radius of gyration of the path according to the formula

$$Rg = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t^2 + y_t^2)}.$$

It provides a view of the extension of the territory range covered by the path.

7 Path straightness (`straightness`).- This is the index $\frac{\text{distance}}{\text{length}}$, which provides a first-order magnitude of the tortuosity of the path, with values between 0 (extremely tortuous) and 1 (a straight line).

8 Turning angles (`turningAngles`).- These are the angles θ_t denoting the change of direction at each time instant t. See figure 10.

9 Directional change (`directionalChange`).- This is a measure of the speed of angular change of direction, defined as

$$\text{directionalChange}_{t_i} = \frac{\theta_{t_i}}{t_i - t_{i-1}}.$$

10 r Index (`r`).- This is another more complex measure of the tortuosity of the path, defined as

$$r = \left| \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} e^{i\bar{\theta}_t} \right|$$

where $\bar{\theta}_t$ denotes the turning angle (see figure 10) at time t of the rediscretized path obtained by sampling the path at equal-length steps.

11 Maximum expected displacement (`EmaxA` and `EmaxB`).- These two related indicators provide a measure of the maximum expected displacement according to

$$E^a_{max} = \frac{\xi}{1-\xi},$$

where $\xi \equiv \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \cos\left(\bar{\theta}_t\right)$, with $\bar{\theta}_t$ being the turning angles of the rediscretised path obtained by sampling the path at equal-length steps.

The related indicator E^b_{max} is defined as

$$E^b_{max} = \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \bar{\Delta}_t \times \frac{\xi}{1-\xi},$$

where $\bar{\Delta}_t$ is the step length at time t of the rediscretised path.

12 Path sinuosity (`sinuosity` and `sinuosity2`).- The original path sinuosity index is defined as

$$\text{sinuosity} = 1.18 \times \frac{\sigma_\theta}{\sqrt{\bar{\Delta}}},$$

where $\sigma_\theta = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\theta_t - \bar{\theta})^2}$, $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ and $\bar{\Delta} = \frac{1}{T}\sum_{t=1}^{T}\Delta_t$. A second version using rediscretised paths is given by:

$$\text{sinuosity2} = \frac{2}{\sqrt{\bar{\Delta} \times \left(\frac{1+\xi}{1-\xi}\right) + \left(\frac{\sigma_\Delta}{\bar{\Delta}}\right)^2}}.$$

We have computed these indicators on our simulated scenario producing the values represented in figures 11 and 12. This list of indicators is not exhaustive (even some alternative forms for them can be found in the literature as for $E^{a,b}_{max}$ or the sinuosity index). Our main argument is that filtering, comprising identification of usual environment, home/work detection, second home detection, etc., must be based on detailed algorithms using these indicators avoiding as much as possible extremely simplistic approaches such as a home is a location where devices are between 23:00 and 06:00 or similar. Ultimately, findings thereof should be connected to other sociodemographic variables producing thus novel insights.

As a simple example, for each given path we can identify the time instants where the observed speed is below a given threshold for a consecutive number of time intervals thus identifying potential home/work locations (see figure 13). Then, different indicators can be computed for this subpath so that further distinction between activities could be unravelled (shopping, sporting, etc.). Notice that the limit imposed by the spatial resolution of the HMM and the accuracy of the emission model establish a bound in this regard.

The reader immediately will realise how more complex and realistic displacement patterns in the simulator are needed to go deep into this analysis in practical terms. In the example in figure 13 the displacement pattern does not correspond to a realistic human displacement whatsoever, so that no reasonable detection algorithm can be proposed using this data. This remains for further work in the future.

Finally, let us close this section by calling reader's attention on the positive feedback arising from this statistical filtering exercise. Once concepts such as usual environment, home/work location, second home location, etc. are computed, the definition of state for the HMM could be enhanced thus incorporating more information into the geolocation estimation. As a final suggestion widening the possibilities, instead of defining indicators such as above, deep learning techniques could be also tested to extract different characteristics of the paths.

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 57, 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 section 3) and if the duplicity of a given device (carried by an individual with another device) is also probabilistic in nature (see section 4), then a priori it is impossible to provide a certain number of individuals[1] 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

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

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 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 based 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 section 4. Should we consider a more general situation, the generalization is obvious, although more computationally demanding.

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

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(1)}|\mathbf{E}_{1:D}, \mathbf{I}^{\text{aux}}\right) = \gamma_{dti} \cdot p_d^{(1)} \tag{14a}$$

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(2)}|\mathbf{E}_{1:D}, \mathbf{I}^{\text{aux}}\right) = \gamma_{dti} \cdot p_d^{(2)} \tag{14b}$$

where $p_d^{(1)}$ and $p_d^{(2)}$ ($p_d^{(1)} + p_d^{(2)} = 1$) are the device duplicity probabilities introduced in section 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 $i = 1, \ldots, N_T$ at time instant t:

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

The sum spans over the number of devices filtered as members of the target population according to section 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. 58] 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

$$\bar{\gamma}_{dtr} = \sum_{i \in \mathcal{I}_r} \gamma_{dti}. \tag{16}$$

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:

$$\bar{\mathbf{N}}_t^{\text{net}} = \sum_{d=1}^{D} \bar{\mathbf{T}}_{dt}. \tag{17}$$

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

$$\mathbb{E}\left[\bar{\mathbf{N}}_t^{\text{net}}\right] = \sum_{d=1}^{D} \sum_{r=1}^{R} \bar{\gamma}_{dtr}^{\text{dedup}} \mathbf{e}_r, \tag{18}$$

$$\mathbb{V}\left[\bar{\mathbf{N}}_t^{\text{net}}\right] = \sum_{d=1}^{D} \sum_{r=1}^{R} \bar{\gamma}_{dtr}^{\text{dedup}} \cdot (1 - \bar{\gamma}_{dtr}^{\text{dedup}}) E_{rr}. \tag{19}$$

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

$$\mathbb{P}\left(E_{dt} = E_{rs}^{(1)}\right) = \tilde{\gamma}_{dt,sr} \cdot p_d^{(1)}, \tag{20a}$$

$$\mathbb{P}\left(E_{dt} = E_{rs}^{(2)}\right) = \tilde{\gamma}_{dt,sr} \cdot p_d^{(2)}, \tag{20b}$$

where $\gamma_{dt,sr}$ 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

$$\bar{\mathbf{N}}_t^{\text{net}} = \sum_{d=1}^{D} E_{dt}, \tag{21}$$

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

$$\mathbb{E}\left[\bar{\mathbf{N}}_t^{\text{net}}\right] = \sum_{d=1}^{D} \sum_{r,s=1}^{R} \tilde{\gamma}_{dt,sr}^{\text{dedup}} \cdot E_{rs}. \tag{22}$$

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 10km × 10km divided into 10 regions as in figure 14. 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 figure 15. Once we have posterior distributions we can also compute credible intervals for each region and each time instant (see figure 16). 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 section 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 figure 17).

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 section 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 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 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 [59]. 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.

The first level makes use of the detection probability p_r of individuals of a network in each region r. We shall concentrate first on the observation process. We model

$$N_r^{\text{net}} \simeq \text{Binomial}\,(N_r, p_r)\,. \tag{23}$$

This model 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. At this first level, we shall consider this as an external parameter taken e.g. from the national telecommunication regulator. The posterior probability distribution for N_r in terms of N_r^{net} 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

$$\widehat{N}_r^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{(1-p_r)\cdot N_r^{\text{net}}}{p_r} \right\rfloor, \tag{24a}$$

$$\widehat{N}_r^{\text{mean}} = N_r^{\text{net}} + \frac{(1-p_r)\cdot (N_r^{\text{net}} + 1)}{p_r}. \tag{24b}$$

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 computed in section 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

$$\Omega_r^{(1)} = \frac{\sum_{d=1}^{D} \bar{\gamma}_{dr} \cdot p_d^{(1)}}{\sum_{d=1}^{D} \bar{\gamma}_{dr}}, \tag{25a}$$

$$\Omega_r^{(2)} = \frac{\sum_{d=1}^{D} \bar{\gamma}_{dr} \cdot p_d^{(2)}}{\sum_{d=1}^{D} \bar{\gamma}_{dr}}. \tag{25b}$$

The deduplicated penetration rates are defined as

$$\tilde{P}_r^{\text{net}} = \left(\Omega_r^{(1)} + \frac{\Omega_r^{(2)}}{2} \right) \cdot P_r^{\text{net}}. \tag{25c}$$

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

$$\alpha_r + \beta_r = N_r^{\text{reg}}, \tag{26a}$$
$$\frac{\alpha_r}{\alpha_r + \beta_r} = \tilde{P}_r^{\text{net}}, \tag{26b}$$

which immediately implies that

$$\alpha_r = \tilde{P}_r^{\text{net}} \cdot N_r^{\text{reg}}, \tag{27a}$$
$$\beta_r = \left(1 - \tilde{P}_r^{\text{net}}\right) \cdot N_r^{\text{reg}}. \tag{27b}$$

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 [60, 61]. 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:

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

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

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

The uncertainty is accounted for by computing the posterior variance, the posterior coefficient of variation, or credible intervals.

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 (28) reduces to the negative binomial distribution

$$\mathbb{P}\left(N_r | N_r^{\text{net}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}\left(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r+\beta_r-1}, N_r^{\text{net}}+1\right) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases} \tag{30}$$

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

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

So far, the inference has been conducted independently in each region r. We can introduce another layer in the hierarchy 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. For example, we can go all the way down the hierarchy, assume a cross-cutting relationship between parameters and some hyperparameters and postulate

$$N_r^{\text{net}} \simeq \text{Bin}\left(N_r, p_r\right), \quad \text{for all } r = 1, \ldots, R, \tag{31a}$$

$$p_r \simeq \text{Beta}\left(\alpha_r, \beta_r\right), \quad \text{for all } r = 1, \ldots, R, \tag{31b}$$

$$\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r+\beta_r}\right), \alpha_r+\beta_r\right) \simeq \text{N}\left(\mu_{\gamma r}(\gamma_0, \gamma_1; \bar{P}_r^{\text{net}}), \tau_\gamma^2\right) \times \text{Gamma}\left(1+\xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1, \ldots, R, \tag{31c}$$

$$\left(\log\gamma_0, \gamma_1, \tau_\gamma^2, \xi\right) \simeq \text{f}_\gamma\left(\log\gamma_0, \gamma_1, \tau_\gamma^2\right) \times \text{f}_\xi(\xi), \tag{31d}$$

where we have denoted $\mu_{\gamma r}(\gamma_0, \gamma_1; \bar{P}_r^{\text{net}}) \equiv \log\left(\gamma_0 \left[\frac{\bar{P}_r^{\text{net}}}{1-\bar{P}_r^{\text{net}}}\right]^{\gamma_1}\right)$ 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 $\gamma_0, \gamma_1, \tau_\gamma^2$. 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, as above.

Under the specifications (31), 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:

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) \propto \int_{\mathbb{R}^R} d^R\mathbf{y}\; \omega_{\text{obs}}\left(\mathbf{y}; \bar{\mathbf{P}}^{\text{net}}\right) \prod_{r=1}^{R} \frac{\text{negbin}\left(N_r - N_r^{\text{net}}; 1-p(y_r), N_r^{\text{net}}+1\right)}{p(y_r)}, \tag{32}$$

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

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 (32) reduces to

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; 1-\tilde{P}_r^{\text{net}}, N_r^{\text{net}}+1\right). \tag{33}$$

The marginal distribution for region r reduces to (30), 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.

$$N_r \simeq \text{Poisson}\left(A_r \sigma_r\right), \tag{34}$$

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:

$$\begin{aligned}
N_r^{\text{net}} &\simeq \text{Bin}\left(N_r, p_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(35a)}\\
N_r &\simeq \text{Poisson}\left(A_r \sigma_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(35b)}\\
p_r &\simeq \text{Beta}\left(\alpha_r, \beta_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(35c)}\\
\sigma_r &\simeq \text{Gamma}\left(1 + \zeta_r, \theta_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(35d)}
\end{aligned}$$

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 (27). 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 $\mathbb{V}\left(\tau_r\right) = \left(\zeta_r + 1\right) \cdot \theta_r^2$. We shall parametrise these gamma distributions in terms of the register-based population densities σ_r^{reg} as

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

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

$$\begin{aligned}
\theta_r(\Delta\sigma_r, \epsilon_r) &= \frac{\sigma_r^{\text{reg}}}{2}\left(1 + \frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}\right)\left[\sqrt{1 + \left(\frac{2\epsilon_r}{1 + \frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}}\right)^2} - 1\right],\\
\zeta_r(\Delta\sigma_r, \epsilon_r) &= \frac{2}{\sqrt{1 + \left(\frac{2\epsilon_r}{1 + \frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}}\right)^2} - 1}. \qquad \text{(36)}
\end{aligned}$$

Under assumptions (35) and assuming $\alpha_r, \beta_r \gg 1$, as above, we get

$$\mathbf{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r + \beta_r} \cdot Q(\theta_r), N_r^{\text{net}} + 1 + \zeta_r\right) \quad (37)$$

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 (37):

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

Accuracy indicators such as posterior variance or credible intervals are computed from the distribution (37).

Expression (37) 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 (27), we have the Poisson limit of the negative binomial distribution so that

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{poisson}\left(N_r - N_r^{\text{net}}; (1-\bar{P}_r^{\text{net}}) \cdot A_r\sigma_r^{\text{reg}}\right). \quad (38)$$

The MAP estimator is trivially $\hat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor (1-\bar{P}_r)A_r\sigma_r^{\text{reg}} \right\rfloor$ and the mean estimator is trivially $\hat{N}^{\text{MAP}} = N_r^{\text{net}} + (1-\bar{P}_r)A_r\sigma_r^{\text{reg}}$, 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

$$\mathbb{P}\left(\mathbf{N}\middle|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; 1-\bar{P}_r, N_r^{\text{net}} + 1\right), \quad (39)$$

which is the same expression as (33), as expected, since having no information about the state process is equivalent to having only the observation process.

We can also introduce more levels in the hierarchy:

$$N_r^{\text{net}} \simeq \text{Binomial}\left(N_r, p_r\right), \quad \text{for all } r = 1, \ldots, R, \tag{40a}$$

$$N_r \simeq \text{Poisson}\left(A_r \sigma_r\right), \quad \text{for all } r = 1, \ldots, R, \tag{40b}$$

$$p_r \simeq \text{Beta}\left(\alpha_r, \beta_r\right), \quad \text{for all } r = 1, \ldots, R, \tag{40c}$$

$$\sigma_r \simeq \text{Gamma}\left(\zeta + 1, \frac{e^{\theta_r}}{\zeta}\right), \quad \text{for all } r = 1, \ldots, R, \tag{40d}$$

$$\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r + \beta_r}\right), \alpha_r + \beta_r\right) \simeq \text{N}\left(\mu_{\gamma r}(\gamma_0, \gamma_1; \bar{P}_r^{\text{net}}), \tau_\gamma^2\right) \times \text{Gamma}\left(1 + \xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1, \ldots, R, \tag{40e}$$

$$\theta_r \simeq \text{N}\left(\mu_{\delta r}(\delta_0, \delta_1; \sigma_r^{\text{reg}}), \tau_\delta^2\right), \quad \text{for all } r = 1, \ldots, R, \tag{40f}$$

$$\left(\log\gamma_0, \gamma_1, \tau_\gamma^2, \xi\right) \simeq \text{f}_\gamma\left(\log\gamma_0, \gamma_1, \tau_\gamma^2\right) \times \text{f}_\xi(\xi) \tag{40g}$$

$$\left(\log\delta_0, \delta_1, \delta_\delta^2, \zeta\right) \simeq \text{f}_\delta\left(\log\delta_0, \delta_1, \delta_\delta^2\right) \times \text{f}_\zeta(\zeta), \tag{40h}$$

where we have denoted $\mu_{\delta r}(\delta_0, \delta_1; \sigma_r^{\text{reg}}) \equiv \log\left(\delta_0 \left[\sigma_r^{\text{reg}}\right]^{\delta_1}\right)$ and f_γ, f_ξ, f_δ, f_ζ 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 (40), 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

$$\begin{aligned} \mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) \propto & \int_{\mathbb{R}^R} d^R\mathbf{y}\ \omega_{\text{obs}}\left(\mathbf{y}; \bar{\mathbf{P}}^{\text{net}}\right) \prod_{r=1}^{R} \frac{\text{negbin}\left(N_r - N_r^{\text{net}}; 1 - p(y_r), N_r^{\text{net}} + 1,\right)}{p(y_r)} \\ \times & \int_{\mathbb{R}^R} d^R\mathbf{z}\ \omega_{\text{state}}\left(\mathbf{z}; \boldsymbol{\sigma}^{\text{reg}}\right) \prod_{r=1}^{R} \text{negbin}\left(N_r; q\left(\frac{A_r e^{z_r}}{\zeta}\right), 1 + \zeta\right), \end{aligned} \tag{41}$$

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

- $q\left(\frac{A_r e^{z_r}}{\zeta}\right) \equiv \frac{\frac{A_r e^{z_r}}{\zeta}}{1+\frac{A_r e^{z_r}}{\zeta}}$;
- $\omega_{\text{state}}(\mathbf{z};\boldsymbol{\sigma}^{\text{reg}}) = \int_{\Omega_{\delta,\zeta}} \mathrm{d}\log\delta_0 \mathrm{d}\delta_1 \mathrm{d}\delta_\delta^2 \mathrm{d}\zeta\ \mathrm{f}_\delta\left(\log\delta_0, \delta_1, \delta_\delta^2\right) \times \mathrm{f}_\zeta(\zeta)\ n\left(\mathbf{z}; \boldsymbol{\mu}_\delta(\delta_0, \delta_1; \boldsymbol{\sigma}^{\text{net}}), \boldsymbol{\Sigma}_\delta\right)$ 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 $\boldsymbol{\Sigma}$.
  - $\mu_{\delta r}(\delta_0, \delta_1; \sigma_r^{\text{reg}}) = \log\left(\delta_0 \left[\sigma_r^{\text{reg}}\right]^{\delta_1}\right)$.
  - $\boldsymbol{\Sigma}_\delta = \tau_\delta^2\ \mathbb{I}_{R\times R}$.

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 (41) reduces to

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; (1-\bar{P}_r)\cdot Q_r(\zeta^*), N_r^{\text{net}} + \zeta^* + 1\right), \quad (42)$$

where we have denoted $Q_r(\zeta) \equiv q(\frac{A_r \sigma_r^{\text{reg}}}{\zeta})$, which is indeed again equation (37).

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

$$\begin{aligned} N_{tr} &= N_{t-1r} + \sum_{\substack{r_t=1 \\ r_t\neq r}}^{N_T} N_{t,rr_t} - \sum_{\substack{r_t=1 \\ r_t\neq r}}^{N_r} N_{t,r_t r} \\ &= \sum_{r_t=1}^{N_T} \tau_{t,rr_t} \cdot N_{t-1r_t}, \end{aligned} \quad (43)$$

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 (section 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 (43) 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 subscriber 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 10km $\times$ 10km divided into 10 regions as in figure 14. 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 (28), (30), and (37), 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 (37), 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 figures 18, 19, and 20 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 (37) is also evident, thus inviting not to model the state process. 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 data from each module is the input data for the next. Mathematically, for the present population use case this can be expressed as

$$\mathbb{P}\left(\mathbf{N}_t|\mathbf{E}_{0:T}, \mathbf{N}^{\text{reg}}, \mathbf{P}^{\text{net}}\right) = \sum_{N_{tr}^{\text{net}} \geq 0} \mathbb{P}\left(\mathbf{N}_t|\mathbf{N}_t^{\text{net}}, \mathbf{N}^{\text{reg}}, \mathbf{P}^{\text{net}}\right) \mathbb{P}\left(\mathbf{N}_t^{\text{net}}|\mathbf{E}_{0:T}\right). \quad (44)$$

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 figures 21, 22, and 23 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 [43] 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.

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 figure 24 for $\text{cv}^{\text{reg}} = 0.01$ and $\text{rb}^{\text{reg}} = 0$ and in figure 25 for $\text{cv}^{\text{reg}} = 0.20$ and $\text{rb}^{\text{reg}} = 0.20$ for the beta negative binomial inference model (see [43] 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 Conclusions and future prospects

To produce official statistics in a sustainable and routinely way in a statistical office, we need to put in place a modular and evolvable statistical production process providing valid for diverse statistical domains. We propose an end-to-end process with these characteristics composed of several modules: (i) a geolocation module providing location probabilities for each device according to the information provided by the telecommunication network; (ii) a deduplication module providing device duplicity probabilities to disambiguate those devices carried by the same individual; (iii) a statistical filtering module providing an identification of those devices comprised by the target population; (iv) an aggregation module providing distributions for the number of individuals detected by network, and (v) an inference module providing distributions for the number of individuals in the target population.

All modules are integrated into a production chain in which the output data from each module is the input data for the next, apart from auxiliary information integrated from external data sources such as official data and telco market information. The language of probability used throughout the end-to-end process allows us to integrate auxiliary information in a natural way and to account for uncertainty all along the process, thus providing accuracy indicators of both the intermediate and final estimates.

The main result of this work is not in the details themselves of each module but on the whole process as a modular structure. Indeed, this modularity will allow us to further investigate the statistical methodology underlying each of the module. 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 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 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.

The whole methodology for the use of mobile network data in official statistical production needs further research and testing. 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 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.

To generate the illustrative examples included above, apart from the network data event simulator [41], we have developed independent prototyping R packages for each module. Package `destim` for geolocation [62]. Package `deduplication` for deduplicating devices [63]. Package `aggregation` to get the probability distributions of the aggregate number of individuals detected by the network [64]. Package `inference` to get the probability distributions of the aggregate number of individuals in the target population [65]. 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.

Declaration

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.

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).

Authors' contributions

All authors have contributed equally.

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.

Author details

[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Paseo de la Castellana, 183, 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.

References

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

28. Ricciato, F.: Towards a Reference Methodological Framework for processing MNO data for Official Statistics. In: 15th World Forum on Tourism Statistics (2018)
29. Tennekes, M., Gootzen, Y.A.P.M., Shah, S.H.: A Bayesian approach to location estimation of mobile devices from mobile network operator data. resreport, Statistics Netherlands (CBS) (May 2020). https://www.cbs.nl/-/media/_pdf/2020/22/cbds_working_paper_location_estimation.pdf
30. UNECE: 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 (2011). https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/1.e.pdf
31. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S.: Statistical Disclosure Control. John Wiley & Sons, Ltd, Chichester (2012). doi:10.1002/9781118348239
32. Templ, M.: Statistical Disclosure Control for Microdata. Springer, Berlin (2017). doi:10.1007/978-3-319-50272-4
33. de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific Reports **3**(1) (2013). doi:10.1038/srep01376
34. Xu, F., Tu, Z., Li, Y., Zhang, P., Fu, X., Jin, D.: Trajectory recovery from ash. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, ??? (2017). doi:10.1145/3038912.3052620
35. UNECE: Fundamental Principles of Official Statistics. Technical report, United Nations (1992). https://www.unece.org/stats/fps.html
36. Commission, E.: Shaping Europe's digital future. https://ec.europa.eu/digital-single-market/en (2020)
37. Salgado, D., Oancea, B.: On new data sources for the production of official statistics. Statistics Spain (INE) Working Paper 01/2020 (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
38. Ucar, I., Gramaglia, M., Fiore, M., Smoreda, Z., Moro, E.: Netflix or Youtube? Regional Income Patterns of Mobile Service Consumption. In: NetMob 2019, Oxford, UK (2019)
39. Barabási, A.-L.: Network Science. Cambridge University Press, Cambridge (2016). http://networksciencebook.com/
40. Salgado, D., Esteban, M.E., Novás, M., Saldaña, S., Sanguiao, L.: Data organisation and process design based on functional modularity for a standard production process. Journal of Official Statistics **34**(4), 811–833 (2018). doi:10.2478/jos-2018-0041
41. Oancea, B., Necula, M., Sanguiao, L., Salgado, D., Barragán, S.: A simulator for network event data. Technical report, Statistics Romania (INS) and Statistics Spain (INE) (December 2019). Deliverable I.2 of Work Package I of ESSnet on Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b9/WPI_Deliverable_I2_Data_Simulator_-_A_simulator_for_network_event_data.pdf
42. Shabbir, N., Sadiq, M.T., Kashif, H., Ullah, R.: Comparison of radio propagation models for long term evolution (LTE) network. International Journal of Next-Generation Networks **3**(3), 27–41 (2011). doi:10.5121/ijngn.2011.3303
43. Salgado, D., Sanguiao, L., Oancea, B., Barragán, S., Necula, M.: Collection of data sets and scripts for "An end-to-end statistical process with mobile network data for Official Statistics"(2020). https://figshare.com/articles/dataset/_/12861095
44. Caffery, J.J., Stuber, G.L.: Overview of radiolocation in CDMA cellular systems. IEEE Communications Magazine **36**(4), 38–45 (1998). doi:10.1109/35.667411
45. Dye, M., Baylin, F.: Mobile Positioning. Mobile Lifestreams Ltd, London (2001)
46. Gustafsson, F., Gunnarsson, F.: Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements. IEEE Signal Processing Magazine **22**(4), 41–53 (2005). doi:10.1109/msp.2005.1458284
47. Gezici, S.: A survey on wireless position estimation. Wireless Personal Communications **44**(3), 263–282 (2007). doi:10.1007/s11277-007-9375-z
48. Mohammadi, M., Molaei, E., Naserasadi, A.: A survey on location based services and positioning techniques. International Journal of Computer Applications **24**(5), 1–5 (2011). doi:10.5120/2946-3928
49. Liu, D., Sheng, B., Hou, F., Rao, W., Liu, H.: From wireless positioning to mobile positioning: An overview of recent advances. IEEE Systems Journal **8**(4), 1249–1259 (2014). doi:10.1109/jsyst.2013.2295136
50. Mahyuddin, M.F.M., M.Isa, A.A., Zin, M.S.I.M., A.H, A.M., Manap, Z., Abstract, M.K.I.: Overview of positioning techniques for LTE technology. Journal of Telecommunication, Electronics and Computer Engineering **9**(2-13), 43–50 (2017)
51. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989). doi:10.1109/5.18626
52. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Cambridge (2006)
53. Vanhoof, M., Reis, F., Ploetz, T., Smoreda, Z.: Assessing the quality of home detection from mobile phone data for official statistics. Journal of Official Statistics **34**(4), 935–960 (2018). doi:10.2478/jos-2018-0046
54. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Boston, MA (2012)
55. Long, J.A., Nelson, T.A.: A review of quantitative methods for movement data. International Journal of Geographical Information Science **27**(2), 292–318 (2013). doi:10.1080/13658816.2012.682578
56. McLean, D.J., Volponi, M.A.S.: trajr: An R package for characterisation of animal trajectories. Ethology **124**(6), 440–448 (2018). doi:10.1111/eth.12739

57. Lehmann, E.L., Casella, G.: Theory of Point Estimation. Springer, New York (2003)
58. Daskalakis, C., Kamath, G., Tzamos, C.: On the structure, covering, and learning of Poisson multinomial distributions (2015). doi:10.1109/FOCS.2015.77
59. Royle, A.J., Dorazio, R.M.: Hierarchical Modelling and Inference in Ecology, p. . Elsevier, New York (2009)
60. Bryant, J.R., Graham, P.J.: Bayesian demographic accounts: Subnational population estimation using multiple data sources. Bayesian Analysis **8**(3), 591–622 (2013). doi:10.1214/13-ba820
61. Bryant, J.R., Graham, P.: A Bayesian approach to population estimation with administrative data. Journal of Official Statistics **31**(3), 475–487 (2015). doi:10.1515/jos-2015-0028
62. Sanguiao, L., Barragán, S., Salgado, D.: destim: An R Package for Mobile Devices Position Estimation. (2020). R package version 0.1.0. https://github.com/Luis-Sanguiao/destim
63. Oancea, B., Barragán, S., Salgado, D.: deduplication: An R Package for Deduplicating Mobile Device Counts Into Population Individual Counts. (2020). R package version 0.1.0. https://github.com/bogdanoancea/deduplication
64. Oancea, B., Barragán, S., Salgado, D.: aggregation: An R Package to Produce Probability Distributions of Aggregate Number of Mobile Devices. (2020). R package version 0.1.0. https://github.org/bogdanoancea/aggregation
65. Oancea, B., Barragán, S., Salgado, D.: inference: R Package for Computing the Probability Distribution of the Number of Individuals in the Target Population. (2020). R package version 0.1.0. https://github.com/bogdanoancea/inference

Figures

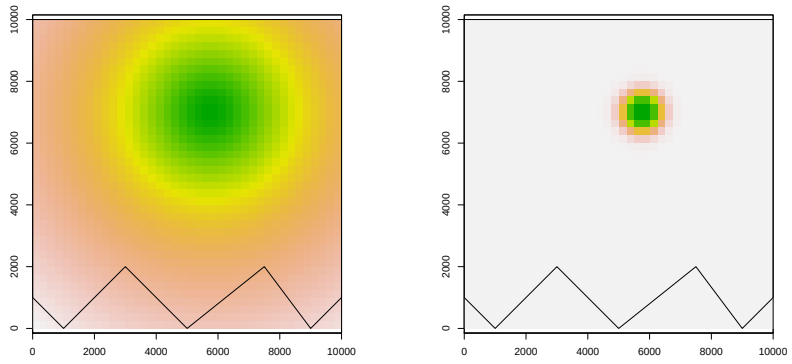

Figure 1 Received signal strength and signal dominance measure. Received signal strength and signal dominance measure for an omnidirectional antenna according to models (1) and (2).

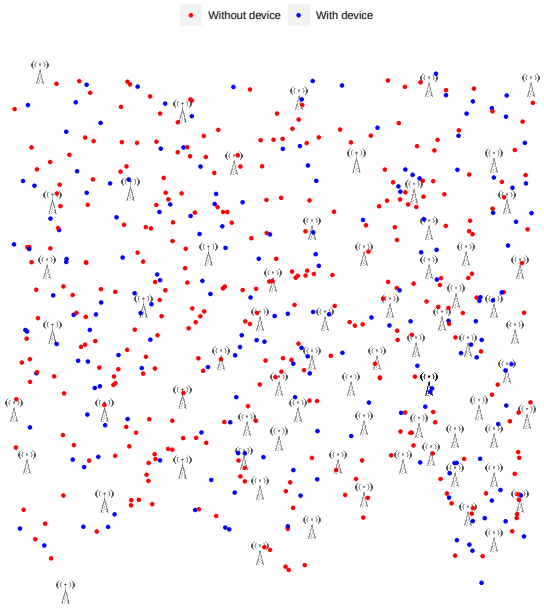

Figure 2 Network configuration. Positions of 70 antennas, positions of individuals at time $t = 0$. (See complementary online material for an animated gif for the displacement of individuals [43].

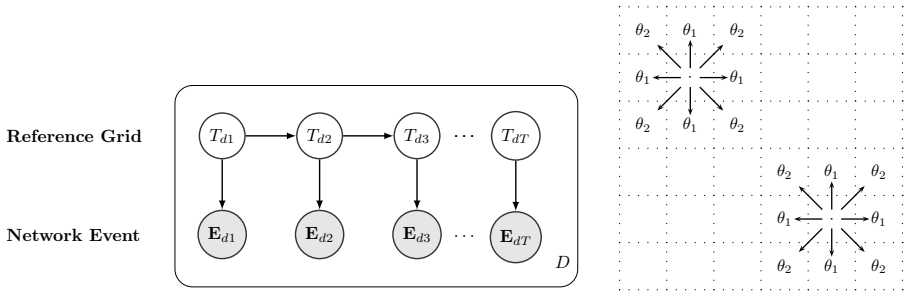

Figure 3 Graphical representation of a hidden Markov model. (Left) Observed variables are network variables $\mathbf{E}_{dt}$ and unobserved variables are tile locations T_{dt}. (Right) Transition probabilities in the reference grid for adjacent cells.

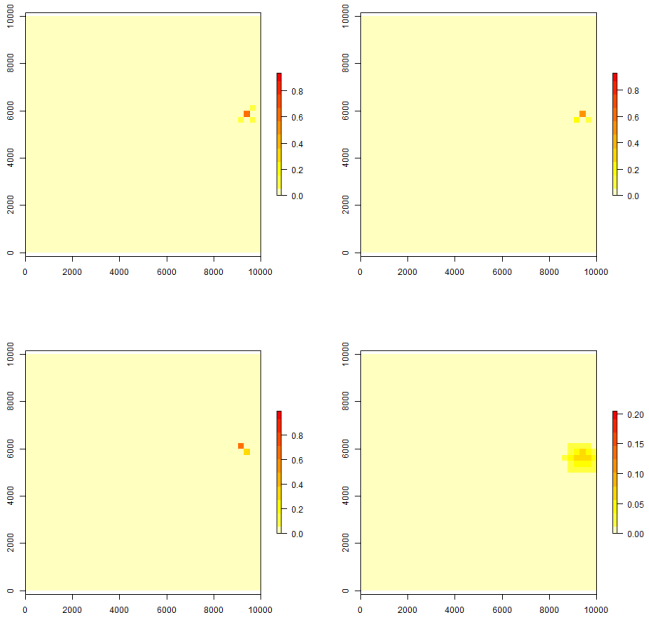

Figure 4 Posterior location probabilities. Posterior location probabilities at time $t = 0$ for a given device under models RSS with uniform prior (top left), RSS with network prior (top right), SDM with uniform prior (bottom left), and SDM with network prior (bottom right).

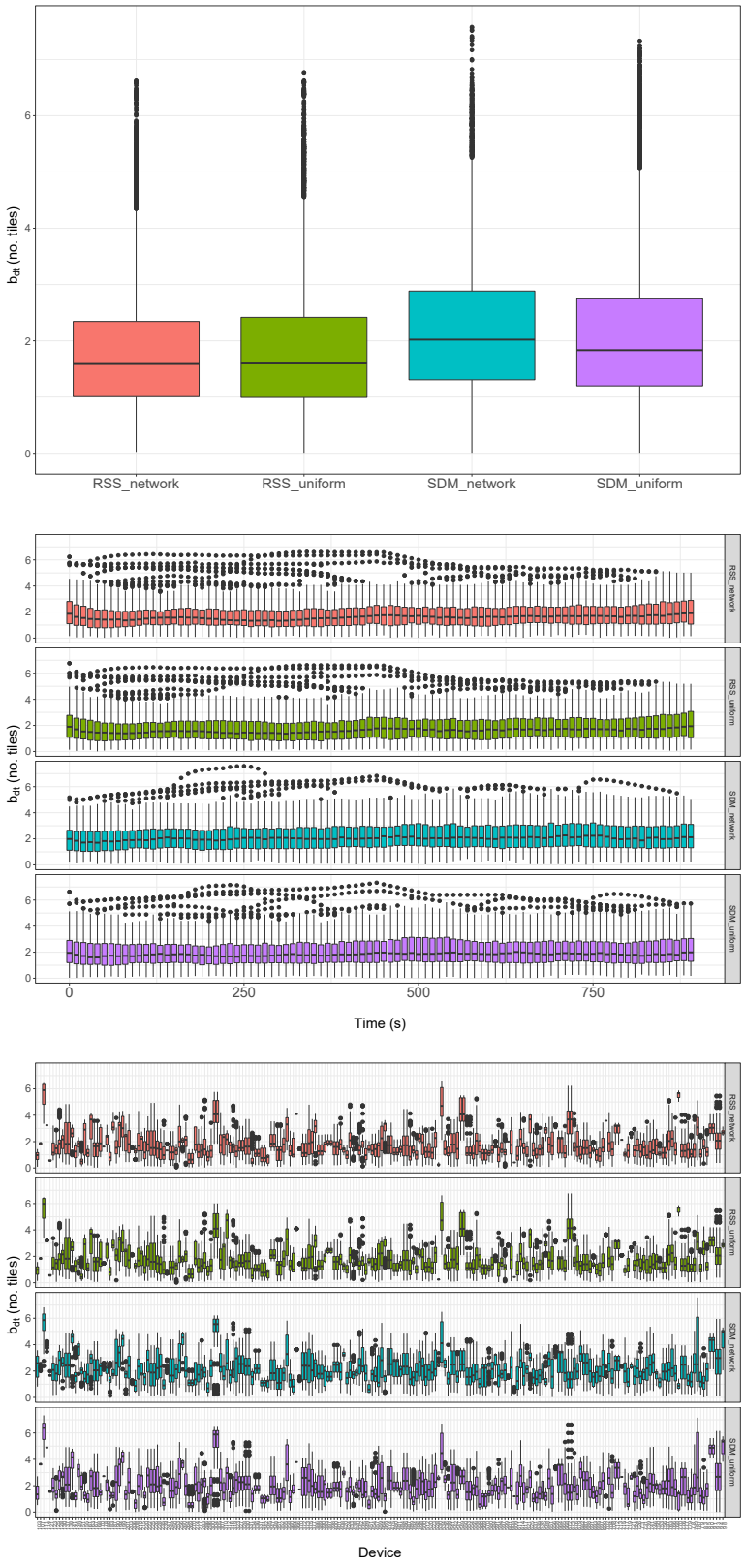

Figure 5 Distance between centers of location probabilities and true positions. (Top) Distribution of distance b_{dt} for all times t and all devices d for models RSS and SDM with uniform and network priors. (Middle) Time evolution of distributions of distance b_{dt} for all devices d for the same models. (Bottom) Distribution per device of distance b_{dt} for all times t for the same models.

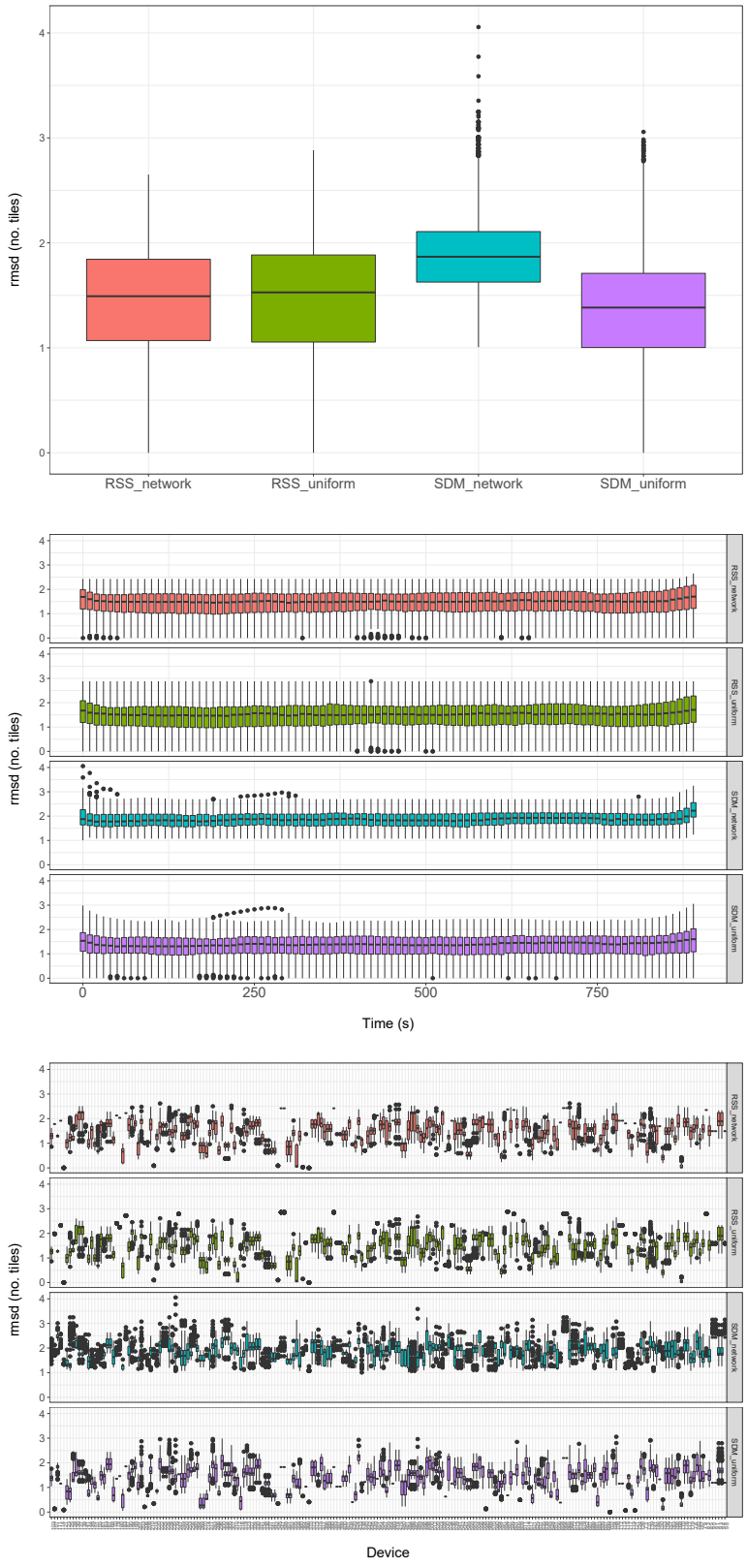

Figure 6 Root mean squared dispersions of location probabilities. (Top) Distribution of root mean squared dispersions rmsd_{dt} for models RSS and SDM with uniform and network priors. (Middle) Time evolution of distributions of distance rmsd_{dt} for the same models. (Bottom) Distribution per device of root mean squared dispersion rmsd_{dt} for all times t for the same models.

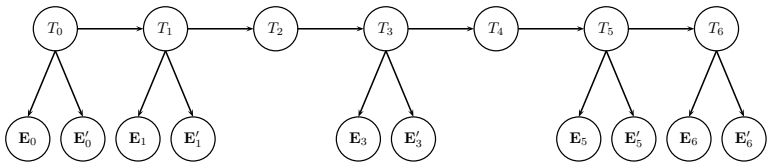

Figure 7 HMM duplicity model. Graphical model for an individual carrying two devices, thus generating pair of events $\mathbf{E}_{d1:T}$ and $\mathbf{E}_{d'1:T}$.

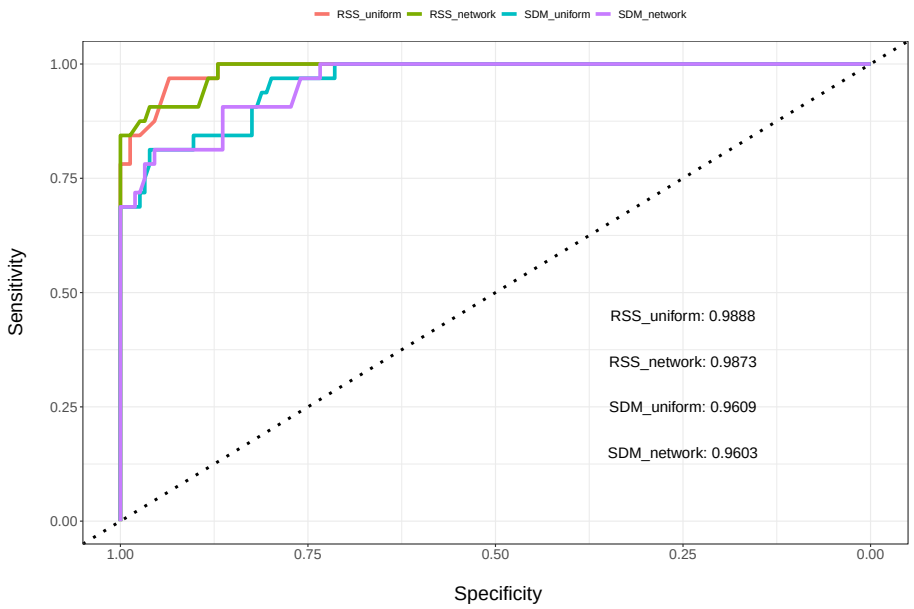

Figure 8 ROC curves for the duplicity probabilities $p_d^{(2)}$. ROC curves for the duplicity probabilities $p_d^{(2)} = 1 - p_d^{(1)}$ computed according to eq. (9) for HMMs with two priors and two emission models.

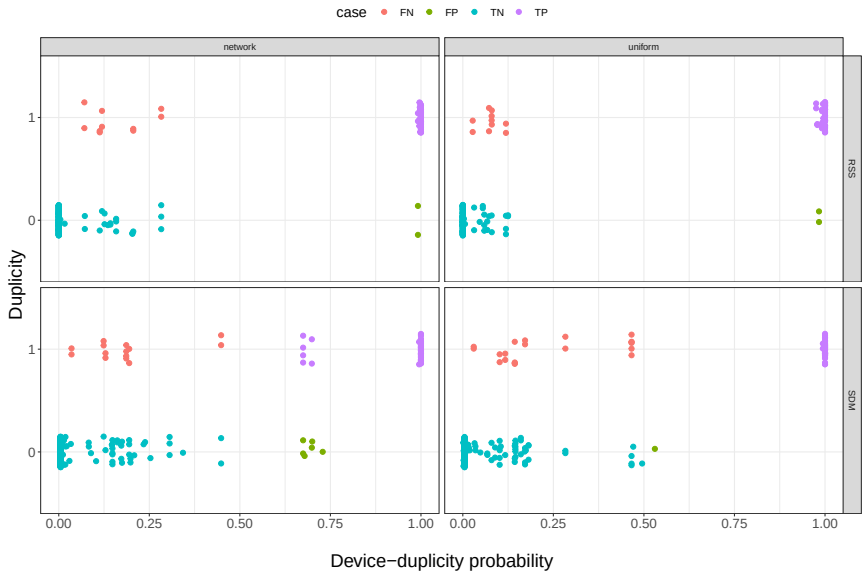

Figure 9 Cases for the duplicity probabilities $p_d^{(2)}$. True/false positive/negative cases for the duplicity probabilities $p_d^{(2)} = 1 - p_d^{(1)}$ computed according to eq. (9) for HMMs with two priors and two emission models.

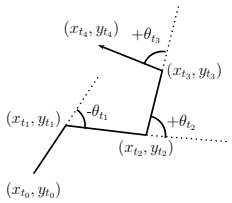

Figure 10 Turning angles of a path. Schematic representation of the sequence of turning angles $\pm\theta_t$ of a probability-based path.

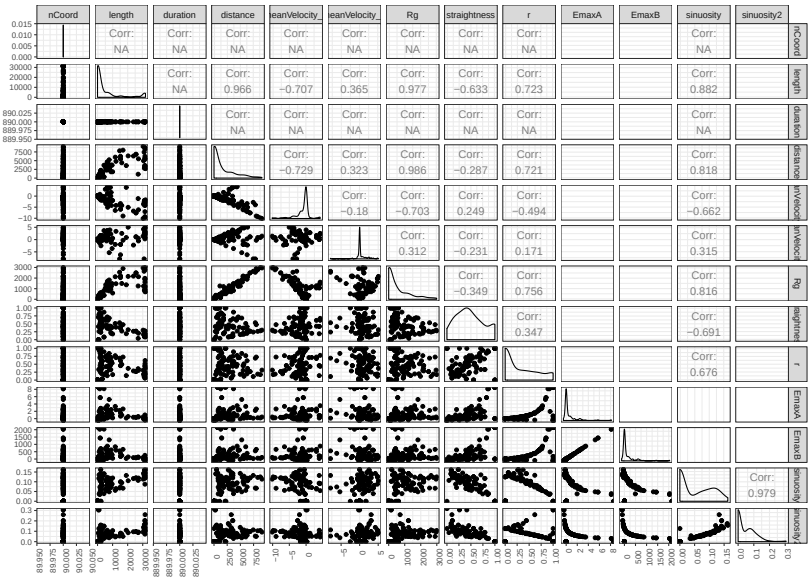

Figure 11 Path indicator distributions. I. Path indicators (`NA` correlations indicate static devices; blank correlations indicate infinite values involved, i.e. straight line paths).

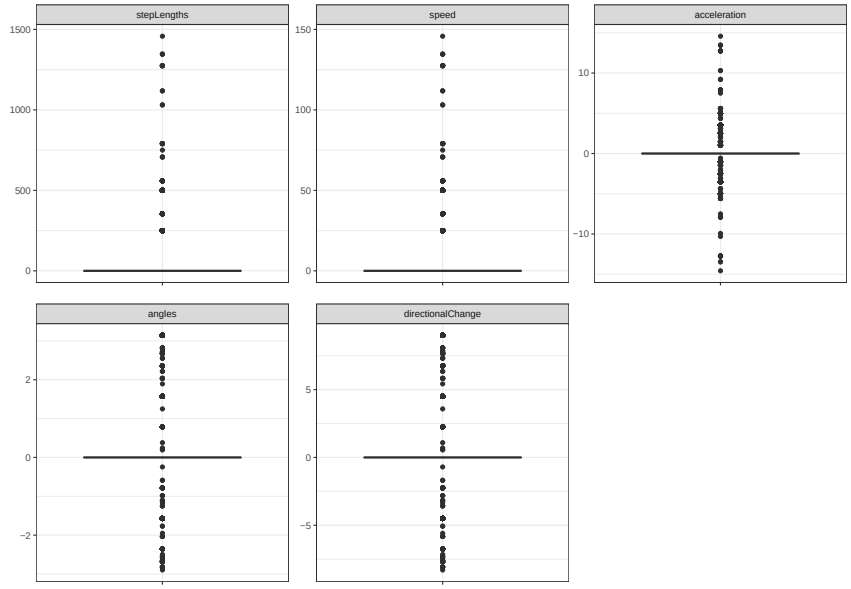

Figure 12 Path indicator distributions. II. Boxplots of the distributions of the five path indicators: (i) the step length, (ii) the speed, (iii) the acceleration, (iv) the turning angles, and (v) the directional change.

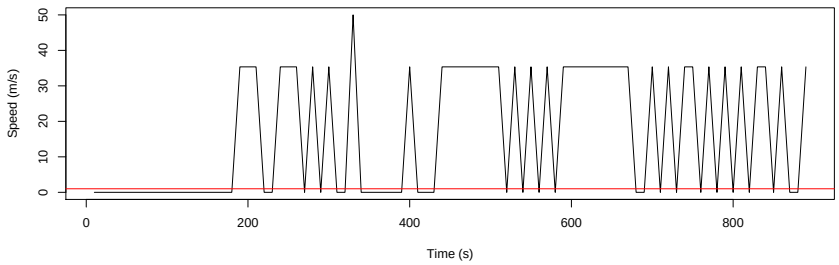

Figure 13 Time windows of stays. Time instants of a given device path with a speed below $1\ ms^{-1}$.

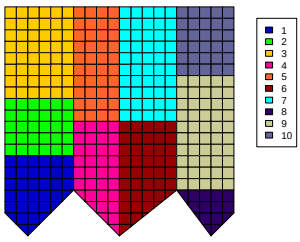

Figure 14 Regions as aggregated territorial units of analysis. Regions are obtained by aggregation of tiles of the reference grid.

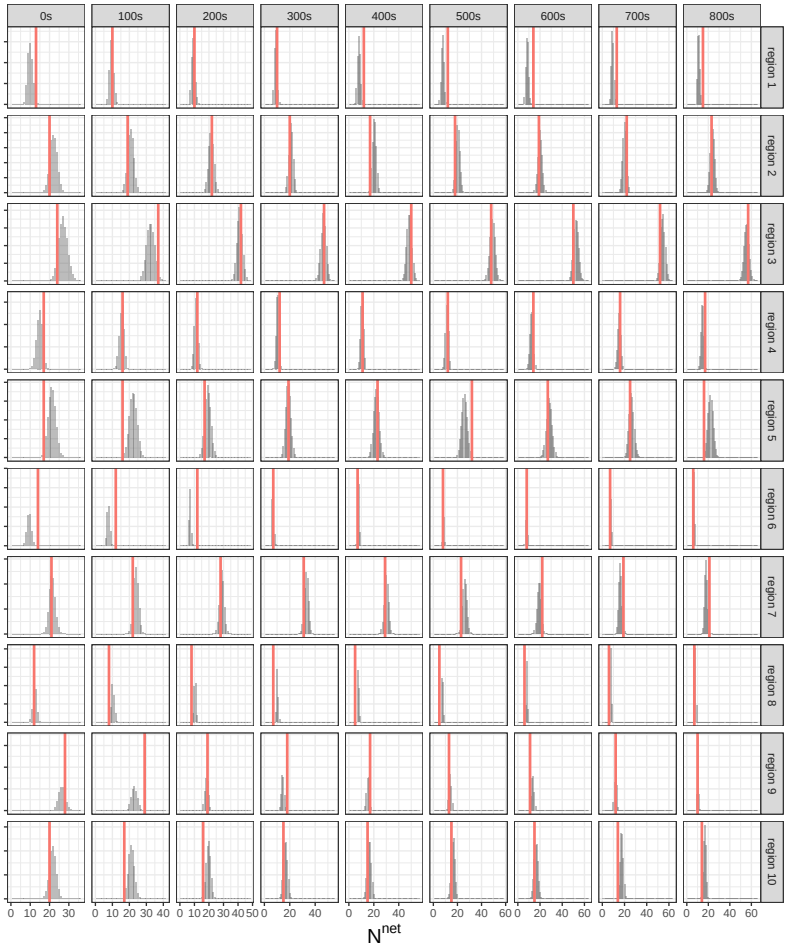

Figure 15 Posterior distributions for the number of individuals $\bar{N}_t^{\mathbf{net}}$ per region. Posterior distributions for the number of individuals $\bar{N}_t^{\mathrm{net}}$ per region detected by the network compared to true values (in red) computed according to eq. (17) using the RSS geolocation model with uniform prior. Only a sample of time instants is shown for visibility's sake.

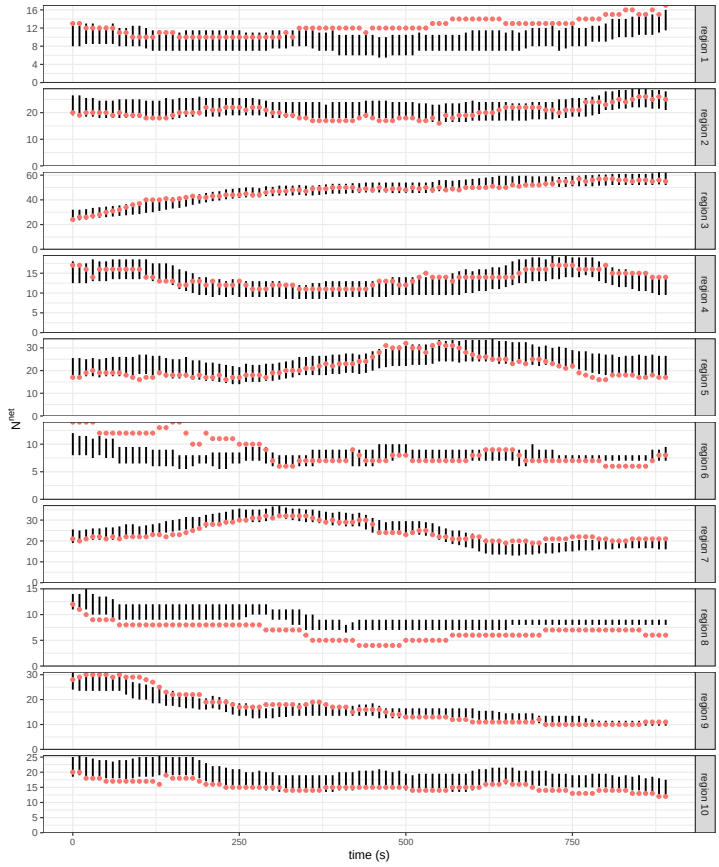

Figure 16 Credible intervals for the number of individuals $\bar{N}_t^{\mathrm{net}}$ per region. Credible intervals (prob $= 95\%$) for the number of individuals $\bar{N}_t^{\mathrm{net}}$ per region detected by the network compared to true values (in red) computed according to the distribution in eq. (17) using the RSS geolocation model with uniform prior.

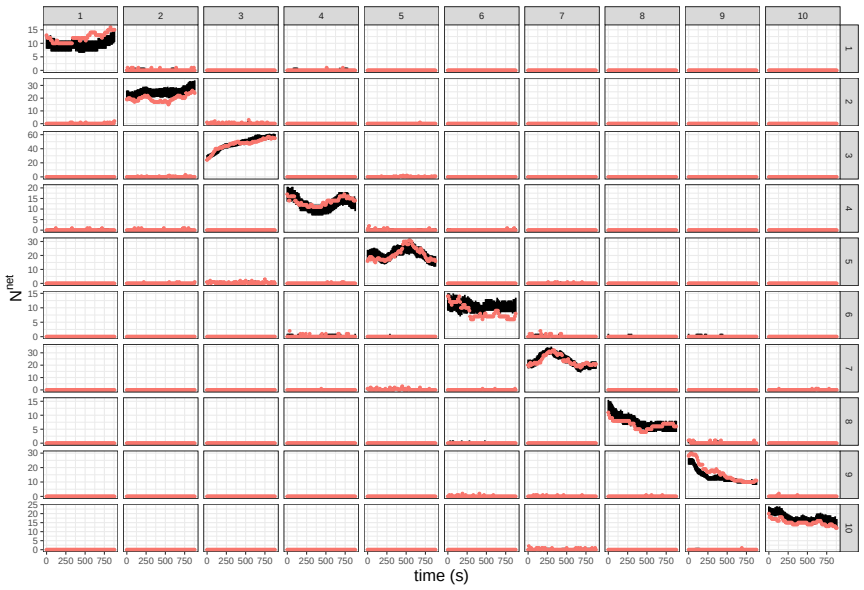

Figure 17 Credible intervals for the OD matrix of the number of individuals N_t^{net} per region. Credible intervals (prob $= 95\%$) for the OD matrix of number of individuals N_t^{net} per region detected by the network compared to true values (in red) computed according to the distribution in eq. (21) using the RSS geolocation model with uniform prior.

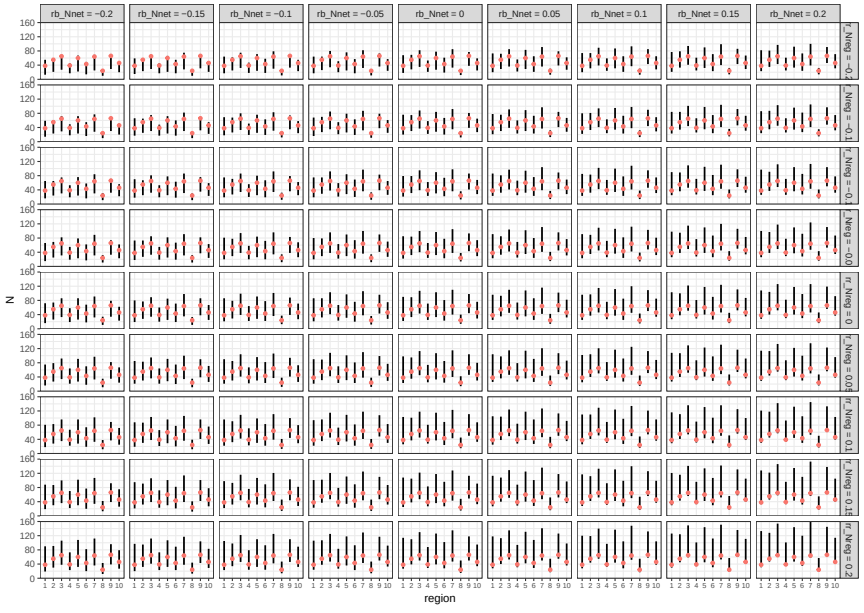

Figure 18 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob $= 95\%$) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the distribution in eq. (28) with the RSS geolocation model and uniform prior and different values of rb^{net} and rb^{reg}.

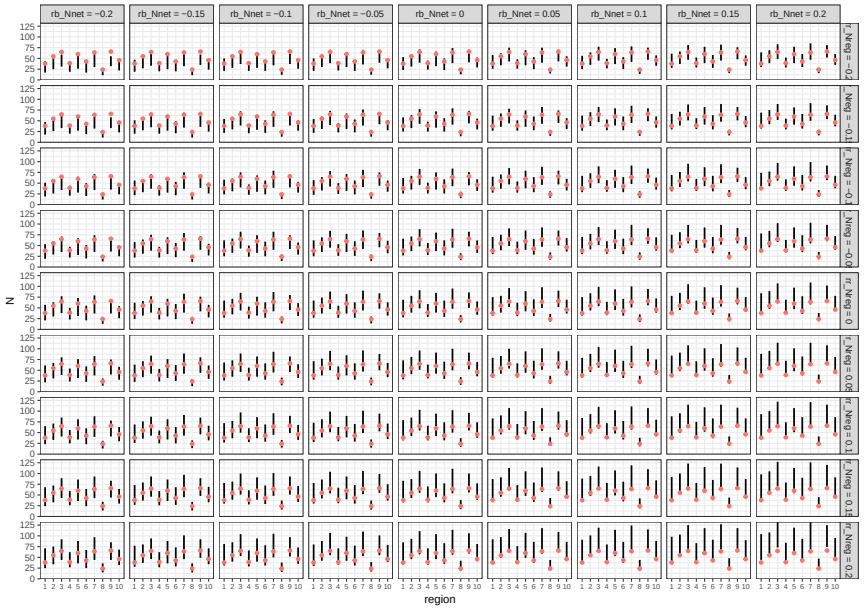

Figure 19 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob $= 95\%$) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the distribution in eq. (30) with the RSS geolocation model and uniform prior and different values of $\mathrm{rb}^{\mathrm{net}}$ and $\mathrm{rb}^{\mathrm{reg}}$.

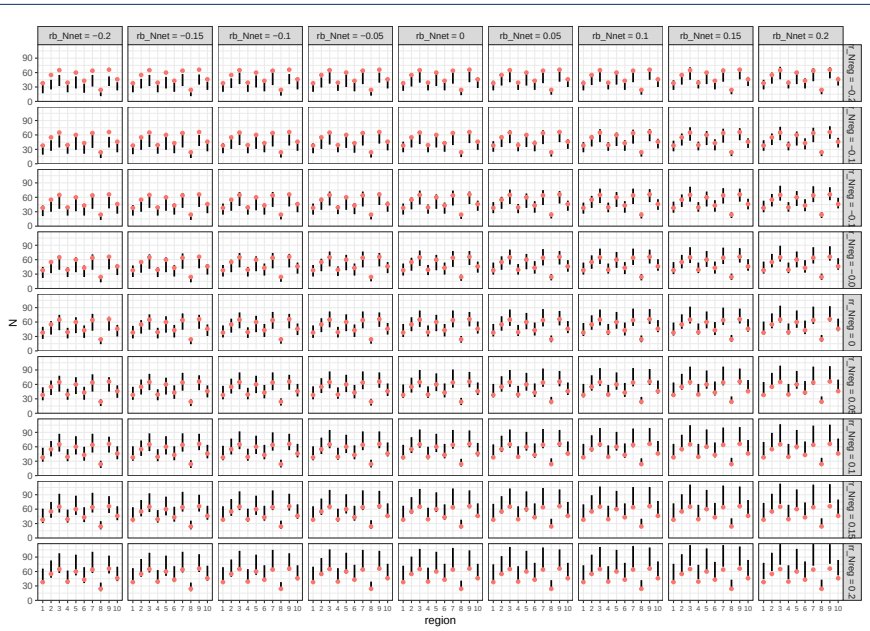

Figure 20 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob $= 95\%$) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the distribution in eq. (37) with the RSS geolocation model and uniform prior and different values of $\mathrm{rb}^{\mathrm{net}}$ and $\mathrm{rb}^{\mathrm{reg}}$.

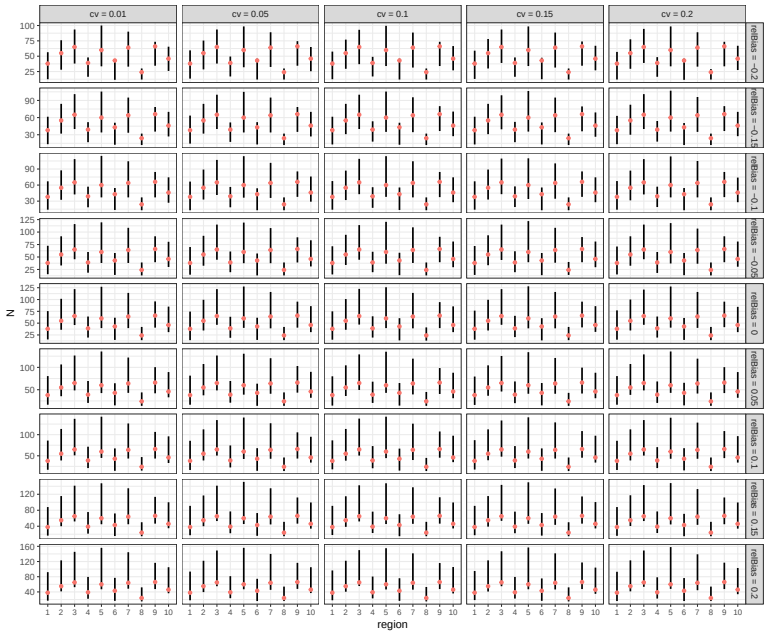

Figure 21 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob $= 95\%$) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the whole module integration using the RSS geolocation model with uniform prior and the beta negative binomial inference model. See animated gif `N_CI_bnb_t.gif` in [43] for the full time sequence of credible intervals and different values of $\mathrm{rb}^{\mathrm{reg}}$ and $\mathrm{cv}^{\mathrm{reg}}$.

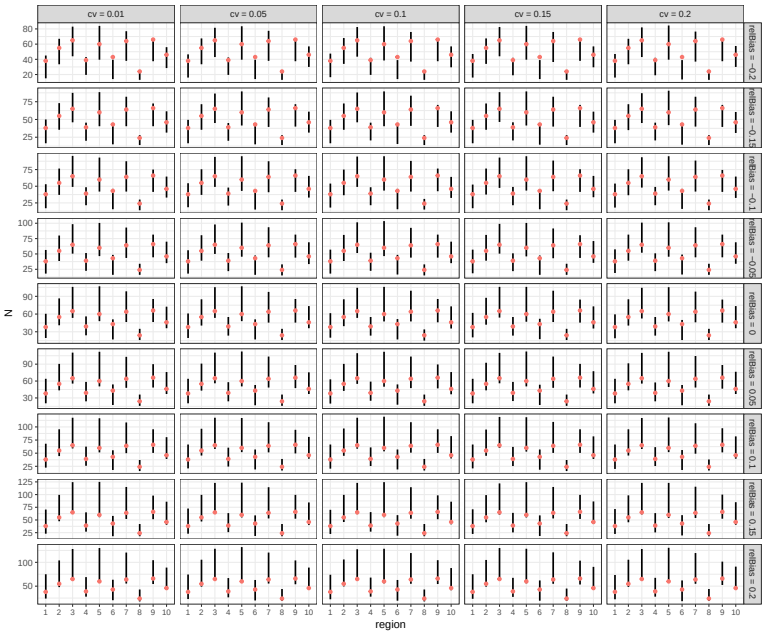

Figure 22 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob $= 95\%$) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the whole module integration using the RSS geolocation model with uniform prior and the negative binomial inference model and different values of $\mathrm{rb}^{\mathrm{reg}}$ and $\mathrm{cv}^{\mathrm{reg}}$.

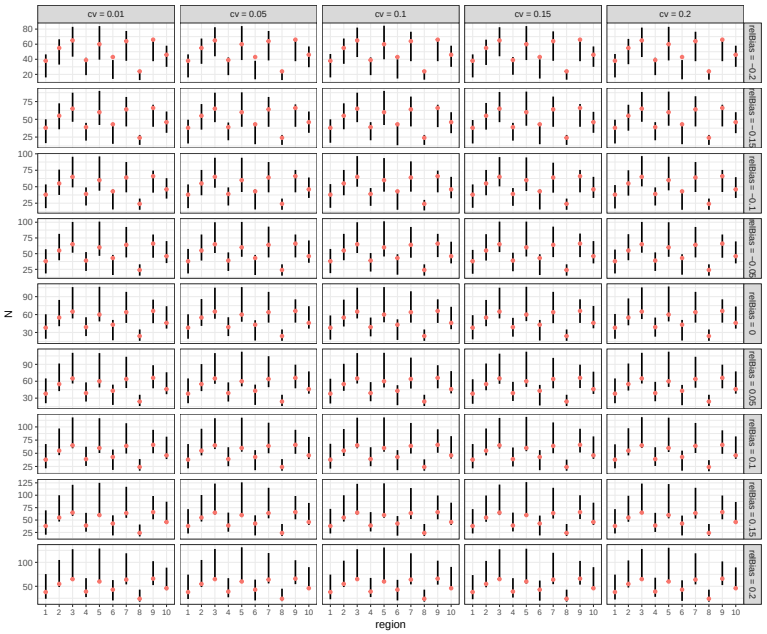

Figure 23 Credible intervals for the initial number of individuals N_0 per region. Credible intervals (prob $= 95\%$) for the initial number of individuals N_0 per region compared to true values (in red) computed according to the whole module integration using the RSS geolocation model with uniform prior and the negative binomial state process inference model and different values of $\mathrm{rb}^{\mathrm{reg}}$ and $\mathrm{cv}^{\mathrm{reg}}$.

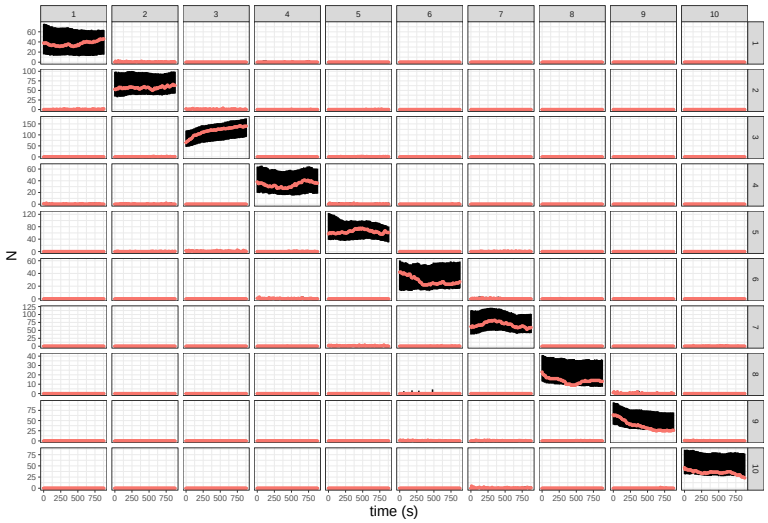

Figure 24 Credible intervals for the OD matrix. Credible intervals (prob $= 95\%$) for OD matrix compared to true values (in red) computed according to the whole module integration with the RSS geolocation model with uniform prior and the beta negative binomial with $\mathrm{rb}^{\mathrm{reg}} = 0$ and $\mathrm{cv}^{\mathrm{reg}} = 0.01$.

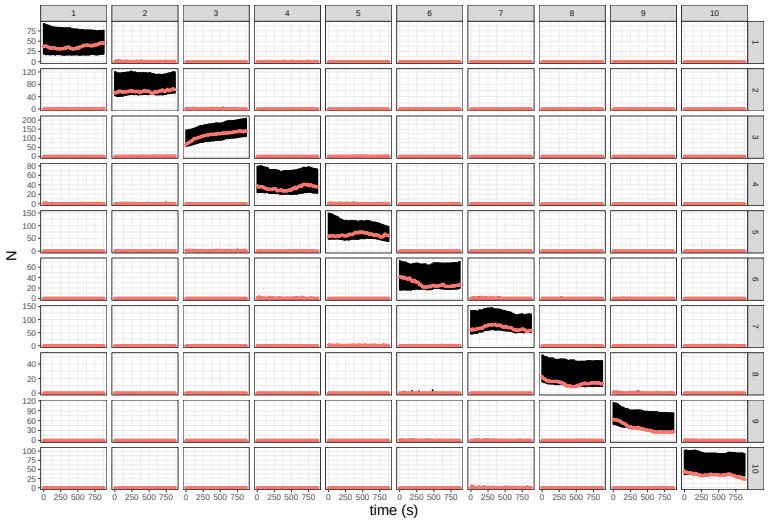

Figure 25 Credible intervals for the OD matrix. Credible intervals (prob $= 95\%$) for OD matrix compared to true values (in red) computed according to the whole module integration with the RSS geolocation model with uniform prior and the beta negative binomial with $\mathrm{rb}^{\mathrm{reg}} = 0.20$ and $\mathrm{cv}^{\mathrm{reg}} = 0.20$.

Additional Files

Additional file 1 — Supplementary material

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.

RESEARCH

Supplementary material for "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†]

*Correspondence: david.salgado.fernandez@ine.es
[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Paseo de la Castellana, 183, Madrid, Spain
Full list of author information is available at the end of the article
†The views expressed in this working paper are those of the authors and do not necessarily reflect the views of their affiliating institutions.

1 Introduction

A more complete list of published articles with statistical uses of mobile network data is provided in the references [1–73]. A great deal of unpublished work can also be found. We specifically recommend the conference series NetMob (`http://www.netmob.org`).

2 Data description

2.1 Scenario data

All input and output data for the simulator can be found at [URL]. Input data for the simulator are specified as xml files (`persons.xml`, `simulation.xml`, `antennas.xml`) and a wkt file for the irregular polygon (territory map). For the scenario used in the article, we have selected:

- `simulation.xml` contains general parameters for the simulation: see table 1.
- `persons.xml` contains general parameters for the displacement patterns: see table 2.
- `antennas.xml` contains parameters to configure each antenna: see table 3. We have configured 70 antennas with the marginal distributions included in table 4.

Output data from the simulator are obtained in csv format (we comment only the basic ones):

- `persons.csv` contains the real evolution of each individual, i.e. the ground truth. For each time instant t and each individual k, the simulator records the position coordinates x and y, the tile and the device(s) carried by the individual.
- `SignalMeasure_MNO1.csv` contains for each antenna the RSS in each tile of the reference grid.
- `AntennaInfo_MNO_MNO1.csv` contains the time sequence of connections. For each time instant t and each device, the simulator records the antenna to which it is connected, its true position coordinates (x, y) and tile, and a network event code for the type of connection.

For details about other parameters and files see [74].

3 Geolocation of mobile devices

3.1 Model construction

We include the mathematical details to compute the posterior location probabilities from the input data. This is conducted in steps:

1 Time discretization and padding.
2 Construction of the transition model.
3 Construction of the emission model.
4 Construction of the initial state (prior) distribution.
5 Computation of the likelihood function.
6 Parameter estimation (likelihood maximization).
7 Application of the forward-backward algorithm.

3.1.1 Time discretization and padding

We shall work in discrete times. To do this we need to relate three parameters, namely (i) the tile dimension l (we assume a square grid for simplicity), (ii) the time increment Δt between two consecutive time instants, and (iii) an upper bound $v_{\max}$ for the velocity of the individuals in the population. As we argued in the main text, we impose that in the time interval Δt, the device d at most can displace from one tile to an adjacent tile. Under this condition, we can trivially set $\Delta t \lesssim \frac{l}{v_{\max}}$. For example, if $v_{\max} = 150\text{km/h} \approx 42\text{ms}^{-1}$, then $\Delta t \lesssim \frac{100}{42} \approx 2\text{s}$. Conversely, if the time increment Δt is fixed, then the maximum distance in terms of the number of tiles will be $\lceil \frac{v_{\max} \cdot \Delta t}{l} \rceil$, which expresses the number of time instants to insert in the time sequence to guarantee the maximum one-tile distance restriction.

If in the dataset the device d is detected at longer time periods, e.g. once in a minute, then we artificially introduce missing values at intervals Δt between every two observed values. This artificial non-response allows us to work with parsimonious models easier to estimate instead of using more complex transition matrices.

Notice that we have used an a priori value for $v_{\max}$, but we can also possibly make an estimation using the observed values $\mathbf{E}_d(t)$ and geometrical considerations about the respective coverage areas and their mutual distance.

Additionally, each observed time instance t is approximated to its closest multiple integer of Δt. Thus, we will have as input data a sequence of time instants at multiples $t_n = \Delta t \cdot n, (n \geq 0)$ and a randomly alternate sequence of missing values and of observed event variables $\mathbf{E}_{t_n}$ (hereafter for ease of notation we drop out any reference to mobile device d since we are only focusing on one device).

3.1.2 Construction of the transition model

Now we specify a model for the transition between tiles (states) $\{T_i\}_{i=1,\ldots,N_T}$. For ease of explanation and notation, let us change the notation of each tile T_i to a two-dimensional index $T_{(i,j)}$. Accordingly, each tile will be specified in this section by a pair of integer coordinates. The correspondence between both enumerations is arbitrary, but fixed once it has been chosen. We again assume time homogeneity for simplicity. Thus, $\mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right)$ will denote $\mathbb{P}\left(T_{(r,s)}(t_n + \Delta t) \middle| T_{(i,j)}(t_n)\right)$ for any $t_n = 0, 1, \ldots$ We assume a square regular grid for simplicity.

Now, we make use of our preceding imposition by which an individual can at most reach an adjacent tile in time Δt. Thus,

$$\mathbb{P}\left(T_{(r,s)}\middle|T_{(i,j)}\right) = 0 \qquad \max\{|r-i|,|s-j|\} \geq 2, \qquad r,s,i,j = 1,\ldots,\sqrt{N_T}. \quad (1a)$$

Now, we assume that we have no further auxiliary information to model these transitions and impose rectangular isotropic conditions:

$$\begin{aligned} \mathbb{P}\left(T_{(i\pm1,j)}\middle|T_{(i,j)}\right) = \mathbb{P}\left(T_{(i,j\pm1)}\middle|T_{(i,j)}\right) &= \theta_1 \qquad i,j = 1,\ldots,\sqrt{N_T}, \quad (1b)\\ \mathbb{P}\left(T_{(i\pm1,j\pm1)}\middle|T_{(i,j)}\right) &= \theta_2 \qquad i,j = 1,\ldots,\sqrt{N_T}. \quad (1c)\end{aligned}$$

The last set of conditions is row-stochasticity:

$$\begin{aligned} \sum_{r,s=1}^{N_T} \mathbb{P}\left(T_{(r,s)}\middle|T_{(i,j)}\right) &= 1, \qquad i,j = 1,\ldots,\sqrt{N_T}, \quad (1d)\\ \mathbb{P}\left(T_{(r,s)}\middle|T_{(i,j)}\right) &\geq 0, \qquad i,j,r,s = 1,\ldots,\sqrt{N_T}.\end{aligned}$$

Now back to the original notation for tiles and using the usual notation for the transition matrix $A = [a_{ij}]$, with $a_{ij} = \mathbb{P}\left(T_{jt}\middle|T_{it}\right)$, conditions (1) amounts to having a highly sparse transition matrix A with up to 4 terms equal to θ_1 and θ_2 (each) per row and diagonal entries guaranteeing row-stochasticity.

For the generic case of a square grid with size N_T, we have

$$A(\theta_1,\theta_2) = \begin{bmatrix} D_1(\theta_1,\theta_2) & M(\theta_1,\theta_2) & O & O & \cdots & O \\ M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) & O & \cdots & O \\ O & M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) & \cdots & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) \\ O & O & O & O & M(\theta_1,\theta_2) & D_1(\theta_1,\theta_2) \end{bmatrix}, \quad (2)$$

where

$$D_1(\theta_1,\theta_2) = \begin{pmatrix} 1-2\theta_1-\theta_2 & \theta_1 & 0 & 0 & \cdots & 0 \\ \theta_1 & 1-3\theta_1-2\theta_2 & \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_1 & 1-3\theta_1-2\theta_2 & \theta_1 & \cdots & 0 \\ 0 & 0 & \theta_1 & 1-3\theta_1-2\theta_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-2\theta_1-\theta_2 \end{pmatrix}_{N_T\times N_T},$$

$$D_2(\theta_1,\theta_2) = \begin{pmatrix} 1-3\theta_1-2\theta_2 & \theta_1 & 0 & 0 & \cdots & 0 \\ \theta_1 & 1-4\theta_1-4\theta_2 & \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_1 & 1-4\theta_1-4\theta_2 & \theta_1 & \cdots & 0 \\ 0 & 0 & \theta_1 & 1-4\theta_1-4\theta_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-3\theta_1-2\theta_2 \end{pmatrix}_{N_T\times N_T},$$

$$M(\theta_1,\theta_2) = \begin{pmatrix} \theta_1 & \theta_2 & 0 & 0 & \cdots & 0 \\ \theta_2 & \theta_1 & \theta_2 & 0 & \cdots & 0 \\ 0 & \theta_2 & \theta_1 & \theta_2 & \cdots & 0 \\ 0 & 0 & \theta_2 & \theta_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \theta_1 \end{pmatrix}_{N_T\times N_T},$$

$$O = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}_{N_T\times N_T}.$$

Notice that $A(\theta_1,\theta_2)$ fulfills all restrictions (1). Indeed, in our proposed implementation, in order to seek future generalization, we will work with a generic block-tridiagonal matrix (2), where the restrictions (1a) leading to 0 have been included, and complemented with the rest of restrictions (1b)-(1d) in matrix form. Thus, we write $C\cdot \mathrm{vec}(\tilde{A}) = \mathbf{b}$, where $\mathrm{vec}(\tilde{A})$ stands for the non-null elements of A in vector form. The rows of $[C\ \mathbf{b}]$ encode each of the restrictions (1b), (1c), and (1d). For example, $a_{12}=\theta_1$ and $a_{21}=\theta_1$ produce a row like this

$$C_i\cdot \mathrm{vec}(\tilde{A}) = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \cdot \left(\ \cdots \cdot a_{12} \cdot \cdots \cdot a_{21} \cdot \cdots\ \right)^T = b_i = 0. \quad (3)$$

Thus, in our software implementation to test this proposal, we have considered a block-tridiagonal matrix like (2) together with a set of linear restrictions of the form $C\cdot \mathrm{vec}\left(\tilde{A}\right) = \mathbf{b}$.

3.1.3 Construction of the emission model

The emission model is specified by the HMM emission probabilities $b_{ik} = \mathbb{P}\left(\mathbf{E}_{t_n} = \mathbf{E}_k \middle| T_{t_n} = i\right)$, where $\mathbf{E}_k$ is a possible value for the observed event variables

$\mathbf{E}_{t_n}$ and i denotes the tile index. We assume time homogeneity. This conditional probability is computed using the radio wave propagation model of our choice:

$$b_{ik}^{\mathrm{RSS}} \propto \mathrm{RSS}(d(\mathbf{E}_k, T_i)) \tag{4}$$
$$b_{ik}^{\mathrm{SDM}} \propto \mathrm{SDM}(d(\mathbf{E}_k, T_i)), \tag{5}$$

where $d(\mathbf{E}_k, T_i)$ stands for the distance between the antenna generating the event $\mathbf{E}_k$ and tile T_i. The proportional constant is fixed to normalize the probability functions.

Up to this point we have as input data the sequence of observed and missing values $a_{t_n} \in \{0, 1, \ldots, N_A\}$ for $t_n = 0, 1, \ldots, T$. We already have the emission matrix B, too.

3.1.4 Construction of the initial state (prior) distribution

For illustrative purposes, we consider two choices: (i) uniform prior, i.e. $\pi_i^{\mathrm{uniform}} = \frac{1}{N_T}$ and (ii) $\pi_i^{\mathrm{network}} \propto \sum_k (\mathrm{RSS}(d(\mathbf{E}_k, T_i)))$ (where RSS is expressed in watts) or $\pi_i^{\mathrm{network}} \propto \sum_k (\mathrm{SDM}(d(\mathbf{E}_k, T_i)))$, depending on the emission model.

3.1.5 Computation of the likelihood

The likelihood is trivially computed using the numerical proviso of setting emission probabilities equal to 1 when there is a missing value in the observed variables (e.g. due to time padding). The general expression for the likelihood is

$$\begin{aligned} L(\mathbf{E}) &= \sum_{i_0=1}^{N_T} \cdots \sum_{i_T=1}^{N_T} \mathbb{P}(T_{t_0} = i_0) \prod_{n=1}^{N} \mathbb{P}\left(T_{t_n} = i_n | T_{t_{n-1}} = i_{n-1}\right) \mathbb{P}\left(E_{t_n} \middle| T_{t_n} = i_n\right) \\ &= \sum_{i_0=1}^{N_T} \cdots \sum_{i_T=1}^{N_T} \mathbb{P}(T_{t_0} = i_0) \prod_{n=1}^{N} a_{i_{n-1} i_n}(\boldsymbol{\theta}) b_{i_n k_{t_n}} \end{aligned} \tag{6a}$$

Notice that the emission probabilities only contribute numerically providing no parameter whatsoever to be estimated.

3.1.6 Parameter estimation

The estimation of the unknown parameters $\boldsymbol{\theta}$ is conducted maximizing the likelihood. The restrictions coming from the transition model (1) makes the optimization problem not trivial. Notice that the EM algorithm is not useful. Instead, we provide a taylor-made solution seeking for future generalizations with more realistic choices of transition probabilities incorporating land use information. Formally, the optimization problem is given by:

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \\ & a_k \in [0, 1], \end{aligned} \tag{7}$$

where $\mathbf{a}$ stands for the nonnull entries of the transition probability matrix A, the objective function $h(\mathbf{a})$ is derived from the likelihood L expressed in terms of the nonnull entries of the transition matrix A, and the system $C \cdot \mathbf{a} = \mathbf{b}$ expresses the sets of restrictions from the transition model (1) not involving null rhs terms (restrictions (1b), (1c), and (1d)).

Let us quantify the number of variables and restrictions in order to propose an abstract procedure possibly generalized to other situations. We illustrate this procedure for a square regular grid of size N_T. The number of zeroes in the transition matrix A can be computed as follows:

- There exist 4 rows in A corresponding to the 4 vertices in the grid. Each of these rows contains $N_T - 4$ zeroes.
- There exist 4 sets of $\sqrt{N_T} - 2$ rows in A corresponding to boundary tiles not being vertices. Each of these rows contains $N_T - 6$ zeroes.
- There exist $(\sqrt{N_T} - 2)^2$ rows in A corresponding to this same number of inner tiles. Each of these rows contains $N_T - 9$ zeroes.

Thus, the total number of zeroes in A is given by $4 \times (N_T - 4) + 4 \times (\sqrt{N_T} - 2) \times (N_T - 6) + (\sqrt{N_T} - 2)^2 \times (N_T - 9) = N_T^2 - 9 \cdot N_T + 12\sqrt{N_T} - 4$. The number of non-null components of $\mathbf{a}$ in problem (7) is $d = 9 \cdot N_T - 12\sqrt{N_T} + 4$.

The number of restrictions n_r not involving zeroes depends very sensitively on the particular transition model chosen for the displacements. In the rectangular isotropic model considered above, we need to identify the number of entries (i) equal to θ_1, (ii) equal to θ_2, and (iii) in the diagonal (thus guaranteeing the row-stochasticity restriction). Using the same counting procedure as above, the number of entries equal to θ_1 will be given by $4 \times 2 + 4 \times (\sqrt{N_T} - 2) \times 3 + (\sqrt{N_T} - 2)^2 \times 4 = 4 \cdot N_T - 4\sqrt{N_T}$. Since θ_1 is a free parameters we get $4 \cdot N_T - 4\sqrt{N_T} - 1$ rows. For θ_2, we get $4 \times (\sqrt{N_T} - 1)^2 - 1$ rows. From the row-stochasticity restriction we get N_T rows. Thus, the matrix C will have dimensions $n_r \times d$, with $n_r = 4 \cdot N_T - 4\sqrt{N_T} - 1 + 4 \times (\sqrt{N_T} - 1)^2 - 1 + N_T = 9 \cdot N_T - 12\sqrt{N_T} + 2$. Notice that $d - n_r = 2$, as expected, since we have two free parameters θ_1 and θ_2.

The abstract optimization problem is thus

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \ , \\ & \mathbf{a} \in [0,1]^d, \end{aligned} \tag{8}$$

where $C \in \mathbb{R}^{n_r \times d}$ and $\mathbf{b} \in \mathbb{R}^d$. The objective function $h(\mathbf{a})$ is indeed a polynomial in the non-null entries $\mathbf{a}$. This problem can be further simplified using the matrix QR decomposition. Write $C = Q \cdot R$, where Q is an orthogonal matrix of dimensions $n_r \times n_r$ and R is an upper triangular matrix of dimensions $n_r \times d$. Then we can rewrite the linear system as $R \cdot \mathbf{a} = Q^T \cdot \mathbf{b}$ and we can linearly solve variables $a_1, \ldots, a_{n_r}$ in terms of variables $a_{n_r+1}, \ldots, a_d$:

$$\begin{pmatrix} a_1 & \cdots & a_{n_r} \end{pmatrix}^T = \tilde{C}_{n_r \times (d-n_r)} \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T .$$

The system (8) then reduces to

$$\begin{array}{ll} \max & \tilde{h}(a_{n_r+1}, \ldots, a_d) \\ \text{s.t.} & 0 \leq \tilde{C} \cdot \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T \leq 1. \end{array} \tag{9}$$

In our current software implementation we resort to general-purpose optimizers. It remains for future work to find an optimised algorithm to solve (9). The solution $\mathbf{a}^*$ to problem (8) will be introduced in the transition probability matrix, which will thus be denoted by $\widehat{A}$.

3.1.7 Application of the forward-backward algorithm

Once the HMM has been fitted, we can readily apply the well-known forward-backward algorithm [see e.g. 75] to compute the target location probabilities γ_{it} and γ_{tij}. No novel methodological content is introduced at this point. For our implementation, we have used the scaled version of the algorithm (see [75]).

3.1.8 Software implementation

To carry out the computation described above upon the synthetic scenario generated by the network event data simulator we have used the prototyping R package called `destim` developed for these purposes by the authors [76]. This package contains a specific implementation of the rectangular geolocation model described in the preceding sections.

3.2 Model evaluation

The center of location probabilities and the root mean squared dispersion can be obtained naturally from a bias-variance decomposition of a mean squared distance. Let us denote by $\mathbf{R}_{dt} \in \{\mathbf{r}_i^{(c)}\}_{i=1,\ldots,N_T}$ the random vector for the position of a device according to the distribution of posterior location probabilities γ_{dti}. Let us shortly denote $\bar{\mathbf{R}}_{dt} \equiv \mathbb{E}\mathbf{R}_{dt} = \sum_{i=1}^{N_T} \gamma_{dti} \mathbf{r}_i^{(c)}$. Let us also denote the true position of device d at time t by $\mathbf{r}_{dt}^*$. Then, we can decompose

$$\begin{aligned} \mathrm{msd}_{dt} \equiv \mathbb{E}\|\mathbf{R}_{dt} - \mathbf{r}_{dt}^*\|^2 &= \mathbb{E}\|(\mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}) + (\bar{\mathbf{R}}_{dt} - \mathbf{r}_{dt}^*)\|^2 \\ &= \mathbb{E}\left[\langle \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}, \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt} \rangle\right] + \\ &\quad 2 \cdot \mathbb{E}\left[\langle \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}, \bar{\mathbf{R}}_{dt} - \mathbf{r}_{dt}^* \rangle\right] + \\ &\quad \mathbb{E}\left[\langle \bar{\mathbf{R}}_{dt} - \mathbf{r}_{dt}^*, \bar{\mathbf{R}}_{dt} - \mathbf{r}_{dt}^* \rangle\right] \\ &= \mathrm{rmsd}_{dt}^2 + \mathrm{b}_{dt}^2. \end{aligned} \tag{10}$$

This decomposition motivates the definition of the figures of merit proposed in the main text. We can also compare directly the mean squared distance (see figure 1). The overall performance is similar for the four models.

4 Device duplicity

4.1 The double-device emission model

To apply formulas for the computation of the device duplicity probabilities we need to compute the likelihood for the HMM model described above for each device separately and for each pair of devices according to figure 7 in the main text. To do this, we just need to have a new emission model producing a double event data sequence. The emission probabilities in this augmented model are computed using the original emission probabilities:

$$\mathbb{P}\left(\mathbf{E}_{dt}, \mathbf{E}_{d't} \middle| T_{dt}, \mathbf{I}^{\text{aux}}\right) = \mathbb{P}\left(\mathbf{E}_{dt} \middle| T_{dt}, \mathbf{I}^{\text{aux}}\right) \cdot \mathbb{P}\left(\mathbf{E}_{d't} \middle| T_{d't}, \mathbf{I}^{\text{aux}}\right). \tag{11}$$

Once these emission probabilities are computed, the computation of the likelihoods $\ell_{dd'}$ runs similar to the single-device case.

The computation depends on prior choice of the parameters λ_d, i.e. the ratio between the prior probability of no duplicity to the prior probability of duplicity. For the computation in the main text, this was chosen according to the parameters in the network event data simulator. In practice, this is not the case, but the MNO can provide a prior estimation of the number of subscribers with more than one device. In any case, we run the computation of $p_d^{(2)}$ for all d and checked the number of true/false positive/negative cases obtained. This is represented in figure 2, where we observed that around the chosen value, the classification is robust.

4.2 Software implementation

To carry out the computation described above upon the synthetic scenario generated by the network event data simulator we have used the prototyping R package called `deduplication` developed for these purposes by the authors [77]. This package implements the computation of the device duplicity probabilities described in the main text, including the computation of the double-device emission model for the underlying HMM. This package contains another two deduplication procedures based on pairwise comparisons and trajectory comparisons. In this work we have included only the alternative producing the best disambiguating method on our scenario.

5 Statistical filtering

To apply the proposed trajectory indicators to the synthetic scenario generated by the network event data simulator we have profusely used the R package `trajr` [78], with slight modifications on some functions to adequate to our trajectories.

6 Aggregation of individuals detected by a network

The core of the aggregation module is the generation of random multidimensional variates according to the Poisson-multinomial distribution as a sum of categorical (multinoulli) variables. This is directly implemented in the prototyping R package called `aggregation` developed for these purposes by the authors [79]. This package takes as input both the posterior location probabilities γ_{dti}, the device duplicity

probabilities $p_d^{(2)}$ for all devices d, and the spatial aggregation of tiles i into larger territorial units and produce n random multidimensional variates according to the Poisson-multinomial distribution defined by equation (17) in the main text. The package also implements the similar computation for the origin-destination matrix according to equation (21) in the main text.

7 Inference

The different models proposed for the inference module have been directly implemented using standard distributions in base R and package `extraDistr` [80], except for the continuous mixtures integrating the full hierarchy of levels for the observation and/or the state processes. The credible interval computations, both for the inference and the aggregation module, have been carried out using the R package `bayestestR` [81]. All credible intervals included in this work are high-density intervals [see e.g. 82]. This is all wrapped into the prototyping R package called `inference` developed for these purposes by the authors [83].

Declaration

Availability of data and materials
Data, scripts, and source code are freely available at [URL].

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

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).

Authors' contributions
All authors have contributed equally.

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.

Author details
[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Paseo de la Castellana, 183, 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.

References

1. Cáceres, N., Wideberg, J.P., ıtez, F.G.B.: Deriving origin–destination data from a mobile phone network. IET Intelligent Transport Systems **1**(1), 15 (2007). doi:10.1049/iet-its:20060020
2. Ahas, R., Aasa, A., Ülar Mark, Pae, T., Kull, A.: Seasonal tourism spaces in Estonia: Case study with mobile positioning data. Tourism Management **28**(3), 898–910 (2007). doi:10.1016/j.tourman.2006.05.010
3. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L.: Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical **41**(22), 224015 (2008). doi:10.1088/1751-8113/41/22/224015
4. González, M.C., Hidalgo, C.A., Barabási, A.-L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008). doi:10.1038/nature06958
5. Farrahi, K., Gatica-Perez, D.: Daily routine classification from mobile phone data. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) Machine Learning for Multimodal Interaction. Lecture Notes in Computer Science, vol. 5237, pp. 173–184. Springer, Berlin Heidelberg (2008). doi:10.1007/978-3-540-85853-9_16
6. Ahas, R., Aasa, A., Roose, A., Mark, U., Silm, S.: Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management **29**(3), 469–486 (2008). doi:10.1016/j.tourman.2007.05.014
7. Eagle, N., Pentland, A., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences **106**(36), 15274–15278 (2009). doi:10.1073/pnas.0900282106

8. Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M.: Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of Urban Technology **17**(1), 3–27 (2010). doi:10.1080/10630731003597306
9. Farrahi, K., Gatica-Perez, D.: Probabilistic mining of socio-geographic routines from mobile phone data. IEEE Journal of Selected Topics in Signal Processing **4**(4), 746–755 (2010). doi:10.1109/jstsp.2010.2049513
10. Sevtsuk, A., Ratti, C.: Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. Journal of Urban Technology **17**(1), 41–60 (2010). doi:10.1080/10630731003597322
11. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A.: Identifying important places in people's lives from cellular network data. In: Lecture Notes in Computer Science, pp. 133–151. Springer, ??? (2011). doi:10.1007/978-3-642-21726-5_9
12. Becker, R.A., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A Tale of One City: Using Cellular Network Data for Urban Planning. IEEE Pervasive Computing **10**(4), 18–26 (2011). doi:10.1109/MPRV.2011.44
13. Steenbruggen, J., Borzacchiello, M.T., Nijkamp, P., Scholten, H.: Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal **78**(2), 223–243 (2011). doi:10.1007/s10708-011-9413-y
14. Couronné, T., Smoreda, Z., Raimond, A.-M.O.: Chatty mobiles: individual mobility and communication patterns. CoRR **abs/1301.6553** (2011)
15. Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Prediction of socioeconomic levels using cell phone records. In: User Modeling, Adaption and Personalization, pp. 377–388. Springer, ??? (2011). doi:10.1007/978-3-642-22362-4_35
16. Blumenstock, J.E.: Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. Information Technology for Development **18**(2), 107–125 (2012). doi:10.1080/02681102.2011.643209
17. Cáceres, R., Rowland, J., Small, C., Urbanek, S.: Exploring the use of urban greenspace through cellular network activity. In: The Second Workshop on Pervasive Urban Applications (PURBA), in Conjunction with Pervasive (2012). http://www.kiskeya.net/ramon/work/pubs/purba12.pdf
18. Phithakkitnukoon, S., Smoreda, Z., Olivier, P.: Socio-geography of human mobility: A study using longitudinal mobile phone data. PLoS ONE **7**(6), 39253 (2012). doi:10.1371/journal.pone.0039253
19. Palmer, J.R.B., Espenshade, T.J., Bartumeus, F., Chung, C.Y., Ozgencil, N.E., Li, K.: New approaches to human mobility: Using mobile phones for demographic research. Demography **50**(3), 1105–1128 (2012). doi:10.1007/s13524-012-0175-z
20. Ferrari, L., Mamei, M., Colonna, M.: Discovering events in the city via mobile network analysis. Journal of Ambient Intelligence and Humanized Computing **5**(3), 265–277 (2012). doi:10.1007/s12652-012-0169-0
21. Smoreda, Z., Olteanu-Raimond, A.-M., Couronné, T.: Spatiotemporal data from mobile phones for personal mobility assessment. In: Zmud, J., Lee-Gosselin, M., Munizaga, M., Carrasco, J.A. (eds.) Transport Survey MethodsM Best Practice for Decision Making. Emerald, ??? (2013)
22. Becker, R., Volinsky, C., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A.: Human mobility characterization from cellular network data. Communications of the ACM **56**(1), 74 (2013). doi:10.1145/2398356.2398375
23. Calabrese, F., Diao, M., Lorenzo, G.D., Ferreira, J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies **26**, 301–313 (2013). doi:10.1016/j.trc.2012.09.009
24. Demissie, M.G., de Almeida Correia, G.H., Bento, C.: Exploring cellular network handover information for urban mobility analysis. Journal of Transport Geography **31**, 164–170 (2013). doi:10.1016/j.jtrangeo.2013.06.016
25. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences **111**(45), 15888–15893 (2014). doi:10.1073/pnas.1408439111
26. Louail, T., Lenormand, M., Ros, O.G.C., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M.: From mobile phone data to the spatial structure of cities. Scientific Reports **4**(1) (2014). doi:10.1038/srep05276
27. Li, W., Cheng, X., Duan, Z., Yang, D., Guo, G.: A framework for spatial interaction analysis based on large-scale mobile phone data. Computational Intelligence and Neuroscience **2014**, 1–11 (2014). doi:10.1155/2014/363502
28. Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C.: Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies **40**, 63–74 (2014). doi:10.1016/j.trc.2014.01.002
29. Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., Pujolle, G.: Estimating human trajectories and hotspots through mobile phone data. Computer Networks **64**, 296–307 (2014). doi:10.1016/j.comnet.2014.02.011
30. Calabrese, F., Ferrari, L., Blondel, V.D.: Urban sensing using mobile phone network data: A survey of research. ACM Computing Surveys **47**(2), 1–20 (2014). doi:10.1145/2655691
31. Chi, G., Thill, J.-C., Tong, D., Shi, L., Liu, Y.: Uncovering regional characteristics from mobile phone data: A network science approach. Papers in Regional Science **95**(3), 613–631 (2014). doi:10.1111/pirs.12149
32. Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M., Zook, M.: Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn. International Journal of Geographical Information Science **29**(11), 2017–2039 (2015). doi:10.1080/13658816.2015.1063151
33. Alexander, L., Jiang, S., Murga, M., González, M.C.: Origin-destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies **58**, 240–250 (2015). doi:10.1016/j.trc.2015.02.018
34. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ Data Science **4**(1) (2015). doi:10.1140/epjds/s13688-015-0046-0
35. Horn, C., Kern, R.: Deriving public transportation timetables with large-scale cell phone data. Procedia Computer Science **52**, 67–74 (2015). doi:10.1016/j.procs.2015.05.026

36. Steenbruggen, J., Tranos, E., Nijkamp, P.: Data from mobile phone operators: A tool for smarter cities? Telecommunications Policy **39**(3-4), 335–346 (2015). doi:10.1016/j.telpol.2014.04.001
37. Doyle, J., Hung, P., Farrell, R., McLeone, S.: Population mobility dynamics estimatedfrom mobile telephony data. Journal of Urban Technology **21**, 109–132 (2014)
38. Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J.J., Dubernet, T., Frías-Martínez, E.: Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. Transportation **42**(4), 647–668 (2015). doi:10.1007/s11116-015-9594-1
39. Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., Li, Q.: Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation **42**(4), 625–646 (2015). doi:10.1007/s11116-015-9597-y
40. Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C.: Analyzing cell phone location data for urban travel. Transportation Research Record: Journal of the Transportation Research Board **2526**, 126–135 (2015). doi:10.3141/2526-14
41. Janecek, A., Valerio, D., Hummel, K.A., Ricciato, F., Hlavacs, H.: The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. IEEE Transactions on Intelligent Transportation Systems **16**(5), 2551–2572 (2015). doi:10.1109/tits.2015.2413215
42. Trasarti, R., Olteanu-Raimond, A.-M., Nanni, M., Couronné, T., Furletti, B., Giannotti, F., Smoreda, Z., Ziemlicki, C.: Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. Telecommunications Policy **39**(3-4), 347–362 (2015). doi:10.1016/j.telpol.2013.12.002
43. Tranos, E., Nijkamp, P.: Mobile phone usage in complex urban systems: a space–time, aggregated human activity study. Journal of Geographical Systems **17**(2), 157–185 (2015). doi:10.1007/s10109-015-0211-9
44. Douglass, R.W., Meyer, D.A., Ram, M., Rideout, D., Song, D.: High resolution population estimates from telecommunications data. EPJ Data Science **4**(1) (2015). doi:10.1140/epjds/s13688-015-0040-6
45. Dobra, A., Williams, N.E., Eagle, N.: Spatiotemporal detection of unusual human population behavior using mobile phone data. PLOS ONE **10**(3), 0120449 (2015). doi:10.1371/journal.pone.0120449
46. Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., Smoreda, Z.: Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations. Transportation Research Procedia **11**, 381–398 (2015). doi:10.1016/j.trpro.2015.12.032
47. Bajardi, P., Delfino, M., Panisson, A., Petri, G., Tizzoni, M.: Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Science **4**(1) (2015). doi:10.1140/epjds/s13688-015-0041-5
48. Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., Giannotti, F.: An analytical framework to nowcast well-being using mobile phone data. International Journal of Data Science and Analytics **2**(1-2), 75–92 (2016). doi:10.1007/s41060-016-0013-2
49. Ponieman, N.B., Sarraute, C., Minnoni, M., Travizano, M., Zivic, P.R., Salles, A.: Mobility and sociocultural events in mobile phone data records. AI Communications **29**(1), 77–86 (2016). doi:10.3233/AIC-150687
50. Chua, A., Servillo, L., Marcheggiani, E., Moere, A.V.: Mapping cilento: Using geotagged social media data to characterize tourist flows in southern italy. Tourism Management **57**, 295–310 (2016). doi:10.1016/j.tourman.2016.06.013
51. Raun, J., Ahas, R., Tiru, M.: Measuring tourism destinations using mobile tracking data. Tourism Management **57**, 202–212 (2016). doi:10.1016/j.tourman.2016.06.006
52. Lu, S., Fang, Z., Zhang, X., Shaw, S.-L., Yin, L., Zhao, Z., Yang, X.: Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. ISPRS International Journal of Geo-Information **6**(1), 7 (2017). doi:10.3390/ijgi6010007
53. Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z., Colizza, V.: Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. Royal Society Open Science **4**(5), 160950 (2017). doi:10.1098/rsos.160950
54. Bwambale, A., Choudhury, C.F., Hess, S.: Modelling trip generation using mobile phone data: A latent demographics approach. Journal of Transport Geography (2017). doi:10.1016/j.jtrangeo.2017.08.020
55. Ricciato, F., Widhalm, P., Pantisano, F., Craglia, M.: Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. Pervasive and Mobile Computing **35**, 65–82 (2017). doi:10.1016/j.pmcj.2016.04.009
56. Song, X., Ouyang, Y., Du, B., Wang, J., Xiong, Z.: Recovering individual's commute routes based on mobile phone data. Mobile Information Systems **2017**, 1–11 (2017). doi:10.1155/2017/7653706
57. Tolouei, R., Psarras, S., Prince, R.: Origin-destination trip matrix development: Conventional methods versus mobile phone data. Transportation Research Procedia **26**, 39–52 (2017). doi:10.1016/j.trpro.2017.07.007
58. Fiadino, P., Ponce-Lopez, V., Torrero-Gonzalez, J.A., Torrent-Moreno, M., D'Alconzo, A.: Call detail records for human mobility studies. In: Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks - Big-DAMA '17. ACM Press, ??? (2017). doi:10.1145/3098593.3098601
59. Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., Smoreda, Z.: A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. IEEE Transactions on Mobile Computing **16**(10), 2682–2696 (2017). doi:10.1109/tmc.2016.2637901
60. Celik, S.C., Incel, O.D.: Semantic place prediction from crowd-sensed mobile phone data. Journal of Ambient Intelligence and Humanized Computing **9**(6), 2109–2124 (2017). doi:10.1007/s12652-017-0549-6
61. Jiang, S., Ferreira, J., González, M.C.: Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. IEEE Transactions on Big Data **3**(2), 208–219 (2017). doi:10.1109/tbdata.2016.2631141
62. Masso, A., Silm, S., Ahas, R.: Generational differences in spatial mobility: A study with mobile phone data. Population, Space and Place **25**(2), 2210 (2018). doi:10.1002/psp.2210
63. Anda, C., nez Medina, S.A.O., Fourie, P.: Multi-agent urban transport simulations using OD matrices from mobile phone data. Procedia Computer Science **130**, 803–809 (2018). doi:10.1016/j.procs.2018.04.139
64. Graells-Garrido, E., Caro, D., Parra, D.: Inferring modes of transportation using mobile phone data. EPJ Data Science **7**(1) (2018). doi:10.1140/epjds/s13688-018-0177-1

65. Thuillier, E., Moalic, L., Lamrous, S., Caminada, A.: Clustering weekly patterns of human mobility through mobile phone data. IEEE Transactions on Mobile Computing **17**(4), 817–830 (2018). doi:10.1109/tmc.2017.2742953
66. Sørensen, A.Ø., Bjelland, J., Bull-Berg, H., Landmark, A.D., Akhtar, M.M., Olsson, N.O.E.: Use of mobile phone data for analysis of number of train travellers. Journal of Rail Transport Planning & Management **8**(2), 123–144 (2018). doi:10.1016/j.jrtpm.2018.06.002
67. Li, Z., Yu, L., Gao, Y., Wu, Y., Song, G., Gong, D.: Identifying temporal and spatial characteristics of residents' trips from cellular signaling data: Case study of beijing. Transportation Research Record: Journal of the Transportation Research Board **2672**(42), 81–90 (2018). doi:10.1177/0361198118793495
68. Liu, Z., Ma, T., Du, Y., Pei, T., Yi, J., Peng, H.: Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. Transactions in GIS **22**(2), 494–513 (2018). doi:10.1111/tgis.12323
69. Wang, Z., He, S.Y., Leung, Y.: Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society **11**, 141–155 (2018). doi:10.1016/j.tbs.2017.02.005
70. Chen, J., Pei, T., Shaw, S.-L., Lu, F., Li, M., Cheng, S., Liu, X., Zhang, H.: Fine-grained prediction of urban population using mobile phone location data. International Journal of Geographical Information Science **32**(9), 1770–1786 (2018). doi:10.1080/13658816.2018.1460753
71. Batran, M., Mejia, M., Kanasugi, H., Sekimoto, Y., Shibasaki, R.: Inferencing human spatiotemporal mobility in greater maputo via mobile phone big data mining. ISPRS International Journal of Geo-Information **7**(7), 259 (2018). doi:10.3390/ijgi7070259
72. Bachir, D., Khodabandelou, G., Gauthier, V., Yacoubi, M.E., Puchinger, J.: Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transportation Research Part C: Emerging Technologies **101**, 254–275 (2019). doi:10.1016/j.trc.2019.02.013
73. Demissie, M.G., Phithakkitnukoon, S., Kattan, L., Farhan, A.: Understanding human mobility patterns in a developing country using mobile phone data. Data Science Journal **18** (2019). doi:10.5334/dsj-2019-001
74. Oancea, B., Necula, M., Sanguiao, L., Salgado, D., Barragán, S.: A simulator for network event data. Technical report, Statistics Romania (INS) and Statistics Spain (INE) (December 2019). Deliverable I.2 of Work Package I of ESSnet on Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b9/WPI_Deliverable_I2_Data_Simulator_-_A_simulator_for_network_event_data.pdf
75. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Cambridge (2006)
76. Sanguiao, L., Barragán, S., Salgado, D.: destim: An R Package for Mobile Devices Position Estimation. (2020). R package version 0.1.0. https://github.com/Luis-Sanguiao/destim
77. Oancea, B., Barragán, S., Salgado, D.: deduplication: An R Package for Deduplicating Mobile Device Counts Into Population Individual Counts. (2020). R package version 0.1.0. https://github.com/bogdanoancea/deduplication
78. McLean, D.J., Volponi, M.A.S.: trajr: An R package for characterisation of animal trajectories. Ethology **124**(6), 440–448 (2018). doi:10.1111/eth.12739
79. Oancea, B., Barragán, S., Salgado, D.: aggregation: An R Package to Produce Probability Distributions of Aggregate Number of Mobile Devices. (2020). R package version 0.1.0. https://github.org/bogdanoancea/aggregation
80. Wolodzko, T.: extraDistr: Additional Univariate and Multivariate Distributions. (2019). R package version 1.8.11. https://CRAN.R-project.org/package=extraDistr
81. Makowski, D., Ben-Shachar, M., Lüdecke, D.: bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. Journal of Open Source Software **4**(40), 1541 (2019). doi:10.21105/joss.01541
82. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A.: Bayesian Data Analysis. Taylor & Francis Ltd, ??? (2013)
83. Oancea, B., Barragán, S., Salgado, D.: inference: R Package for Computing the Probability Distribution of the Number of Individuals in the Target Population. (2020). R package version 0.1.0. https://github.com/bogdanoancea/inference

Figures

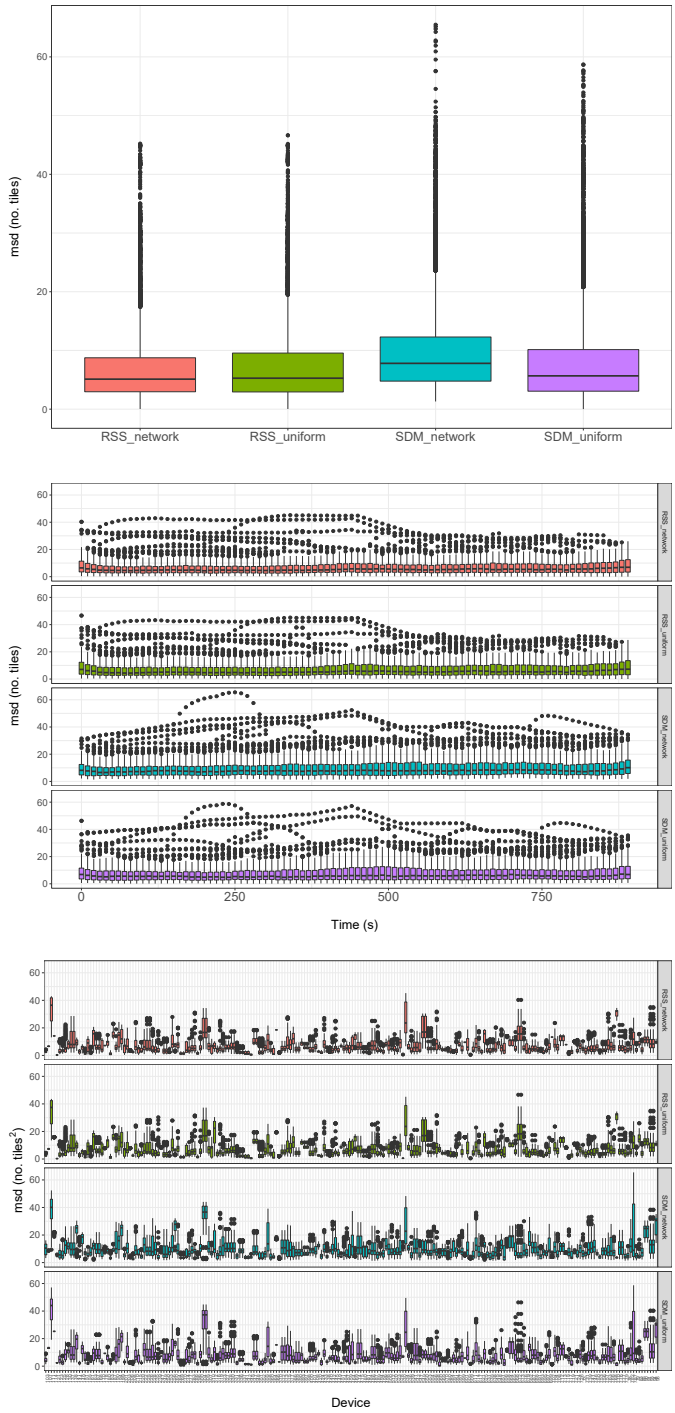

Figure 1 Mean squared distance. (Top) Distribution of mean squared distances msd_{dt} for models RSS and SDM with uniform and network priors. (Middle) Time evolution of distributions of mean squared distances msd_{dt} for the same models. (Bottom) Distribution per device of mean square distance msd_{dt} for all times t for the same models.

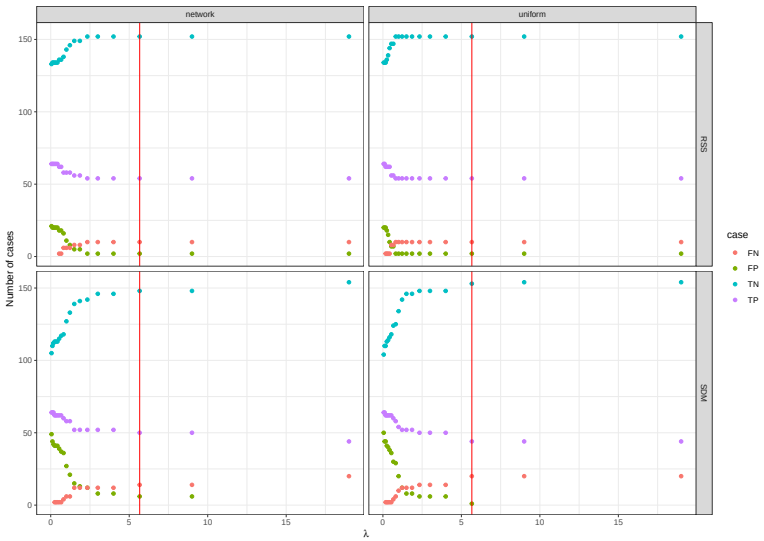

Figure 2 Device duplicity classification for different λ. Number of true/false positive/negative cases in the classification of device duplicity for different values of the prior ratio $\lambda_d = \lambda$ for all devices d. The red vertical lines represent the value used in the main text.

Tables

Table 1 Simulation parameters. Generic parameters included in `simulation.xml`.

Time (s)	
start_time	0
end_time	900
time_increment	10
time_stay	20
interval_btw_stays	120

MNO	
name	MNO1
device_share	0.35

Others	
displacement	random walk w/ drift
connection_type	strength
connection_threshold	-85dBm
grid_tile_dimensions	250 m×250 m

Table 2 Persons parameters. Parameters included in `persons.xml` (not exhaustive).

Persons	
num_persons	500
speed_walk	3 ms^{-1}
speed_car	16 ms^{-1}

Table 3 Antennas parameters. Parameters per antenna included in `antennas.xml`.

Antenna	
MNO_name	MNO1
max_connections	56
power	10
attenuationfactor	3.8
type	omnidirectional
Smin (thrsh_RSS)	-85 dBm
Qmin (thrsh_SDM)	0.3
S_{mid}	-76 dBm
S_{steep}	0.5
coords	(500 m, 10000 m)

Table 4 Antenna configuration parameters. Marginal distributions of network configuration parameters included in `antenna.xml`.

Parameter	min	q1	q2	mean	q3	max
Power (W)	5.000	10.000	10.000	9.574	10.000	10.000
Path Loss	3.800	3.900	3.900	3.939	4.000	4.000
Radius CoverArea (m)	1121.353	1333.521	1530.999	1483.766	1603.719	1947.483
S_{steep}	0.500	0.900	0.900	0.959	0.900	3.000
S_{mid} (dBm)	-94.000	-80.000	-80.000	-80.871	-79.000	-76.000