Towards a methodological framework for estimating present population density from mobile network operator data

Fabio Ricciato*[1], Giampaolo Lanzieri[2], Albrecht Wirthmann[1], and Gerdy Seynaeve[3]

[1]Eurostat Unit B1 - Methodology; Innovation in Official Statistics.
[2]Eurostat Unit F2 - Population and migration
[3]Proximus EBU Innovation

Abstract

The concept of *present population* is gaining increasing attention in official statistics. One possible approach to measure present population exploits data collected by Mobile Network Operators (MNO), from simple Call Detail Records (CDR) to more informative and complex signalling records. Such data, collected primarily for network operation processes, can be repurposed to infer patterns of human mobility. Two decades of research literature have produced several case studies, mostly focused on to CDR data, and a variety of ad-hoc methodologies tailored to specific datasets. Moving beyond the stage of explorative research, the regular production of official statistics across different MNO requires a more systematic approach to methodological development. Towards this aim, Eurostat and other members of the European Statistical System are working towards the definition of a general Reference Methodological Framework for processing MNO data for official statistics. In this contribution we report on the methodological aspects related to the estimation of present population density, for which we present a general and modular methodological structure that generalises previous proposals found in the academic literature. Along the way, we define a number of specific research problems requiring further attention by the research community. We stress the importance of comparing different methodological options at various points in the data workflow, e.g. in the geolocation of individual observations and in the inference method. Finally, we present illustrative preliminary results from a case-study based on real signalling data from a European operational network and highlight some lessons learned.

*Corresponding author. Email: fabio.ricciato@ec.europa.eu.

1 Measuring population size

Enumerating the population is one of the oldest statistical activities, but also one with many challenges, starting from the very definition of "population". The current international recommendations on population statistics favour the adoption of the concept of "usually resident population", based on a 12-month period of actual stay in the geographic area of interest. Therefore, observations must last for at least one year before assessing the inclusion or exclusion of an individual into the population of interest, regardless of which method or technology is adopted to perform the observations. However, this is not the only population concept in official statistics (e.g., see [1, 2]) and other alternative population concepts exist that are prone to be measured in more timely ways, with shorter observation periods. The most prominent alternative concept is the "present population", also known as *de facto* population. According to this concept, the target population is composed by all individuals who are physically present in the geographic area of interest at a selected reference time. The recent discussions on the population concept to be adopted in future official statistics may lead to increased relevance of the *de facto* population as a complementary source of information on population dynamics (see [3, 4]). Indeed, the evolution of concepts and data sources used for measuring the population might eventually lead to break the relation between physical presence on a territory and inclusion in its population count. In any case, the observation of physical presence of individuals on the territory could provide valuable inputs to the estimation of population based on different concepts.

The above considerations motivate the attention of statisticians for methods that allow measuring – or at least estimating – the size of the present population at a given reference time, over large territories (e.g., a whole country) and in a timely manner. To this aim, mobile network operator (MNO) data represent a promising data source, as evidenced by several research studies and academic literature during the last two decades. Moving from proof-of-concept case study towards an official statistics production setting requires addressing a number of issues, including privacy protection and sustainability of data provision. It also requires a more systematic methodological approach, calling for the development of a proper reference methodological framework that is the focus of the present contribution.

2 Why a Reference Methodological Framework?

A reference methodological *framework* is an abstract organisation of the data flow that is logically antecedent to the development of particular *methods*. The framework defines the structure of the data transformation flow, from raw input data through several stages of intermediate data until the final output, and defines *what* function is placed at each stage. Once the general framework is defined, different particular methods can be developed by instantiating each stage with a particular solution defining in detail *how* that function is implemented.

In other words, we may consider the methodological framework as a subspace of the whole methodological space, and particular methods as points within that subspace. In principle, every methodological design choice introduces an additional dimension. The analysis of complex data calls for involved methodologies embedding many (implicit or explicit) design choices, leading to a highly dimensional methodological space. Two similar methods that differ only in one or a few methodological design choices may be thought as neighbouring points in the methodological space. In an abstract sense, establishing a methodological framework gives some structure to the methodological space.

A specific method can be directly developed without the previous definition of a reference framework This approach suits well for case studies proof-of-concept and academic work focusing on a specific input data set and for a well understood use-case (desired output). However, when designing a statistical production process we must cope with a number of methodological design challenges. Each design choice represent a new dimension in the methodological space, and along each dimensions we may need to consider different solutions. In other words, in order to achieve a sound methodology we may have to explore a local neighbourhood in the methodological subspace. The role of a reference methodological framework, before and above the identification of particular solution points (i.e., specific methods), is essential to this purpose: it helps to reason about the problem, to implement different options and to communicate them to other users and co-developers.

The design of a reference methodological framework should take into account the following characteristics of the MNO data in input:

- **Heterogeneity**. Data are highly heterogeneous between different MNO in many respects. First, while most previous studies have considered Call Detail Records (CDR), an increasing number of MNO is now able to acquire signalling data that are more informative but also more complex than CDR. Furthermore, the data format and the detailed data generation process (hence their information content) are dependent on the particular configuration and operational conditions of the network infrastructure, all aspects that vary across MNOs;

- **Multi-purpose**. MNO data can be used in principle to extract information serving multiple statistical applications and use-cases. Therefore, a modular approach is needed to organise the analytic flow into processing modules that can be modified and adapted to different statistical purposes.

In order to address these challenges, a general Reference Methodological Framework (RMF for short) is under development in the European Statistical System (ESS). In this contribution we outline the general design principles of the RMF as initially proposed by Eurostat. While the RMF development work is still ongoing within ESS team and the final specifications are not yet ready[1], we believe that the fundamental concepts and core ideas presented in this work

[1]For updates, documents and deliverables refer to `https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPI_Milestones_and_deliverables`

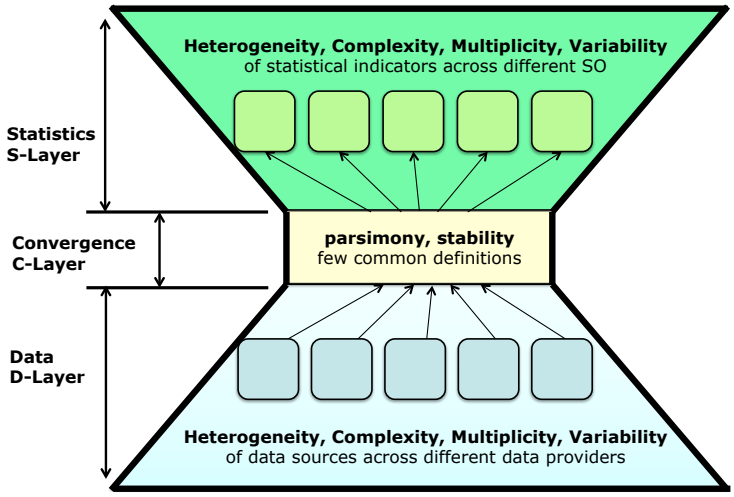


Figure 1: The layered hourglass model at the foundation of the Reference Methodological Framework under development by Eurostat

will be useful for the wider research community dealing with the analysis of MNO data. Along the way, we also formulate a number of research problems that deserve further attention by the research community.

The methodological aspect is only one of many challenges to be solved in order to move towards regular production of experimental statistics based on MNO data. Other non-methodological issues need to be solved, including legal and business aspects related to data access, data privacy and security aspects etc. Developing a RMF might be helpful to reason about these other non-methodological aspects, but the main goal of RMF is strictly focused on the methodological aspects. For this reason, issues of data access and privacy, though very compelling in practice, are left out of the scope of this contribution that is intended to be strictly focused on the methodological aspects. For a broader discussion of related to the use of new data sources for official statistics the interested reader is referred to [5] and [6].

The RMF design follows the principles of functional layering and the so-called *hourglass model* that lie at the foundation of the Internet (see [7, 8, 9]). The processing flow, from raw input data up to the desired statistical indicators (output data), is organised into three layers as sketched in Fig. 1. At the bottom, the *Data Layer* (D-layer) embeds processing modules whose implementation logic is highly specific to the particular MNO infrastructure, to the type and format of the available input data, or anyway dependent on technology-specific details. For each module in this layer, the implementation logic must be developed in close cooperation with MNO engineers and telecommunications experts in order to maximise the information content and mitigate some sources of error. At the top, the *Statistics Layer* (S-layer) includes modules that are de-

pendent on the particular statistical application and desired indicators. Between the input-specific D-layer and the output-specific S-layer, the intermediate *Convergence Layer* (C-layer) is designed to be input-agnostic and output-agnostic. In this way, the C-layer decouples the complexity and heterogeneity of the other domains, enabling independent development, hence evolvability and portability of the processing methods. The hourglass model is reminiscent of the Internet architecture, where the Internet Protocol decouples the (lower) network access layers from the (upper) application layers.

For a particular target application — for example, the problem of spatial density estimation that is the focus of this work — establishing a unified modular framework brings several further advantages. It helps the *definition and refinement* of novel estimation methods or variants thereof. It facilitates the *comparison* between different methods. By establishing a common terminology, formalism and conceptual frame, it facilitates the discussion and organisation of work among different researchers and development teams.

We remark that the RMF depicted in Fig. 1 is relevant only during the phase of *methodological development*, i.e., for building the processing methods and their software implementation. The RMF is not intended to prescribe *where* each function is to be physically executed during the computation phase. In one possible deployment scenario, the lower part of the processing workflow – including the whole D-layer, C-layer and the bottom part of the S-layer – could be physically executed at the MNO premises. The intermediate data that are produced at some logical point within the S-layer are then passed to the statistical office where the upper part of the S-layer functions are executed. From such (purely illustrative) example it should be clear that the layer interfaces do not necessarily map to borders between organisations. In other words, the proposed RMF helps to determine, at the methodological development stage, *what* logical function takes place in each module and *how* such function is implemented. It is not meant to define *where* and *by whom* the function is physically executed at the production stage: these are issues that attain more closely the legal, organisational and technical issues around data access (including but not limited to the privacy aspects) that fall beyond the scope of this contribution.

The general RMF concept and design principles are elaborated elsewhere (see [10, 11]). In this contribution we focus on a particular use case, namely the estimation of present population density. A modular view of the main processing stages is sketched in Fig. 2. Each module is briefly described in the next section.

Before proceeding further, it is useful to remind that processing MNO data for inferring human presence and mobility involves coping with three main dimensions of uncertainty:

- **Spatial uncertainty**: a generic MNO observation (signalling event) does not report to a point position, but rather relates to an extended location representing the expected (or assumed) radio coverage area of a particular radio cell[2] or slice thereof at that time;

[2]The "radio cell" represents a fundamental building block of mobile networks – that in fact are also called *cellular networks*. Every radio cell is uniquely identified by the Cell Global

- **Temporal uncertainty**: the location of a generic mobile user can not be observed continuously, but only at discrete event-generation times;
- **Population coverage uncertainty**[3]: the target population units, i.e., humans, do not correspond exactly to the observed units, i.e. mobile devices (see [12, Fig. 1.7]). Consequently, under-coverage, over-coverage and double-counting errors apply, corresponding respectively to people carrying no mobile device, devices not carried by humans (often called Internet-of-Things (IoT) or Machine-to-Machine (M2M) devices), and people carrying multiple devices or subscriptions in dual-sim phones. From the perspective of a single MNO, under-coverage errors are further magnified due to the limited market share of its subscriber basis.

3 A modular methodology for density estimation

The methodological framework sketched in Fig. 2 is designed to take in input multiple sources of MNO data: event-based records (CDR or, preferably, signalling records), network topology data (position and type of radio cells, radio coverage maps, transmit antenna configuration parameters[4]) and possibly other auxiliary information about subscriber groups, terminal types, etc.

3.1 Lower functions

The first processing stage (filtering and transformation box in Fig. 2) aims at excluding non-personal mobile devices like Internet-of-Things (IoT) and Machine-to-Machine (M2M) in order to reduce the over-coverage error. The identification of such devices is achieved through a set of filtering rules based on various network-level identifiers[5]. The detailed logic and their practical implementation are highly network-specific. Likewise all other modules at the D-layer, the RMF specifications should provide a set of guidelines, leaving to MNO engineers to particularise and adapt the implementation to the particular network configuration and available data.

Identity (CGI). Every radio cell is associated to a transmit antenna, but the association is not 1:1 as a single antenna can be used to transmit multiple radio cells (multiplexing).

[3]Throughout the paper we use the term "coverage" in two distinct and independent ways, in the spatial domain (radio coverage, coverage area) and in terms of statistical units (population coverage, coverage errors). The meaning of each occurrence should be clear from the context.

[4]Note that antenna configuration can be static or dynamic (adaptive). In the latter case, the power and direction of transmission (and consequently the size and location of the coverage area) can be varied in time in order to adapt to the current traffic condition. Antenna steering can be implemented mechanically or electronically (adaptive arrays).

[5]For example IMSI, IMEI, TAC, APN, as defined in the 3GPP standards (for more details see e.g. `https://www.etsi.org/deliver/etsi_ts/123000_123099/123003/10.05.00_60/ts_123003v100500p.pdf`).

3.2 Geo-location module

The second stage (event geo-location in Fig. 2) is central for the reduction of spatial uncertainty. Every event record contains information about the radio cell originating the message (call, SMS or signalling exchange). The coverage area of individual radio cells can be determined, at least approximately, based on auxiliary information about radio cell configuration and topology that are normally available in some form to MNO engineers (antenna position, orientation, beamwidth, transmit power, radio cell type, etc.). Therefore, every individual signalling event can be referred to a specific region, called *event location* hereafter, that basically corresponds to the nominal cell coverage area or part thereof[6]. Several different strategies can be chosen to calculate (or predict) the cell coverage area, depending on which approach is taken for modelling the physical process of device-to-radio cell association, leading to different geo-location options within the proposed framework. The systematic comparison of alternative geo-location methods is still an open research task. At one extreme, the simplest approach is to assume that the mobile device always connects to the closest antenna, leading to the well known approach based on Voronoi tessellation seeded by tower location [14, 15]. Despite its popularity in academic literature, such modelling approach is overly simplistic: it does not take into account some very basic aspects of mobile network operation, *in primis* the fact that radio cells with very different transmit power and coverage ranges are superimposed in so-called *multi-layer radio deployments*, with small cells and large cells (operating in different frequency bands and with different radio technologies 2G/3G/4G) co-existing together in the same area. Not always the mobile terminal selects the strongest received signal, and anyway the strongest received signal does not always correspond to the closest antenna (e.g. due to different transmit power) as implicitly assumed by adopting the Voronoi approach. Missing such fundamental phenomenological aspect represents a gross simplification and a possible source of modelling error.

Other geo-location variants mitigate this problem by taking into account, implicitly or explicitly, the multi-layer nature of radio network deployments, and heterogeneity of radio cell size. Examples include e.g. the approach proposed in [16] and those based on so-called Best Server Area (BSA) maps[7] as explored in

[6] The definition of "event location" depends on which variables can be observed by the MNO data at hand. If only the radio cell identifier is observed, then the event location corresponds to the radio cell coverage area. However, if additional variables can be observed, then the event location can be further narrowed down. For instance, in case of signalling messages extracted from the Radio Access Network (RAN), one may extract the so-called Timing Advance (TA) that provides a direct indication about the distance between the mobile device and the radio antenna. In other cases, proprietary systems are deployed in the network to extract accurate point positions of mobile devices obtained through multilateration methods (e.g. the LocHNESSs platform in [13] or commercial solutions for Location-Based Services (LBS)). However, such precise data are in general available only for selected groups of opt-in users and/or in limited geographical areas.

[7] BSA maps associate every point in space to the single radio cell with the strongest received signal strength. The latter is typically predicted by ad-hoc software tools based on radio propagation models, terrain data, radio cell configuration parameters, etc. In some cases,

[17]. Still, also such improved approaches are conceived to build event locations that are mutually disjoint, leading to alternative forms of *tessellations* of the geographical space.

Only a few pioneering work have started to consider models of overlapping locations that embrace the multi-layer nature of radio cell deployment, and the fact that radio cell coverage areas overlap by design [18, 19, 20, 21]. The choice between overlapping locations vs non-overlapping event locations (tessellations) has important consequences for the choice of the inference method, as discussed below.

For a generic mobile device, the output of the geo-location stage is basically a sequence of event observations referred to discrete points in time (event timestamps). Each observation is associated to a more or less extended area called *event location* hereafter. The task of determining the event location is performed by the geo-location module.

It is important to consider that within the interval between two consecutive observation times the mobile network does not know where the mobile device was located, whether it moved or where it moved. However, if signalling data are available from the D-layer, we can assume that during such interval the mobile devices remained confined within a certain area, called *bounding area* in our RMF (see Fig. 3) consisting of a predefined set of neighbouring radio cells[8]. The collection of event timestamps, event locations and possibly bounding areas for the same mobile devices constitutes the so-called C-path (for C-layer path) in the proposed RMF.

To maximise portability and inter-operability, a common C-path format should be adopted to represent data from different MNO. The definition of a standard format would allow algorithm producers (including statisticians and researchers) to develop software implementation that can run on data from different MNO.

In order to mitigate double-counting errors, one possibility is to resort to probabilistic matching: heuristic algorithms could be developed to identify pairs of strongly similar C-paths along a sufficiently long observation interval, that are likely to be associated to the same individual[9] (if present, such a module would be placed at the point marked with an asterisk in Fig. 2). These approaches can mitigate, but not completely eliminate population coverage errors. The

field measurements are used to improve the prediction of BSA predictions. BSA maps are typically used for radio planning and radio optimisation. An extension of the BSA concept is given by the N-Best Servers Area (N-BSA) maps, where each point is associated to $N \geq 1$ strongest radio cells. BSA can be seen as a special case of N-BSA for $N = 1$. BSA maps represent tessellations (non-overlapping locations), while N-BSA maps with $N \geq 2$ lead to overlapping locations.

[8]The concept of Bounding Area corresponds to the notions of Location Area, Routing Area, Tracking Area List and Registration Area, respectively, in 2G, 3G, 4G and 5G technology.

[9]This kind of processing would be extremely sensitive from a privacy point of view. In principle, Privacy Enhancing Technologies (PET) [22] may be adopted for inter-MNO data processing in general, and the probabilistic identification of paired subscriptions across different MNO based on spatio-temporal similarities represent an intriguing application of for PET, open for further research. Needless to say, establishing the *technical feasibility* of this approach is a separate task from establishing its *legal feasibility.*

statistical model at the upper S-layer should take coverage errors into account, and possibly quantify them in order to adjust the estimates, e.g. by resorting to external reference data from administrative records (as done e.g. in [14, 15]) or ad-hoc surveys.

3.3 Space-time interpolation

The next processing task is to determine the device location at the reference time t^* from the available observations at neighbouring observation times. This function represents a sort of *interpolation* in the joint space-time domain. Again, different strategies can be considered for this module. At one extreme, we might just pick the event location closest in time to t^* (zero-order interpolation) to serve as reference location, and this simplistic approach would be probably sufficient for initial implementations. More sophisticated interpolation methods might take into account external information about e.g. road network maps, urban layout, transportation network schedules, and of course the bounding areas introduced above, if available. If sufficiently frequent event data are available, interpolation might be combined with transport mode inference (as done e.g. in [18] for highway drivers).

The elaboration of more advanced interpolation strategies represents an interesting research problem *per se*. A possible research approach points in the direction of building a stochastic agent for each mobile user, considering a certain number of (quantitative or categorical) state variables such as e.g., position, velocity, transport mode etc. The stochastic nature of the agent means that state variables are expressed in the form of probabilistic distributions, and observations from MNO events — possibly coupled with auxiliary information (e.g., road maps, public transport schedules, etc. — is used to update the distribution shapes and impose constraints on their admissible ranges.

3.4 Density inference

From the observed (or interpolated) event locations for the selected set of mobile devices, the next task is to compute an estimation of the (unknown) spatial density. We propose to refer the estimates to a regular grid of small units, or tiles[10], e.g. the INSPIRE grid at Level 11 with tile size 100 m $\times$ 100 m [23]. From fine-grained estimates at the tile level, density estimates at any desired (coarser) level of administrative units can be computed straightforwardly by simple aggregation. The resulting workflow is exemplified in Fig. 4. By splitting the estimation workflow into two stages – namely *(i)* from event locations to fixed tiles in the estimation grid, and *(ii)* from tiles to the final desired level of statistical/administrative units — we decouple the most complex estimation task *(i)* from the particular use case and reporting administrative level, enabling the reuse of per-tile estimates across different application domains.

[10] We use the term "tile" to refer to the generic grid unit in order to avoid confusion with the term "cell" that we reserved for radio cells.

The estimation task *(i)* is mapped to the module *density estimation* in Fig. 2. External maps (e.g., buildings, land use, roads) can be used at this stage to increase spatial accuracy as depicted in Fig. 4. For instance, they can be used as priors in Bayesian inference methods, or as explicit constraints in other inference approaches.

4 A formalised view of density estimation

Hereafter we describe a compact model for the *data generating process* for the problem at hand. Let the jth element u_j of the column vector $\boldsymbol{u}$ denote the unknown number of mobile devices in grid tile $j = 1 \dots J$ (tile count). Let the ith element c_i of the column vector $\boldsymbol{c}$ denote the observed number of mobile devices associated to event location $i = 1, \dots I$ (location count). Denote by p_{ij} the probability that a mobile device located in grid tile j will be mapped to (observed in) event location i, as sketched in Fig. 5. In other words, p_{ij} represents the following conditional probability:

$$p_{ij} \stackrel{\text{def}}{=} \text{Prob}\left\{\text{user seen in event location } i \mid \text{user placed in tile } j\right\}. \tag{1}$$

For the sake of a more compact notation we gather the individual probabilities p_{ij}'s into matrix $\boldsymbol{P}_{[I \times J]}$. Note that matrix $\boldsymbol{P}$ is column stochastic by design, i.e., its elements sum to one along columns (formally $\boldsymbol{P}^T \mathbf{1}_I = \mathbf{1}_J$).

The (measured) location count $\boldsymbol{c}$ can be interpreted as the single realisation of a random vector $\tilde{\boldsymbol{c}}$ whose expected average value is given by:

$$E\left(\tilde{\boldsymbol{c}}\right) = \boldsymbol{P}\,\boldsymbol{u} \tag{2}$$

In the estimation problem we must solve for estimand $\boldsymbol{u}$ given the vector of measurement data $\boldsymbol{c}$ (representing the single available observation) and the model matrix $\boldsymbol{P}$ (inversion problem). The estimate $\hat{\boldsymbol{u}}$ can be written in general as:

$$\hat{\boldsymbol{u}} = g\left(\boldsymbol{P}, \boldsymbol{c}\right) \tag{3}$$

where $g(\cdot)$ denotes the estimator of choice. It is important to remark that in cases of practical interest the number of tiles is (much) larger than the number of event locations, i.e. $J >> I$, and therefore the associated inversion problem is underdetermined, pointing at issues of structural non-identifiability (see [24] and references therein). Any additional external information that is available to help the estimation process (e.g., prior distributions or spatial constraints derived from maps) can be embedded in the estimator $g(\cdot)$. Furthermore, it is evident that we must constrain the estimand variables to be non-negative, i.e. $u_i \geq 0$, $\forall i$.

Equation (3) shows that establishing an estimation procedure entails two distinct design choices that map to logically sequential sub-problems in the overall framework:

Geo-location approach	**Example**	**Notes**
Non-overlapping event locations (tessellation)	$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$	Fig. 5(a). Applicable to all Voronoi variants [14, 16] and BSA [17].
Overlapping event locations with equal probabilities	$\boldsymbol{P} = \begin{bmatrix} 0 & 0 & 0 & 1/3 & 1/2 \\ 1 & 1/2 & 1/2 & 1/3 & 0 \\ 0 & 1/2 & 1/2 & 1/3 & 1/2 \end{bmatrix}$	Fig. 5(b). Used in [19, 12].
Overlapping event locations with unequal probabilities	$\boldsymbol{P} = \begin{bmatrix} 0 & 0 & 0 & 0.25 & 0.62 \\ 1 & 0.65 & 0.43 & 0.6 & 0 \\ 0 & 0.35 & 0.57 & 0.15 & 0.38 \end{bmatrix}$	Fig. 5(b). Used in [20].

Table 1: Examples of model matrix $\boldsymbol{P}$ (before consolidation) corresponding to different classes of geolocation methods.

- Construction of model matrix $\boldsymbol{P}$ in the **geo-location** block. This task includes the determination of event location maps in the **geolocation** block and their probabilistic binding to tiles;
- Choice of a particular estimator $g(\cdot)$ in the **density inference** block.

Different methods can be adopted in each block. However, all possible methods can be grouped into three classes corresponding to the three rows of Table 1.

In case of non-overlapping locations (tessellation) every tile is assigned to only one single location, as exemplified in Fig. 5(a), therefore the elements of $\boldsymbol{P}$ are binary (refer to the top row of Table 1). In case of overlapping locations (refer to Fig. 5(b)) the elements of $\boldsymbol{P}$ can take any value in the interval $[0, 1]$ depending on the geo-location model of choice. For instance, the method presented in [19, 12] assumes that all radio cells covering a generic tile have the same probability of being selected by a mobile device placed in that tile, resulting in fractional values for the non-zero elements of $\boldsymbol{P}$ that are equal along each column. This approach leads to fractional elements of matrix $\boldsymbol{P}$ that are equal along columns, as shown in the middle row of Table 1. Finally, in the approach proposed by `mobloc` [20, 21] (and adopted also in [25]) the elements of $\boldsymbol{P}$ are tied to the relative signal strength of the received radio signal from the serving cell (relative to other cell signal strengths) resulting in unequal (and in general non-fractional) values for the competing cells in each tile, as exemplified in the bottom row of Table 1.

If two generic tiles j_1 and j_2 have equal assignment probabilities for all cells, i.e. $p_{ij_1} = p_{ij_2}\ \forall i$ — and same values for priors, in case that prior information is used in the estimate — then they are indistinguishable from each other. In other

words, they are perfectly *collinear* and we cannot identify differences between their respective estimates. In this case it makes sense to merge both tiles into a single super-tile, and then compute a single estimate for the whole super-tile[11]. Analytically, this corresponds to merging together identical columns of matrix $\boldsymbol{P}$. We refer to this operation by the term *consolidation*. Note that the consolidation process does not necessarily imply that the resulting (consolidated) matrix is full rank. In other words, it does not guarantee the resulting problem is fully identifiable.

If $\boldsymbol{P}$ has binary elements, the data generating process becomes deterministic and the resulting matrix after consolidation can be conducted to the identity matrix. The estimation problem then becomes trivial. For this reason, the estimation problem defined in the present contribution was not even mentioned in previous work dealing with MNO data that, aside of a few exceptions, ended up considering implicitly or explicitly some form of tessellation (typically based on Voronoi) for the geolocation stage. In fact, any form of *tessellation* divides the space into non-overlapping event locations that necessarily lead to binary elements of $\boldsymbol{P}$.

In case of overlapping locations (Fig. 5(b)) the estimation problem is instead non-trivial. Exploring possible estimation approaches for this case remains an open research problem. Hereafter we briefly list three different solutions that have been proposed in the recent literature. All three methods are fully coherent with the overall methodological workflow described above. To the best of our knowledge they are the only solutions proposed for this problem. More research is needed to better understand the theoretical properties of these estimators and probably develop new ones. Also, all these procedures produce simple point estimates, while it would be desirable to develop estimation procedures that deliver also some measure of uncertainty. Furthermore, we remark that problem formulation outlined in this section targets the density estimation of mobile devices, not people. Recall that mobile devices do not map 1:1 to humans, due to the population coverage errors. Further work is needed to extend the density estimation model(s) to take into account population coverage errors, in addition to spatial uncertainty.

4.1 MLE based on Multinomial distribution

One possible estimation method was elaborated in [19, Section 5.3], where a Maximum Likelihood Estimator (MLE) was developed based on a hierarchical generative model where the (unknown) vector $\boldsymbol{u}$ is modelled as a multivariate random vector with Multinomial distribution. Given the data $\boldsymbol{c}$ and model $\boldsymbol{P}$, the MLE is found by solving the following constrained optimisation problem:

$$\hat{\boldsymbol{u}} = \underset{\substack{\|\boldsymbol{u}\|_1 = c_{tot} \\ \boldsymbol{u} \geq \boldsymbol{0}}}{\arg\max} \sum_{i=1}^{I} c_i \log \sum_{j=1}^{J} u_j p_{ij} = \underset{\substack{\|\overline{\boldsymbol{u}}\|_1 = c_{tot} \\ \boldsymbol{u} \geq \boldsymbol{0}}}{\arg\max}\ \boldsymbol{c}^T \log \boldsymbol{P}\boldsymbol{u} \tag{4}$$

[11] The term *section* was adopted in [19] instead of *supertile*.

wherein $c_{tot} \stackrel{\text{def}}{=} \|\boldsymbol{c}\|_1 = \sum_i c_i$.

The practical implementation of this estimator involves a non-linear numerical optimisation that might be cumbersome to solve for very large problem instances, as typical in our application. Besides the practical resolution aspects, during our work we recognised a more fundamental theoretical issue in that the minimisation problem does not appear to have a unique global solution in the typical case that the number of supertiles (variables) is larger than the number of event locations (observations). This points to a fundamental problem of structural non-identifiability that was not noticed in the original paper [19]. Further research is compellingly needed in order to investigate this aspect.

4.2 MLE based on Poisson distribution

Recently, the authors of [25] have considered to apply to this problem the MLE estimator that was developed earlier by Shepp and Vardi [26] in the field of emission tomography. Like the previous approach, also this method is based on a hierarchical generative model, but here the elements of $\boldsymbol{u}$ are modelled as Poisson (instead of Multinomial) random variables. The MLE is then computed iteratively via an Expectation Maximization (EM) procedure.

At the iteration τ the new estimate $\hat{u}_j^{\tau+1}$ is computed iteratively from the previous estimate $\hat{u}_j^{\tau}$ according to the following formula (see [25, eq. (2)] and [26, eq. (2.13)]):

$$\hat{u}_j^{\tau+1} = \frac{\hat{u}_j^{\tau}}{\sum_{m=1}^{I} p_{mj}} \cdot \sum_{i=1}^{I} \frac{p_{ij} c_i}{\sum_{k=1}^{J} p_{ik} \hat{u}_k^{\tau}} \tag{5}$$

The iterative resolution approach given by (5) enables the application of this method to very large problem instances. However, during our work we recognised that the final solution depends on the initialisation point. In other words, similarly to the previous MLE presented in §4.1, also this estimation approach points to issues of structural non-identifiability (that went unnoticed in both previous works [25] and [26]) and are worth to be further investigated by the research community.

4.3 Direct computation based on Bayes rule

This simple method was proposed by the `mobloc` package developed by Tennekes et al. [20] (see also [21]). The location counters are distributed linearly to cells through a matrix $\boldsymbol{Q}_{[J\times I]}$ that is derived from the elements of $\boldsymbol{P}$ as described below, formally:

$$\hat{\boldsymbol{u}} = \boldsymbol{Q}\boldsymbol{c}. \tag{6}$$

The generic element q_{ji} of matrix $\boldsymbol{Q}$ is interpreted as a conditional probability:

$$q_{ji} \stackrel{\text{def}}{=} \text{Prob}\left\{\text{user placed in tile } j \mid \text{user seen in event location } i\right\}. \tag{7}$$

Note the swapping of the conditioning direction between q_{ji} and p_{ij} defined earlier in (1). Let $a_j \stackrel{\text{def}}{=} \text{Prob}\left\{\text{user placed in tile } j\right\}$ denote the prior distribution.

Recalling the fundamental Bayes rule

$$\text{Prob}\{j \mid i\} = \frac{\text{Prob}\{i \mid j\} \cdot \text{Prob}\{j\}}{\text{Prob}\{i\}}$$

it holds that $q_{ji} \propto p_{ij} \cdot a_j$. In case of non-informative prior, a_j is constant and the posterior term q_{ji} is proportional to the probability p_{ij} up to a renormalisation constant, formally $q_{ji} = \frac{p_{ij}}{\gamma_i}$. The normalisation constant γ_i must guarantee that $\sum_j q_{ji} = 1$ and is easily computed as $\gamma_i = \sum_j p_{ij}$. Therefore, for a non-informative (uniform) prior we can compute the posterior terms simply as

$$q_{ji} = \frac{p_{ij}}{\sum_j p_{ij}} \tag{8}$$

and then plug them into (6) to obtain the per-tile estimates.

This method is particularly appealing due to its simplicity, ease of implementation and computational scalability: through direct computation it provides immediately a single non-ambiguous solution for the given input model $\boldsymbol{P}$ and data $\boldsymbol{c}$.

5 A case study: description

In this section we report preliminary results obtained based on real data from an operational network. Our goal here is not to provide conclusive final results, but rather to illustrate how different choices at a single step in the methodological chain may lead to very different final estimates. We focus specifically on the geolocation module and present four different density maps obtained by applying four different geolocation solutions *to the same input dataset*. The numerical results presented hereafter should be taken both as a warning about the issue of methodological sensitivity, and stimulus to conduct further research on the comparison of different methodological options for each logical module.

5.1 Input data

The input data for this work were collected by the operational network of Proximus, the largest mobile network operator in Belgium, during a single working day in 2017 (exact date undisclosed). All individual signalling records were collected through all mobile technologies (2G, 3G and 4G) across the whole Belgium. The raw data were pre-processed in order to filter out non-personal devices (IoT, M2M devices). All personal information was removed and device identifiers were pseudonymised before storing the data on hard disk. Each individual record contains the user pseudonym, the timestamp and the identifier of the radio cell where the signalling event was generated. The information about bounding areas was not used for this study. Since the focus of this work was *not* on space-time interpolation (ref. Fig. 2) we adopted the most simple option for this module, i.e., zero-order backward interpolation: for a generic user i the

cell identifier of the latest observed event before the target reference time t^* was taken as proxy for the user location at the reference time. Four different reference times were considered for this study at different time-of-day, namely $t_k^* \in \{6\text{h}, 10\text{h}, 17\text{h}, 22\text{h}\}$. For each radio cell j, a vector of four counters was built representing the number of mobile users mapped to that cell at different reference times. Radio cells for which all counters were zero at all four reference times were discarded. The whole data processing was conducted locally at the MNO premises, i.e., individual signalling records never left the MNO domain and only aggregate (non-personal) data were passed to Eurostat. This approach of *moving computation towards the data* (instead of moving the data) is in line with the fundamental principles of *Trusted Smart Statistics* [6] and with the operational paradigm of other projects[12].

5.2 Geo-location methods

For this study we have considered four different geolocation methods: three different variants of tessellations (non-overlapping locations) and one variant of overlapping location (with uniform probabilities). In the rest of the paper the four method are labelled as follows: Naif Voronoi (NV), Bi-layer Voronoi (BV), Proximus Voronoi (PV) and Overlapping Locations (OL).

The differences between the different options will be explained with the support of Fig. 6 and Fig. 7. The most important point to consider is *cell size heterogeneity*. Real-world radio deployments consists of cells of very different sizes, ranging from small cells of a few tens of meters or even less (picocells, femtocells), through microcells of a few hundred meters, up to large macrocells of many kilometres. Generally speaking, overlapping cells of different sizes coexist in the same area and overlap to each other. To exemplify the discussion, we may divide all radio cells into only two categories: small cells and large cells. Small cells may be deployed to fill coverage gaps between large cells, or to serve known high traffic hotspots within the coverage area of larger cells. To illustrate, Fig. 6(a) depicts a synthetic toy scenario with ten radio cells: four large cells (for which antenna towers are labelled by capital letters A-D) and six small cells (labelled by small letters). The circles (in blue and red, respectively, for large and small cells) represent the cell coverage borders. The network geometry depicted in Fig. 6(a) may be translated into different tessellation patterns depending on how the tessellation is built: the three options considered in this study are shown in the remaining Figg. 6(b)-6(d) and explained hereafter.

Besides cell size heterogeneity, another aspect to consider is *cell directionality*. Radio cells, and especially large macro cells, are often configured with directional antennas, meaning that the cell tower location is eccentric to the radio coverage are. A particular popular configuration consists of multiple cells (also called *sectors* in this case) sharing a single antenna mast but pointing at different azimuthal directions. To illustrate, Fig. 7 depicts a popular configuration with three symmetric sectors of 120° beam width, but the following

[12] See e.g. the OPAL project `https://www.opalproject.org`.

discussion applies *mutatis mutandis* to other (possibly non-symmetric) configurations with different beam width values, e.g. 60° or 180°. For the case shown in Fig. 7, different choices are possible as to where the seed(s) of the Voronoi tessellation are placed. In the simplest (and by far most popular in academic literature) option, the seed of each cell is mapped to the cell tower location, as shown in Fig. 7(a). This choice implies that antenna directionality, if present, gets ignored because all three cells (sectors) sharing the same antenna tower will be mapped to a single Voronoi seed, and therefore to the same Voronoi polygon, regardless of their different azimuthal directions. The directional information can be preserved instead by the other two alternative methods shown in Figg. 7(b) and 7(c). In both cases, a different seed is associated to each cell (sector). With the method depicted in Fig. 7(b), which was originally proposed in [16], the seed location is determined by adding a small offset to the tower location in the direction of the cell beam: this simple trick forces the Voronoi tessellation to generate different polygons in the different azimuthal direction. For the sake of completeness we mention here also the method depicted in Fig. 7(c) and considered earlier in [19] and [27], wherein the seed is placed at the barycentre of the cell coverage area. Note that implementing this latter option requires (at least approximate) knowledge of the cell range in addition to the cell azimuth.

5.2.1 Naif Voronoi (NV)

In the Naif Voronoi (NV) option, the reference cell area is built geometrically by means of a simple Voronoi tessellation seeded by all cell tower locations. As explained above commenting Fig. 7(a), antenna directionality is ignored. Most importantly, cell size heterogeneity completely is ignored since small and large cells are treated equally. The resulting NV tessellation corresponding to the scenario of Fig. 6(a) is depicted in Fig. 6(b).

NV is by far the most popular method in past academic literature. It is simple to understand and to implement as it uses minimal information about cell deployment: only cell tower locations must be known. In other words, cell directionality (beam width and azimuth) and cell size (large vs. small) information, even if potentially available, is not used. This approach however suffers from serious limitation. The main problem of NV originates from the mixing of small and large cells, leading to a gross distortion of the network geometry for both groups of cells. The problem can be appreciated by carefully considering the tessellation pattern in Fig. 6(b) against the network geometry in Fig. 6(a): small cells tend to be associated to Voronoi polygons larger (in some cases much lager) than their actual coverage area, while conversely large cells get associated to polygons smaller (in some cases much smaller) than their coverage area. This introduces an artificial inflating and deflating, respectively, of small areas and large areas. This in turn translates into a serious distortion of the spatial density pattern. Recall that with tessellations the spatial density is computed simply by the ratio of the number of users seen in each polygon (given by the sum of the users observed across all cells mapped to that polygon) divided by the polygon area. That leads to artificially expanding [resp. compressing]

the reference geographical area of small cells [resp. large cells] causes artificial dilution [resp. concentration] of the local density. In our dataset we often encountered small cells with very low user counts in zones where mobile traffic is mostly served by large cells. Such uneven distribution of users across cells of different geographical size (small cells with few users, large cells with many users) coupled with the *systematic distortion* of area sizes (expansion of small cells, compression of large cells) produce an artificially high degree of density variations at fine spatial level, with a few very high-density zones scattered among many very low density zones. The resulting pattern appears to be a sort of random mix of "voids and fills" — a sort of noise at high spatial frequencies — that is essentially an artefact of poor geolocation methodology. This pattern will be evident from the empirical results shown later in this section.

5.2.2 Bilayer Voronoi (BV)

In order to avoid the artefact pattern described above, we consider an alternative scheme where large and small cells are treated differently. The two cell groups are interpreted as independent *layers* of radio coverage, motivating the attribute "Bi-layer" that we use to label this option. BV represents a simplified version of the more articulated method proposed by Meersman et al. [16]. Essential to this method is a preliminary classification of radio cells into large and small. In the first step, small cells are ignored and a Voronoi tessellation is built by considering exclusively the large cells. In so doing, antenna tower locations are taken as seeds (as in Fig. 7(b)) therefore the directionality of (large) cells is ignored. As a by-product, all macro cells sharing the same antenna tower are mapped to a single seed, hence to a single polygon. This includes cells of different radio access technologies (2G, 3G and 4G) that are often co-located on the same tower to save on site installation costs. This motivates the label "technology agnostic" used in [16] to refer to this method. The Voronoi polygons built in this way are logically associated to groups of co-located (large) cells. The resulting BV tessellation corresponding to the scenario of Fig. 6(a) is depicted in Fig. 6(c). In BV the polygon size depends exclusively on the local density of large cells, irrespective of the presence and position of small cells: this is the key difference against NV wherein the polygon size depends on the local density of *large and small* cells jointly. For each polygon, the user count is computed by accumulating the user counts of *(i)* all constituting large cells (sharing the same antenna tower) plus *(ii)* the small cells falling inside the polygon.

5.2.3 Proximus Voronoi (PV)

The Proximus Voronoi is similar to BV in that only large cells are used to drive the Voronoi tessellation. The only difference between BV and PV is that the latter takes into account directionality by setting seed locations according to the offset method exemplified in Fig. 7(b) (recall that BV instead uses the simple antenna location method of Fig. 7(a)). This method follows closely the method proposed by the Proximus engineers earlier in [16] (termed "technology

agnostic cell sectors" therein). Taking the reference deployment shown in Fig. 6(a) and assuming that all macro cells are divided into three 120° sectors in parallel azimuthal directions following the same pattern shown in Fig. 7(a)), the resulting PV tessellation would be that shown in Fig. 6(d).

5.2.4 Overlapping Locations (OL)

While the previous three geolocation methods produce (different variants of) tessellations, the fourth considered method assumes overlapping locations, as per the terminology introduced earlier in §3.

For each cell, the nominal coverage area was determined by the expert operator engineers taking into consideration antenna orientation, type of cell and (approximate) transmission range. The latter was determined based on cell traffic statistics that were available from radio equipment. For a generic tile j covered by n (overlapping) cells, the assignment probabilities (i.e., the non-zero elements of model matrix $\boldsymbol{P}$) were set equal to $p_{ij} = \frac{1}{n}$ (ref. to the middle row of Table 1). The procedure adopted to determine the nominal cell coverage area in this specific case study is admittedly heuristic and depends on several network-specific configuration details that may differ across MNO (e.g., availability of up-to-date information about radio cell configuration details, or lack thereof). Providing a detailed description of such heuristic procedure goes beyond the scope of this work and would be anyway unnecessary in the economy of the present contribution. Indeed, our goal here is not to affirm the superiority of the particular procedure adopted for this specific case study, but rather draw attention to the potential opportunity (and challenges) associated to the class of geolocation methods based on overlapping locations. In other words, the procedure adopted in our case study serves merely as a possible representative of this class of methods, not necessarily representing *the* best possible method in all operational conditions. We anticipate that more sophisticated methods to establish the cell coverage area will soon emerge as natural extensions of the N-BSA concept[13].

[13] As the market for MNO data analytics develops, and more value (also commercially) is put on high-precision analytical products, vendors of radio planning tools will probably start to incorporate additional features to their commercial tools, in the direction of exporting more accurate an complete radio coverage maps than merely single BSA, towards N-BSA for arbitrary values of N and augmented with (relative) signal strength information, so as to enable more accurate instantiation of the model matrix $\boldsymbol{P}$. Note in fact that most (if not all) the information needed to run such computation *is already incorporated within such tools* (physical propagation models, radio cell configuration parameters, terrain data etc.) and in some cases the full computation does already takes place, but the necessary data exporting features are often missing since such features were never required for pure radio planning tasks, for which such tools were originally developed. As a further step, the vendors may incorporate the computation of matrix $\boldsymbol{P}$ directly into their commercial tools, and through this path extend the scope of their product towards the raising market of mobile analytics.

5.3 Density inference

Recall that the main focus of the case study presented in this contribution is to compare different geolocation methods. In other words, we intend to show how different final results, i.e density maps, can be obtained from the same input dataset for different choices of the geolocation method. However, the geolocation method is only one module in the longer methodological chain depicted in Fig. 2, and to achieve final results we had to populate also the other modules. In principle, conducting a fair comparison requires that all other modules are instantiated with the very same method, regardless of which geolocation option is considered. This is always possible up to some unavoidable interdependencies that might exist between different modules. An important interdependency exist between the geolocation method (refer again to Fig. 2) and the subsequent density inference method. As discussed earlier in Section 4, if the geolocation method is of tessellation type (non-overlapping locations) the solution to the problem of density inference reduces, trivially, to the ratio between user counter and area of each polygons. Therefore, all three different variants of tessellation methods considered in our study (namely NV, BV and PV) come with an obvious (and trivial) solution for the estimation stage.

Things are different for the fourth method, namely OL. As a matter of fact, the identification of the "best" estimator for the problem at hand is still an open research problem (also because the problem was never explicitly stated in these terms in the past literature). In Section 3.4 we have listed the existing few proposals that have appeared recently in the literature. For this case study, we have decided to adopt the simplest of the three methods, namely the one described earlier in §4.3, based on the simple inversion of the Bayesian rule and first proposed by the Tennekes et al. in the `mobloc` R package [20] (see also [21]). Again, this choice is driven by the quest for simplicity rather than for optimality: its simple structure allows for direct computation (it does not involve numerical minimisation) making it particularly appealing for large scale problem instances as the one considered in this study.

6 A case study: results and discussion

Fig. 8 reports four spatial density maps obtained with the different geo-location methods described above, for the whole Belgium, at one sample reference time (6h) on the considered day. In all cases estimates were obtained for squared tiles of 100 m $\times$ 100 m adhering to the INSPIRE grid at Level 11 [23]. Fig. 9 reports the corresponding zoom-in maps for the region around Brussels.

From Fig. 8, and even more so from Fig. 9, it is immediately apparent how different quantitative results can be obtained by adopting different choices for the geolocation module. Recall that all maps were obtained from the very same input data (!). This picture remind us of the (often neglected) importance of considering methodological sensitivity in so-called "Big data" research. More often than not, academic papers tend to present their results based on

a single method embedding many different choices at various stages along the methodological chain, without considering alternative options. In other words, they consider a single point (or method) of a much wider methodological space, missing to explore alternative options in the neighbourhood. On the other hand, any methodological design choice adds a new dimension to the methodological space, resulting in a highly multidimensional methodological space that is obviously impossible to explore exhaustively. However, *the fact that conducting a complete exhaustive exploration is unfeasible cannot be taken as a legitimate justification for conducting no exploration at all.* Exploring at least *some* alternative methodological points, along directions that are deemed to be of higher priority, should be(come) common practice in all research dealing with big data and, consequently, with complex analytic methodologies. The definition of a general *methodological framework*, before and above the instantiation of a single *method*, is well instrumental to this purpose in that it gives structure (and a reference system) to the methodological space, at the same time helping to identify the most compelling directions for exploration and research subproblems.

A second take-home message from Figg. 9 and 8 is that details matter. The first three geolocation methods (NV, BV and PV) are all based on the same *principle* of Voronoi tessellation, but slight differences in the interpretation of this principle – in this specific case, how the Voronoi seeds are selected – may have important effects on the final estimates. This is particular evident when contrasting the NV map to the other two (BV and PV): the NV maps present many "voids" that are not present in the other maps. We account this pattern to a methodological artefact produced by mixing small and large cells, as explained earlier in §5.2.1. More in general, such differences remind us that there is no single "Voronoi approach" but rather many different ways of applying the basic Voronoi principle to MNO data. As the differences between the final results obtained with different methods may be substantial, the difference between should not be underscored as a matter of minor "implementation detail", but constitutes instead an characterising element of the overall methodology Unfortunately, in most previous literature (including the popular paper by Deville et al. [14]) there is no explicit mentioning as to whether small/ and large cells where handled differently, making it impossible to tell whether their results were obtained with a method more similar to NV or to BV (or with yet another different variant of the Voronoi).

More in general, besides the important quest for ensuring *reproducibility* in methodological research [28, 29], we believe that stronger attention should be put by the research community to the issue of *methodological sensitivity.*

Among all the four methods, the map obtained by OL might appear visually the most plausible compared to the others: the overlapping cell pattern renders a map with smoother gradient across neighbouring tiles, without the sharp transitions across the border of Voronoi polygons that are evident in all tessellation methods. However, we remark that *better visual plausibility does not imply better accuracy.* As a matter of fact, at this stage of the work we cannot conclude which one of the four maps represents more closely the actual distribution of mobile devices at that time, nor we can quantify the accuracy

gain (if any), simply because the "ground truth" distribution is unknown. Comparing the *present population* estimates obtained from MNO data (e.g., at night time) to the official figure of *resident population*based on administrative data (registered population) is surely an interesting exercise — and we do indeed we plan to conduct such comparison as part of our future work — but as matter of fact the latter cannot be taken as ground truth for the former, trivially because (even at night time) people do not all and always stay in the place where they are registered.

Lack of ground truth is a well-known problem in many diverse research fields dealing with measurement or estimation of complex phenomena, including e.g. topology and traffic in the Internet. In all those fields, resorting to synthetic data and simulations, is often the only way to gain insight into the strengths, weaknesses and trade-offs of different measurement/estimation methods. Methodological research on MNO data will be no exception, and the work within the ESS is already moving in this direction with the ongoing development of a modular open-source simulator for MNO data in support of methodological research [30].

Besides the issues related to availability of ground truth (or lack thereof), another important methodological research point relates to the choice of the error metric between the estimated map $\hat{\boldsymbol{u}}$ and the given reference map $\dot{\boldsymbol{u}}$, or more in general between any two given maps. The problem relates to measuring the distance $D\left(\hat{\boldsymbol{u}}, \dot{\boldsymbol{u}}\right)$ between distributions defined over an Euclidean space (bidimensional in our case). Previous work (including [14]) has resorted to standard metrics like Mean Square Error (MSE) computed over all individual element-by-element differences, i.e., $D\left(\hat{\boldsymbol{u}}, \dot{\boldsymbol{u}}\right) = \sum_j \left(\hat{u}_j - \dot{u}_j\right)^2$. Others have considered to apply to this problem the Kullback-Leibler (KL) divergence, the Hellinger distance, or some other kind of f-divergence [31], implicitly interpreting the (rasterised) spatial distribution as a (binned) probabilistic distribution. A problem that is common to all such approaches, and more in general to any f-divergence function, is that they miss completely the *Euclidean proximity* between the individual bins of the distribution (tiles in our case), that is a fundamental aspect of *spatial distributions* as opposite e.g. to categorical distributions. The problem is exemplified in the toy scenario depicted in Fig. 10, where we assume the (unknown) the ground truth distribution is the one labelled by "O". It is trivial to recognise that candidate maps "A", "B" and "C" yield exactly the same distance to "O" if MSE, KL or any other kind of f-divergence is taken to measure "distance" between distributions. This is because — somewhat paradoxically — all such measures of "distance" (in the probabilistic sense) do not take into account the Euclidean distance between the bins (tiles), but only their associated distribution values. In other words, adopting any such measure implies giving away the very fundamental aspect of any spatial distribution, that is its Euclidean support, implicitly interpreting it as a purely categorial distribution.

Going back to the toy example of Fig. 10, if any such measure is adopted to evaluate the goodness of the different maps, hence of their respective estimation methods, it would be impossible to determine the superiority of method "A"

over alternative methods "B" and "C". A similar statement applies to the triple of maps labelled as "E", "F" and "G" in the same figure. This toy example should serve as a warning, or at least a reminder, that a sub-optimal choice of the evaluation procedure between different measurement options carries the risk of leading towards the selection of sub-optimal methodologies.

Generally speaking, the ability to assess quantitatively the goodness of an estimation method is a key task in methodological development for official statistics, not least because it allows to evaluate whether the increased level of accuracy is worth the increased level of implementation and conceptual complexity. For the specific problem at hand, establishing a solid method to quantify spatial distribution error remains an open research problem. Part of our current work is focusing on this aspect. We believe that f-diverge measures should be avoided for this problem, and we consider the application of the so-called Earth Mover Distance (EMD) — also known as Wasserstein distance and with several other names, see [32, 33] and references therein — as a very promising direction for further investigation[14]. However, early attempts to apply EMD on our dataset revealed a number of practical issues, including but not limited to the difficulty of computing EMD between very large maps. Applying EMD computation on local zones (sub-map), or to more coarsely aggregated grids, comes with subtle but potentially serious implications that we are currently investigating. Likewise with Voronoi tessellation, there is no single "EMD approach" but multiple different ways of applying the EMD principle to our problem. Again like Voronoi, more research is needed to understand the implications of different options with the goal to identify the best solution — or at least avoid the most naif ones.

7 Outlook on future work

In this paper we have discussed the motivations and the fundamental ideas underlying the development of a general Reference Methodological Framework for processing MNO data for official statistics, with a focus on the task of estimating the spatial density of present population. While the work is still in progress within the ESS, we believe that the key concepts and problem framing outlined in this paper will be useful also for the wider research community engaged in MNO data processing. Along the way, we have indicated several research problems that deserve further attention by the research community. Among them, our ongoing research in Eurostat is focusing specifically on the problem of density inference in case of overlapping locations, for which we recognised issues of structural non-identifiability, and on new procedures to quantitatively assess the accuracy of spatial estimates in Euclidean spaces. In parallel to the methodological work, we are investigating possibilities to leverage Privacy Enhancing Technologies for the fusion of input data from multiple (and possibly competing) MNO.

[14]The difference between EMD and f-divergence measures is brightly explained in terms of "horizontal" versus "vertical" differences in `https://jeremykun.com/2018/03/05/earthmover-distance`.

Acknowledgments

We are grateful to Janine Eguienta for her contribution to this work during a summer internship at Eurostat. The clarity and readability of an earlier version of this work has improved thanks to interaction with and feedback from various ESS colleagues and most prominently: David Salgado, Martijn Tennekes, Benjam Sakarovitch, Milena Suarez-Castillo, Roberta Radini and Tiziana Tuoto. A discussion held by Fabio Ricciato with Angelo Coluccia was helpful to clarify certain theoretical aspects related to inference method for overlapping locations.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily represent the official position of the European Commission.

References

[1] G. Lanzieri. Population definitions at the 2010 censuses round in the countries of the unece region. In *15th Meeting of the UNECE Group of Experts on Population and Housing Censuses, Geneva, Switzerland*, 2013. `http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/10_E_x20_Aug_WEB.pdf`.

[2] G. Lanzieri. On a new population definition for statistical purposes. In *15th Meeting of the UNECE Group of Experts on Population and Housing Censuses, Geneva, Switzerland*, 2013. `http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/Eurostat_introductory_paper_on_new_population_definition.pdf`.

[3] G. Lanzieri. Alternative definitions of population for future demographic and migration statistics. In *13th Meeting of the Task Force on the future EU censuses of population and housing, Luxembourg*, 2018.

[4] G. Lanzieri. Towards a single population concept for international purposes: definitions and statistical architecture. In *16th Meeting of the Task Force on the future EU censuses of population and housing, Luxembourg*, 2019.

[5] E. Letouzé and J. Jütting. Big data and human development: Towards a new conceptual and operational approach. DATA-POP Alliance White Paper, `https://paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf`, March 2015.

[6] F. Ricciato, A. Wirthmann, K. Giannakouris, F. Reis, and M. Skaliotis. Trusted smart statistics: Motivations and principles. *Statistical Journal of the IAOS*, 35(4), 2019.

[7] M. Chiang, S. Low, A. Calderbank, and J. Doyle. Layering as optimization decomposition. *Proceedings of the IEEE*, 95, 2007.

[8] S. Akhshabi and C Dovrolis. The evolution of layered protocol stacks leads to an hourglass-shaped architecture. In *ACM SIGCOMM'11*, August 2011.

[9] J. Zittrain. Chapter 45: Internet. In *A History of Intellectual Property in 50 Objects*. PCambridge University Press, 2019. `https://dash.harvard.edu/handle/1/40838991`.

[10] F. Ricciato. Towards a reference methodological framework for processing mno data for official statistics. In *15th Global Forum on Tourism Statistics, Cusco, Peru*, November 2018. `https://tinyurl.com/ycgvx4m6`.

[11] F. Ricciato. Towards a reference methodological framework for the processing of mobile network operator data for official statistics. Keynote Talk at Mobile Tartu 2018. `https://tinyurl.com/y75psqs4`.

[12] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. Estimating population density distribution from network-based mobile phone data. JRC Technical Report, 2015. `https://tinyurl.com/ydz4mgaw`.

[13] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1), 2011.

[14] P. Deville et al. Dynamic population mapping using mobile phone data. *PNAS*, 111(45), 2014. `https://doi.org/10.1073/pnas.1408439111`.

[15] B. Sakarovitch, M.P. de Bellefon, P. Givord, and M. Vanhoof. Estimating the residential population from mobile phone data, an initial exploration. *ECONOMIE ET STATISTIQUE / ECONOMICS AND STATISTICS*, 505-506, 2018. `https://www.persee.fr/docAsPDF/estat_0336-1454_2018_num_505_1_10870.pdf`.

[16] F. De Meersman et al. Assessing the quality of mobile phone data as a source of statistics. In *European Conference on Quality in Official Statistics*, June 2016. `https://tinyurl.com/y7pbracn`.

[17] F. De Fausti, M. Savarese, F. Fabbri, M. Spada, R. Radini, T. Tuoto, and L. Valentino. Challenges and opportunities with mobile phone data in official statistics. In *Conference of European Statistics Stakeholders (CESS)*, October 2018.

[18] A. Janecek, D. Valerio, K. Hummel, F. Ricciato, and H. Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 16(5), October 2015.

[19] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. Beyond the "single-operator, cdr-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, May 2016.

[20] M. Tennekes. R package for mobile location algorithms and tools, April 2017. `https://github.com/MobilePhoneESSnetBigData/mobloc`.

[21] M. Tennekes. Statistical inference on mobile phone network data. Presentation at European Forum for Geography and Statistics (EFGS 2018) `https://www.efgs.info/wp-content/uploads/conferences/efgs/2018/presentations/TENNEKES_V1_EFGS2018-Statistical-interference-on-mobile-phone-network-data.pdf`, 2018.

[22] Big Data UN Global Working Group. Un handbook on privacy-preserving computation techniques. `https://tinyurl.com/y3rg5azm`, 2019.

[23] INSPIRE Thematic Working Group Coordinate Reference Systems and Geographical Grid Systems. D2.8.I.2 Data Specification on Geographical Grid Systems — Technical Guidelines (v 3.1), April 2017. `http://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GG_v3.1.pdf`.

[24] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer. Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Phil. Trans. R. Soc.*, 371, 2013. `http://jeti.uni-freiburg.de/papers/20110544.full.pdf`.

[25] J. van der Laan and E. de Jongey. Maximum likelihood reconstruction of population densities from mobile signalling data. In *NetMob'19*, 2019.

[26] L. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1982.

[27] Center for Spatial Information Science — Univ. of Tokyo. A study on urban mobility and dynamic population estimation by using aggregate mobile phone sources. `http://www.csis.u-tokyo.ac.jp/dp/115.pdf`.

[28] V. Stodden. The reproducible research movement in statistics. *Statistical Journal of the IAOS*, 30, 2014. `doi:10.3233/SJI-140818`.

[29] V. Stodden. Enhancing reproducibility for computational methods. *Science*, 354(6317), 2016. `doi:10.1126/science.aah6168`.

[30] B. Oancea, M. Necula, L. Sanguiao, D. Salgado, and S. Barragán. Deliverable i2: A simulator for network event data. `https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b9/WPI_Deliverable_I2_Data_Simulator_-_A_simulator_for_network_event_data.pdf`, December 2019.

[31] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *EEE Transactions on Information Theory*, 52(10), October 2010.

[32] C. Gottschlich and D. Schuhmacher. The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *Plos One*, 2014.

[33] E. Levina and P. Bickel. The earth mover's distance is the mallows distance: some insights from statistics. In *8th IEEE Int. Cong. on Computer Vision (ICCV 2001).*, 2011.

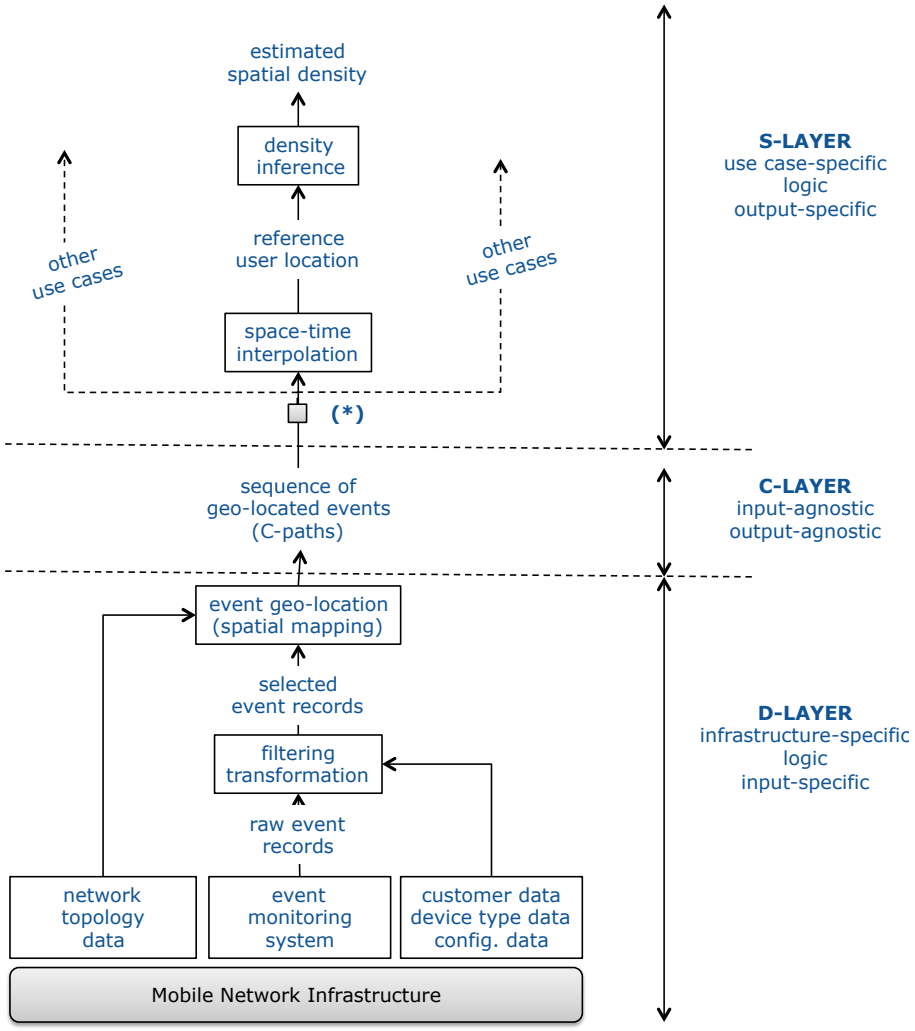


Figure 2: Modular structure of the density estimation procedure.

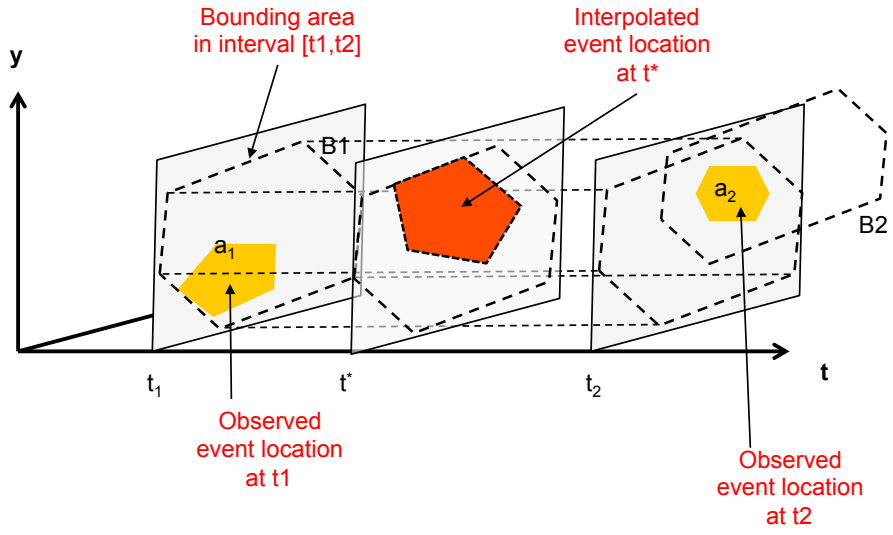


Figure 3: Graphical representation of the C-path components (event locations and bounding areas) and interpolated location.

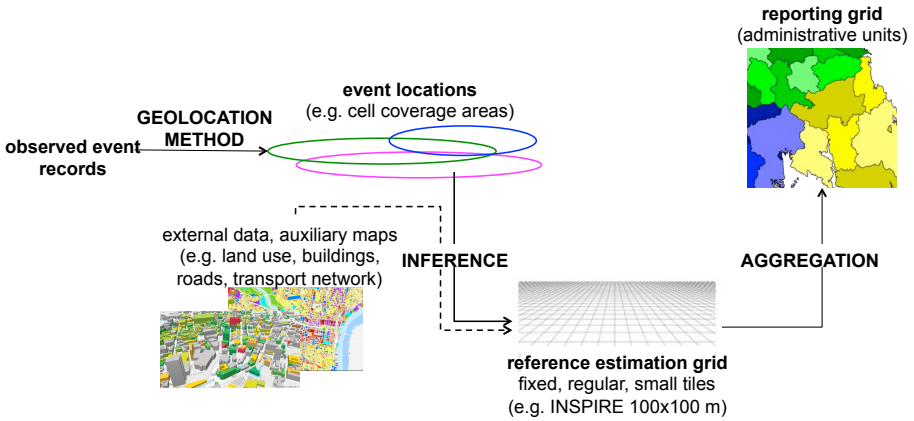
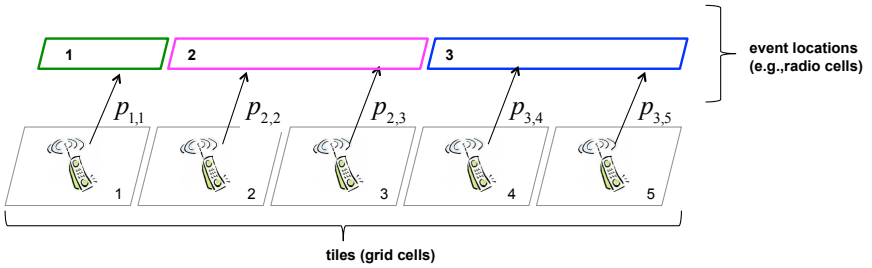
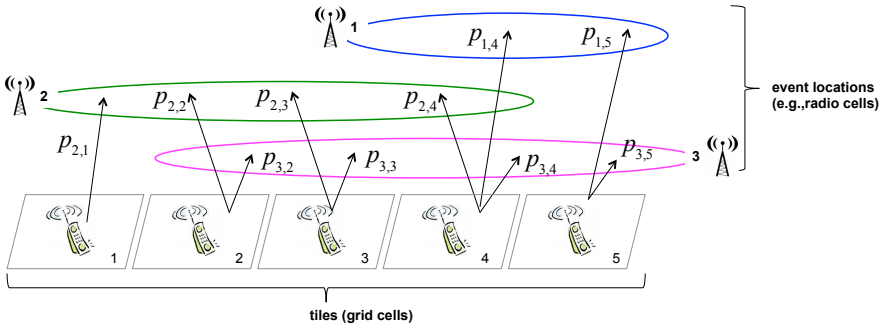


Figure 4: Relation between the different geo-referenced stages. The geo-location stage maps individual records to event locations: it is logically placed in the D-layer and makes the best possible use of the available information from the MNO infrastructure. The inference stage takes in input the event locations and associated counters, and delivers in output the intermediate estimates at the level of individual tiles (or super-tiles). This stage is logically placed at the S-layer, and might take in input also additional information in the form of constraints or priors (e.g., land use maps, transportation network). In the final stage, the final estimates are obtained at the desired level of administrative units by simple aggregation.

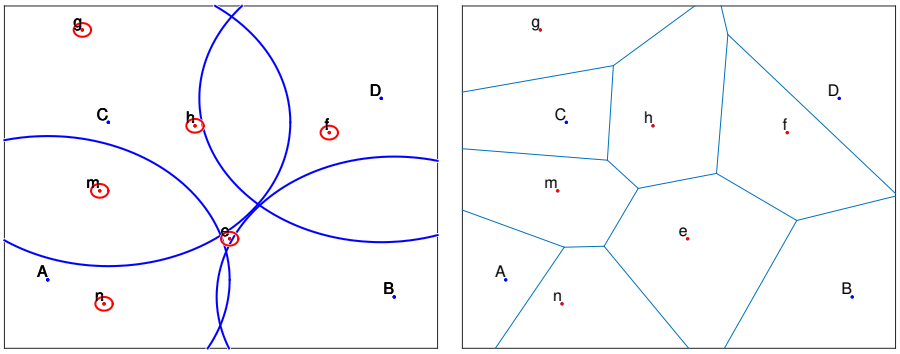


(a) Non-overlapping event locations (tessellation)



(b) Overlapping event locations

Figure 5: Assignment probabilities from tiles to event locations in the data generating process.



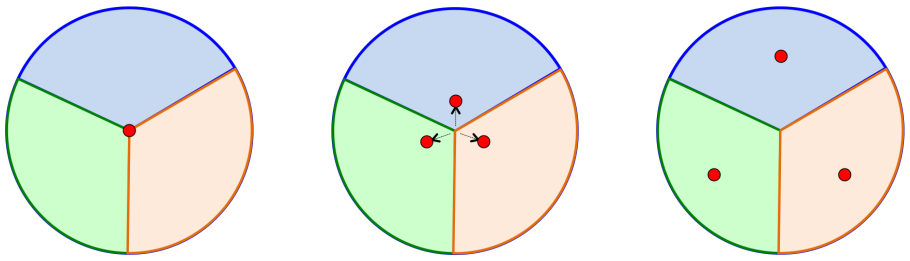
(a) Toy scenario

(b) Naif Voronoi (NV)

(c) Bi-layer Voronoi (BV)

(d) Proximus Voronoi (PV)

Figure 6: Illustration of the three different tessellation options based on a simple toy scenario

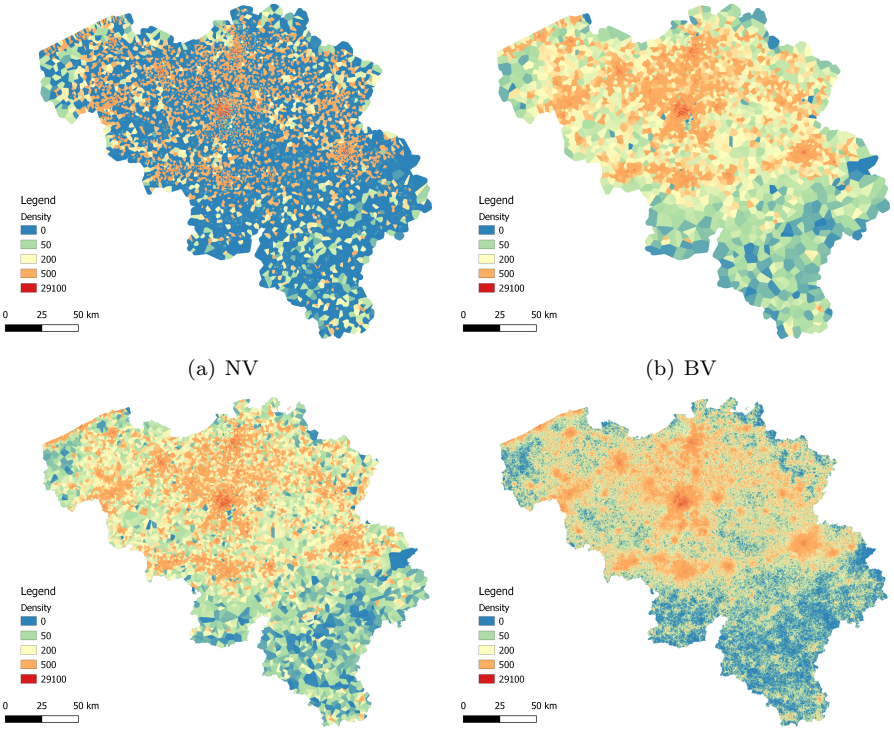


(a) Seed set on antenna tower (adopted in NV, BV).

(b) Seeds set on antenna tower plus small displacements in azimuthal direction (adopted in PV).

(c) Seeds set on barycenters of cell (sector) footprint.

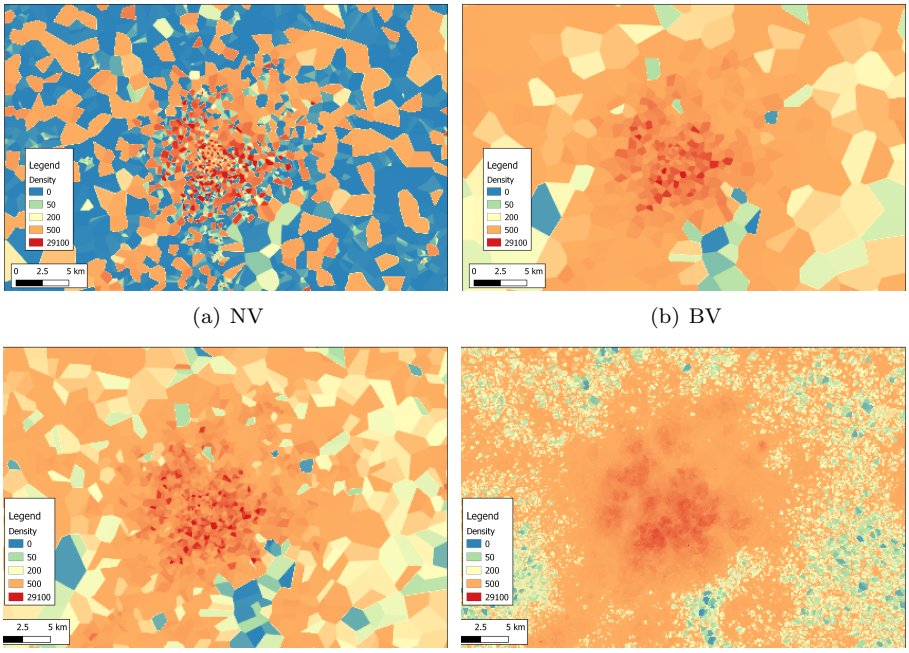
Figure 7: Graphical illustration of the different options for setting seeds in Voronoi tessellation in case of three radio cells (sectors) of 120° beam width sharing a common antenna location.



(a) NV (b) BV

(c) PV (d) OL

Figure 8: Density maps for whole Belgium obtained with different geolocation methods.



(a) NV (b) BV

(c) PV (d) OL

Figure 9: Zoom in around Brussels area (same data as in Fig. 8).

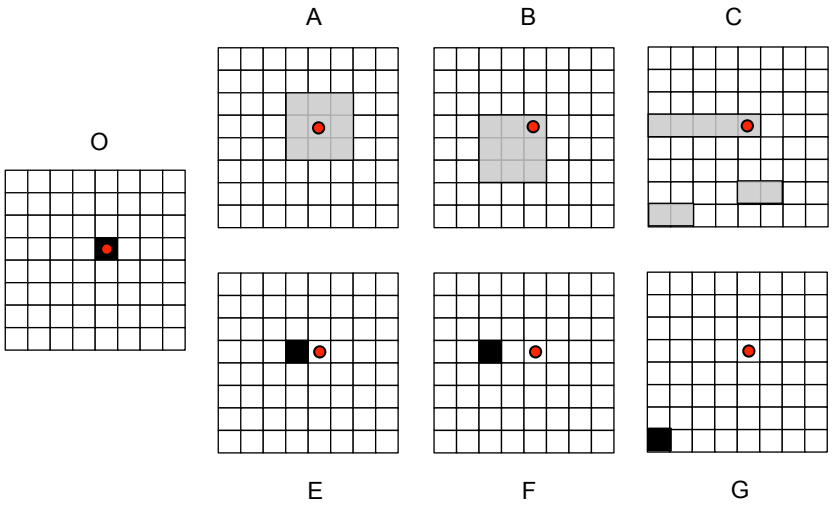


Figure 10: Toy scenario to illustrate the limitations of f-divergence functions to measure distance between spatial distribution. Given the reference distribution labelled as "O", the f-divergence values for maps "A", "B" and "C" are all exactly equal, while obviously map "A" should be preferred. Similarly, the f-divergence values for "E", "G" and "G are all exactly equal, while clearly "E" should be preferred.