JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 25 (2022), pp. 29–66
doi:10.5311/JOSIS.2022.25.166

RESEARCH ARTICLE

Bayesian location estimation of mobile devices using a signal strength model

Martijn Tennekes*

Department of Methodology, Statistics Netherlands, The Netherlands

Yvonne A. P. M. Gootzen

Department of Methodology, Statistics Netherlands, The Netherlands

Received: July 23, 2021; returned: September 24, 2021; revised: December 31, 2021; accepted: October 20, 2022.

Abstract: Mobile network operator (MNO) data are a rich data source for various topics in official statistics, such as present population, mobility, migration, and tourism. Estimating the geographic location of mobile devices is an essential step for statistical inference. Most studies use Voronoi tessellation for this, which is based on the assumption that mobile devices are always connected to the nearest radio cell. We propose an alternative location estimation method following a Bayesian approach and using a physical model for the received signal strength. Our Bayesian framework allows for different modules of prior knowledge about where devices are expected to be, and different modules for the likelihood of connection given a geographic location. We discuss and compare the use of several prior modules, including one that is based on land use. For the likelihood module we propose a signal strength model using radio cell properties such as antenna height, propagation direction, and power. Using Bayes' rule, we derive a posterior probability distribution that is an estimate of the geographic location, which can be used for further statistical inference. We describe the method and provide illustrations of a fictional example that resembles a real-world situation. The method has been implemented in the `R` packages `mobloc` and `mobvis`, which are briefly described.

Keywords: mobile network operator data, mobile phone data, geographic location, present population, Bayesian statistics

*The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

1 Introduction

Mobile network operator (MNO) data have shown to be a rich potential source for official statistics, in particular on present population [2,11,12,26,52,68], mobility [3,10,13,21,23,45, 66,69,72], commuting [29,70], migration [32,35,67], and tourism [12]. The main task in these areas is to estimate human mobility patterns at large scale to support a wide range of policy issues, for instance, sustainable urban transport planning to mitigate climate change [14, 35], or monitoring the spread of the COVID-19 pandemic [4,19,42].

The reason why MNO data is widely considered a gold mine for estimating human mobility is the lack of good alternatives. Traditional surveys on human mobility are useful when the sample size is sufficiently large, but are typically costly and time consuming [8]. Geotagged social media data are a good source for mobility research, but the outcomes may suffer from population bias [34]. MNO data, on the other hand, are readily available, since they are used to facilitate mobile phone communication. Detailed spatial-temporal mobility patterns can be obtained from MNO data [8], also in combination with other sources [8,64].

A key step in statistical inference on MNO data is the estimation of the geographic location of mobile devices [49,52]. Two types of MNO data are required for statistical inference: *cell plan* data, which contain information about the mobile communication network infrastructure, and *event* data, which are data from transaction events between the network and mobile devices. The latter type of data may contain network measurement variables such as Timing Advance which help to estimate the geographic location of devices more accurately.

However, the availability of location related information in event data (as supplied by MNOs) is often limited [58]. The only location related variable that is always available is the identification number of the serving cell. This poses a problem for statistical inference from MNO data based on existing methods which usually require network measurement data [9,24,25,71]. This contradiction between data availability on the one hand, and data requirements of existing methods on the other, implies a need to explore methods with lower data requirements. In particular, we aim to understand whether cell plan data can be sufficient to estimate geographic location.

The main objective of this paper is to design an estimation method that is flexible regarding the type of data used. In other words, the method should work if only cell plan data are provided, and should provide better estimates when network measurement data are available. To this end, we propose a Bayesian location estimation method that, as a primary step, uses a physical model of the received signal strength based on cell plan data. Furthermore, auxiliary data, such as land use, may have predictive value about the location of devices. We explore how such auxiliary data can be used as prior information.

The majority of studies on MNO data use Voronoi tessellation [12] to distribute the geographic location of logged events.[1] However, this approach assumes that a device is always connected to the nearest cell, and therefore ignores cell properties, such as propagation direction, power, height, and tilt, as well as overlap in coverage areas of cells due to *load balancing* (if a cell has reached full capacity, neighbouring cells that also have coverage are able to take over communication with mobile devices). Finally, it does not take into account where devices are expected to be. The Bayesian location estimation approach that we

[1] The geographic area is divided into Voronoi regions such that each Voronoi region corresponds to the geographic location of a cell and each point in that region is closer to that cell tower than to any other cell.

propose allows us to provide a remedy for these shortcomings. Combining this approach with a physical model of the received signal strength additionally avoids the requirement of network measurement data which troubles existing Bayesian approaches.

In sum, the contributions of this paper are as follows. First, we develop a *signal strength model* that uses physical properties of the cells, such as height, propagation direction, and tilt to estimate the propagation per cell. Second, we develop a *signal dominance model* that additionally takes load balancing into account. We use the signal strength and dominance model for the likelihood function. Third, we explore the use of non-uniform prior information. To that end, we develop a prior that is based on land use, a prior that uses (modelled) signal strength data, and a composite prior in which several priors can be mixed. Fourth, we develop a method that updates the posterior using Timing Advance data to ensure that our approach also works with more detailed event data. Fifth, we provide an overview of validation methods that we recommend for further research.[2]

We would like the methods to be applied, tested, and if needed improved as much as possible in future research studies. Therefore we implement the methods as open source software. More specifically, the calculations are implemented in the `R` package `mobloc` [60] and visualization tools that are used to analyse and present the results are implemented in the `R` package `mobvis` [61].

The outline of the paper is as follows. Background and related work are described in Section 2. We introduce the signal strength model in Section 3, and in Section 4 we further introduce the signal dominance model. We present our Bayesian estimation method as a modular framework in Section 5. The likelihood, prior, and timing advance module will be discussed in Sections 5.1, 5.2, and 5.3 respectively. A small example of the framework is provided in Section 5.4. How this framework can be used in statistical inference is discussed in Section 5.5. In Section 6 we provide an overview of validation methods. In Section 7 we illustrate our approach with a fictional application. The implementation is described in Section 8. Conclusion and discussion are provided in Section 9.

2 Background and related work

2.1 Mobile network operator data

A mobile communication network is also called a *cellular network*, where each *cell* enables mobile communication for a specific land area. The cell plan specifies the geographical location of the cells and their physical properties. An MNO often facilitates mobile communication via multiple networks; one for each generation (e.g. 3G, 4G, or 5G), which may also serve at different frequency bands (e.g. 900 , 1800 , and 2100 Hz).

Two types of cells can be distinguished [44]: a *macro cell* that is placed in a cell tower or on top of a roof, which has a range of about 50 kilometers in rural areas, and a *small cell* that is used to enable mobile communication inside buildings and in dense urban areas, and has a theoretical range of two kilometers.[3] An important cell property is the propagation direction. A cell can be directional or omnidirectional. In practice, most real-world

[2] Our original plan was to validate the methods with ground truth data. Although some validation has been done in several internal studies [58], the results could not be shared for reasons of confidentiality.

[3] A small cell is a general name for *micro*, *pico*, and *femto cells*, which have theoretical ranges of respectively 2000, 200, and 20 meters.

deployed macro cells are directional, often covering an angle of about 120 degrees whereas small cells are usually omnidirectional.

We use the following cell plan variables for the signal strength model: cell identification number, geographic location, height, directionality (directional or omnidirectional), azimuth angle (propagation direction; for directional cells only), elevation angle (also known as tilt), power, vertical beam width, and horizontal beam width.

There are generally two types of event data, depending on which transaction events are contained. Event data that are used to calculate the costs in order to bill customers are called *Call Detail Records* (CDR), which contain events related to active mobile phone use, such as initiating a call [4]. Event data that also contains passive events such as location updates, are called *signalling data*. These data are used by the MNO for network analysis and optimization. Signalling data are usually much richer than CDR data and are therefore recommended for statistical inference.

Network measurement variables that may be included in event data are Timing Advance (also known as round-trip time, which is an indication of the distance between cell and device), Received Signal Strength (RSS), and the Signal to Interference plus Noise Ratio (SINR).

2.2 Voronoi tesselation

A couple of variations of the Voronoi algorithm have been proposed to overcome some of the limitations mentioned in Section 1 [11,18,50]. One improvement is to shift the locations of the Voronoi points from the cell tower locations towards the direction of propagation. Alternatively, when the *service area* is known for each cell, i.e. the area which is served by the cell, the location of the centroids of these service areas can be used as Voronoi points. Another improvement is to create a Voronoi tessellation for macro cells, and subsequently assign each small cell to the Voronoi region they are located in. The Voronoi method can be extended with auxiliary data sources, such as land use, to improve the geographic location of devices [22].

2.3 Bayesian location estimation

Several Bayesian estimation methods have been proposed as an alternative to the Voronoi method [24, 25, 41, 71]. The basic principle is that prior information about the expected location of devices is combined with a likelihood, which is a probability that a device is connected to a certain cell given its actual location, taking into account the overlap of nearby cells. The outcome of a Bayesian estimation method is a posterior, which is an estimate of the location of a device given that it is connected to a certain cell. This location is not a single point, but a geospatial distribution that captures the uncertainty.

The Bayesian estimation methods proposed by [24, 25, 71] require specific radio measurement data on event level, namely the variables RSS, SINR, and Timing Advance, but these are often unavailable in practice [39]. In contrast, the estimation method we propose does not require any of these variables, and can therefore be applied in case network

[4] CDR data contain records about calls (initiating and receiving), SMSes (sending and receiving), and mobile data usage. Note that in several studies the term CDR is used for data that only contains call and SMS events, and alternative terms *Data Detail Records* (DDR) or *Event Data Records* (EDR) are used for data that also include mobile data usage events.

measurement data are unavailable. We illustrate how network measurement data (Timing Advance) can be used optionally to improve the estimations.

Recently, another Bayesian estimation method has been proposed [41] that seems similar to ours at first glance. However, a crucial difference is that this method follows an empirical approach; from aggregated Timing Advance data, the geographic distribution of connections in the past are extracted and used in the connection likelihood. Our approach is more theoretical, because it uses a propagation model for the likelihood that does not use any empirical data. Furthermore, we use Timing Advance data (if available) in a different way; we estimate the location of a device given the identification number of the cell and the measured Timing Advance value.

The proposed Bayesian methods assume a uniform distribution for the prior, which means no additional information is used about where devices are expected to be. While a uniform prior generally speaking represents an "objective" prior, we argue this is not always true for location estimation of mobile phones.

2.4 Spatial density estimation

Our method is focused on the location estimation of a single device, including its uncertainty. For static applications, such as the estimation of the present population, the next step in statistical inference is to estimate the spatial density of devices. For this purpose, two methods have been introduced recently [31,48,50] that use the likelihood probabilities presented in this paper (created with the signal strength and signal dominance model), which are also referred to as *emission probabilities*.[5] To estimate the spatial density of devices, rather than the location of a single device, both methods also require the number of connected devices per cell at a certain time, which can be obtained from event data.

The first method is the *Maximum Likelihood Estimator (MLE)* [31, 50]. The locations of the devices are assumed to be generated by a random (multinomial [50] or Poisson [31]) process that uses the emission probabilities. The maximum likelihood is a spatial density that best explains the observed data, which in this case consist of the number of connected devices per cell.

A downside of the MLE is that is does not take prior information into account. An enhancement of the MLE has been proposed that also takes prior information into account [48]. This method uses the emission probabilities to compute multiple maximum likelihoods, and subsequently selects the one that best confirms the prior data. Therefore, this method is called the *Data First (DF)* method.

2.5 Radio propagation

Advanced radio propagation (signal strength) models have been proposed in the field of electrical engineering, e.g. [30,38,51], but these require complex technical specifications and configurations of the antennas. This information is rarely available for purposes other than network analysis and optimization [38, 54]. Our propagation model is simple in comparison, but can be applied with just the cell plan data that is usually available for statistical purposes [58].

[5] Both [31] and [48] refer to earlier work that led to this paper [59, 62].

3 Signal strength model

In the methods that we will describe, we will overlay the geographic area of interest with a grid. The main advantage of using grid tiles is that different geospatial datasets can be combined without the need to calculate spatial intersections, which is a time consuming operation. Moreover, the mathematics described below are easier since all grid tiles have the same area. Henceforth, we denote $\mathcal{G}$ as the set of all grid tiles, and use the letter g to denote one such grid tile. The set of all cells is denoted with $\mathcal{A}$, where each element a is a single cell.

This section describes a physical model of the propagation of signal strength originating from a single cell, by applying basic physical laws for signal propagation [56]. We distinguish two types of cells: omnidirectional and directional, resulting in two different propagation models. Omnidirectional cells have no aimed beam and their coverage area can be thought of as a circular disk. Directional cells point in a certain direction and their coverage area can be thought of as an oval with one axis of symmetry. In practice, small cells are usually omnidirectional and macro cells (i.e. attached to cell towers or placed on rooftops) are often directional [27].

3.1 Omnidirectional cells

For omnidirectional cells, propagation of the signal strength $S(g, a)$ is modelled as

$$S(g, a) := S_0 - S_{\text{dist}}(r_{g,a}), \tag{1}$$

where S_0 is the signal strength at $r_0 = 1$ meter distance from the cell in dBm (decibel-milliwatts) and $r_{g,a}$ is the distance between the middle point of grid tile g and cell a in meters. The value of S_0 can be different for every cell and is assumed to be a known property. In cell plan data, it is common to list the power P of a cell in Watt, rather than the signal strength in dBm. The value of S_0 can be calculated from P using the conversion between Watt and dBm [16]:

$$S_0 = 30 + 10 \log_{10}(P). \tag{2}$$

The function $S_{\text{dist}}(r)$ returns the loss of signal strength as a function of distance r:

$$S_{\text{dist}}(r) := 10 \log_{10}(r^\gamma) = 10\gamma \log_{10}(r), \tag{3}$$

where γ is the *path loss exponent*, which resembles the reduction of propagation due to reflection, diffraction and scattering caused by objects such as buildings and trees [55]. In free space, γ equals 2, but in practice higher values should be used. As a rule of thumb, 4 can be used for urban areas and 6 for indoor environmentsËIJ [51, 55]. Special situations, such as tunnels, could improve the propagation such that a value of less than 2 is applicable. The path loss exponent can be approximated by using the land use register.

In Fig. 1, the signal loss as a function of the distance is shown for a cell with $10\,\mathrm{W}$ power that is standing in an urban environment ($\gamma = 4$).

3.2 Directional cells

A directional cell is a cell that is aimed at a specific angle. Along this angle, the signal strength is received at its best. However, the signal can also be strong in other directions. It

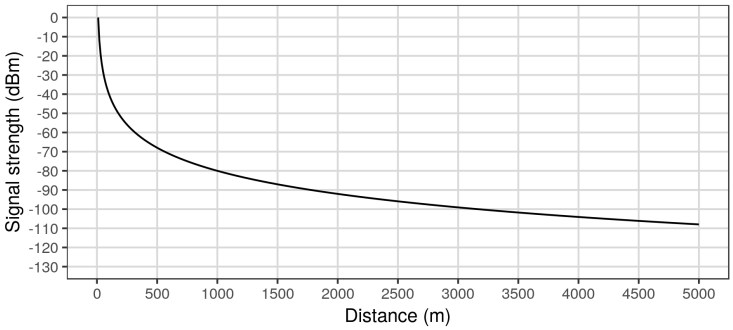


Figure 1: Signal strength as a function of the distance for a specific cell.

is comparable to a speaker producing sound in a specific direction. The sound is audible in many directions, but is much weaker at the sides and the back of the speaker. We specify the beam of a directional cell a by four parameters:

- The azimuth angle φ_a is the angle from the top view between the north and the direction in which the cell is pointed, such that $\varphi_a \in [0, 360)$ degrees. Note that cell towers and rooftop cells often contain three cells with 120 degrees in between.
- The elevation angle θ_a is the angle between the horizon plane and the tilt of the cell. Note that this angle is often very small, typically only four degrees. The plane that is tilt along this angle is called the *elevation plane*.
- The horizontal beam width α_a specifies in which angular difference from the azimuth angle in the elevation plane the signal loss is $3\,\mathrm{dB}$ or less. At $3\,\mathrm{dB}$, the power of the signal is halved. The angles in the elevation plane for which the signal loss is $3\,\mathrm{dB}$ correspond to $\varphi_a \pm \alpha_a/2$. In practice, these angles are around 65 degrees.
- The vertical beam width β_a specifies the angular difference from θ_a in the vertical plane orthogonal to φ_a in which the signal loss is $3\,\mathrm{dB}$. The angles in which the signal loss is $3\,\mathrm{dB}$ correspond to $\theta_a \pm \beta_a/2$. In practice, these angles are around 9 degrees.

Let $\delta_{g,a}$ be the angle in the elevation plane between the azimuth angle φ_a and the orthogonal projection on the elevation plane of the line between the center of cell a and the center of grid tile g. Similarly, let $\varepsilon_{g,a}$ be the angle from the side view between the line along the elevation angle θ_a and the line between the center of cell a and the center of grid tile g. Note that $\varepsilon_{g,a}$ depends on the cell property of the installation height above ground level. We model the signal strength for directional cells as

$$S(g,a) := S_0 - S_{\text{dist}}(r_{g,a}) - S_{\text{azi}}(\delta_{g,a}, \alpha_a) - S_{\text{elev}}(\varepsilon_{g,a}, \beta_a), \tag{4}$$

where S_0 is the signal strength at $r_0 = 1$ meter distance from the cell, in the direction of the beam so that $\delta = 0$ and $\varepsilon = 0$. The signal loss due to distance to the cell, azimuth angle difference and elevation angle difference is specified by S_{dist}, S_{azi} and S_{elev}, respectively. The definition of S_{dist} is similar to the omnidirectional cell and can be found in Eq. (3).

Each cell type has its own signal strength pattern for both the azimuth and elevation angles [56]. These patterns define the relation between signal loss and the offset angles, i.e., $\delta_{g,a}$ for the azimuth and $\varepsilon_{g,a}$ for the elevation angles. We model the radiation pattern for

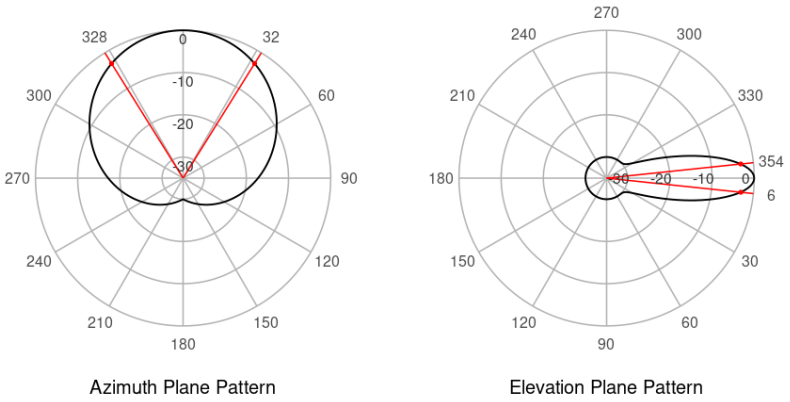


Figure 2: Radiation patterns for the azimuth and elevation planes.

both S_{azi} and S_{elev} by fitting a Gaussian curve using two parameters: the beam width and the signal loss in the opposite propagation direction, which we set to $30\,\text{dB}$ by default. For implementation details we refer to [60] which is briefly described in Section 8.

The resulting patterns are shown in Fig. 2. The black line shows the relation between signal loss and angle in the azimuth plane (left) and elevation plane (right). The grey circles correspond to the signal loss; the outer circle means $0\,\text{dB}$ loss (which is only achieved in the main direction), the next circle corresponds to $5\,\text{dB}$ loss, an so forth. The red lines denote the angles corresponding to $3\,\text{dB}$ loss. The angle between the red lines is α_a in the azimuth plane and β_a in the elevation plane.

Although these models approximate the general curve of real radiation patterns, the radiation patterns are more complex in reality, e.g. they often contain local spikes caused by so-called side and back lobes [56].

Fig. 3 (top row) illustrates the signal strength at the ground level from above for a specific cell. In this case, the cell is placed at $x = 0$, $y = 0$ at 55 meters above ground level in an urban environment ($\gamma = 4$), has a power of $10\,\text{W}$, and is directed eastwards with an elevation angle (tilt) of 5 degrees, a horizontal beam width of 65 degrees and a vertical beam width of 9 degrees. Notice that the signal strength close to the cell, which on ground level translates to almost under the cell, is lower than at a couple of hundred meters distance. This is caused by relatively large ε angles at grid tiles nearby the cell.

Table 1: Indication of quality for signal strength in 4G networks.

Signal strength (dBm)	Quality
-70 or higher	Excellent
-90 to -70	Good
-100 to -90	Fair
-110 to -100	Poor
-110 or less	Bad or no signal

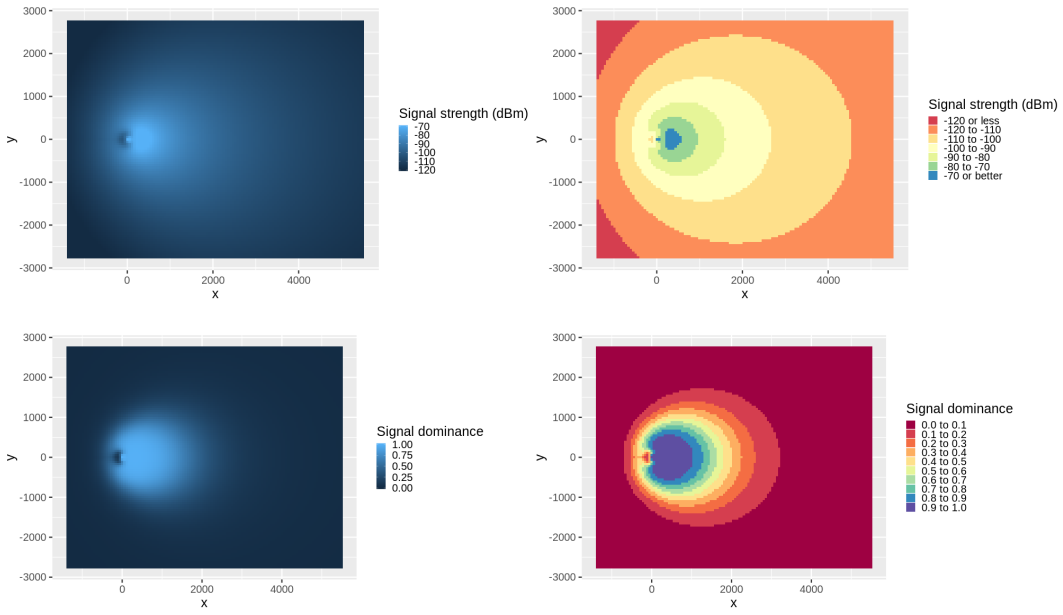


Figure 3: Signal strength (top row) and signal dominance (bottom row) at ground level.

4 Signal dominance model

The assignment of a cell to a mobile device does not only depend on received signal strength, but also on the capacity of the cells. The process of assigning devices to cells while taking into account the capacity of the cells is called *load balancing*. To capture this process we introduce a transformation of the modelled signal strength into a variable that we call *signal dominance*.

With signal dominance we aim to compensate for two phenomena. The first is the switching of a device when it is receiving a bad signal to a cell with a better signal. Table 1 describes how the signal strength can be interpreted in terms of quality for 4G networks [27]. The second phenomenon is the switching between cells that is influenced by some decision making system in the network that tries to optimize the load balancing within the network [54]. The specifics of this system are considered unknown.

What we do know is that MNOs aim to minimize the number of handovers to reduce costs [7]. A *handover* is the process of switching the connection of a device from one cell to another. At the same time, MNOs aim to provide sufficient received signal strength for the connected devices. What an MNO considers to be sufficient is unknown, but we assume that they aim for a certain threshold; a lower signal strength is a trigger for the network to switch to a better cell. Because of handover costs, we assume that in case the received signal strength of the connected cell is already above this threshold, the network only switches to another cell for load balancing reasons. We assume that this threshold value is approximately $-90\,\mathrm{dBm}$, which is the signal strength that is considered as good according to Table 1ËIJ [27].

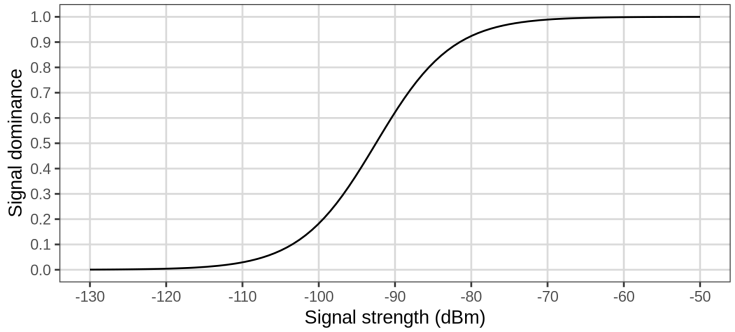


Figure 4: Logistic relation between signal strength (dBm) and signal dominance, where S_{mid} and S_{steep} are set to $-92.5\,\mathrm{dBm}$ and $0.2\,\mathrm{dBm}$ respectively to resemble Table 1.

To model this take on the load balancing mechanism, we use a logistic function to translate the signal strength $S(g, a)$ to the more interpretable signal dominance measure $s_{\text{dom}}(g, a)$. Let us define

$$s_{\text{dom}}(g, a) := \frac{1}{1 + \exp\big(-S_{\text{steep}}\,(S(g, a) - S_{\text{mid}})\big)}, \tag{5}$$

where S_{mid} and S_{steep} are parameters that define the midpoint and the steepness of the curve respectively. Fig. 4 shows an example of Eq. (5).

The signal dominance at ground level is shown in Fig. 3 (bottom row). Compared to signal strength shown in Fig. 3 (top row), signal dominance is more focused on the area that is in close proximity (in this example between 0 and 1500 meter) of the cell eastwards (the propagation direction).

We will apply signal dominance to estimate the probability that a device will connect to a certain cell. Suppose there are only two cells available for connection to an unconnected device with received signal strength values S_1 and S_2, and signal dominance values s_{dom1} and s_{dom2}. Then the probability that the first cell is chosen is estimated as $s_{\text{dom1}}/(s_{\text{dom1}} + s_{\text{dom2}})$.

Fig. 5 shows these probability values in rounded percentages for a sequence of values for S_1 and S_2. A percentage can be interpreted as the number of times out of one hundred cases the network prefers the first cell over the second. This percentage is 50% on the diagonal, i.e. when $S_1 = S_2$. However, the percentages are close to 50 when both S_1 and S_2 are sufficiently large. This follows our earlier made assumption, namely that above some threshold (in this example around $-90\,\mathrm{dBm}$) the cell choice does depend (much) on signal strength, but on other factors, such as cell capacity.

We will use these probabilities in the likelihood function we propose for the Bayesian estimation method that is introduced in the next section.

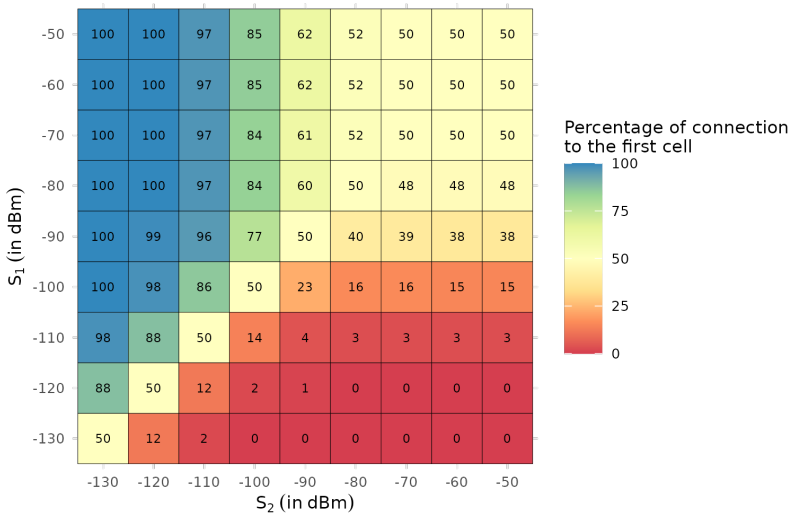


Figure 5: Percentages of connection to the first cell according to the signal dominance model (with parameters $S_{\text{steep}} = 0.2$ and $S_{\text{mid}} = 92.5$).

5 A modular Bayesian framework

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

$$\mathbb{P}(g \mid a) \propto \mathbb{P}(g)\mathbb{P}(a \mid g), \tag{6}$$

where g represents a grid tile and a a cell. The probability $\mathbb{P}(g)$ that a device is located in grid tile g without any connection knowledge represents the location prior about the relative frequency of events at grid tile g. The connection likelihood $\mathbb{P}(a \mid g)$ is the probability that a device is connected to cell a given that the device is located in grid tile g. The location posterior $\mathbb{P}(g \mid a)$ represents the probability that a device is located in grid tile g given that the device is connected to cell a.[6]

The position of Eq. (6) in our modular Bayesian framework is illustrated in Fig. 6. The location prior and connection likelihood are estimates produced from models we call the location prior module and connection module, respectively. For the location prior we can use different input data sources, such as land use data and cell plan data. The connection module uses cell plan data. The location prior is then updated by the connection likelihood to form the location posterior. The localisation method can at this point be considered as complete and this posterior may be used for further statistical inference. Alternatively, the posterior can be seen as a new prior and updated again with other likelihoods if the

[6]We note that the expression $\mathbb{P}(a \mid g)$ plays two distinct roles in our discussion. When we model $\mathbb{P}(g \mid a)$ via Eq. (6), the cell a is considered fixed. Hence $\mathbb{P}(a \mid g)$ is not a probability distribution, but a likelihood function with parameter g. On the other hand, when we will propose a model for $\mathbb{P}(a \mid g)$ in Section 5.1, it will be considered as a probability distribution over a for fixed g. Although this distinction is not critical for the reader's understanding of our model, the intended interpretation should be clear from context. For simplicity, we will refer to $\mathbb{P}(a \mid g)$ as the connection likelihood henceforth.

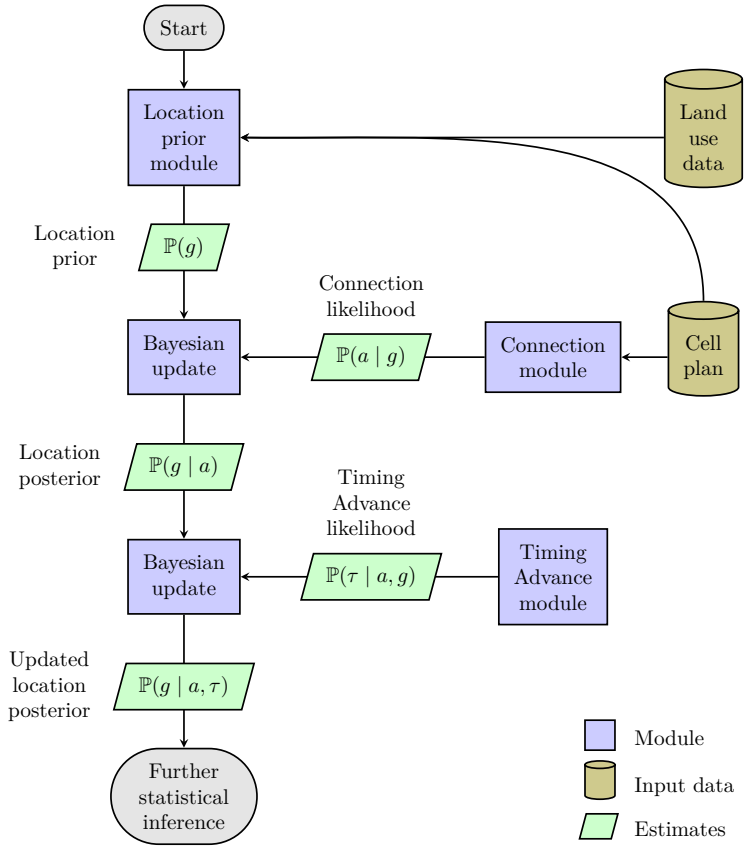


Figure 6: A modular framework for modelling location posteriors.

necessary information is available. Fig. 6 shows how a Timing Advance module can be used to construct an updated location posterior.

The rest of this section is devoted to proposals for, and elaborations of these modules.

5.1 Connection likelihood

We define the *connection likelihood* $\mathbb{P}(a \mid g)$ for a cell a and a grid tile g to be the probability that when a device located in grid tile g generates an event at some cell, it does so at a. We model this probability as

$$\mathbb{P}(a \mid g) := \frac{s(g,a)}{\sum_{a' \in \mathcal{A}} s(g,a')}, \tag{7}$$

where $\mathcal{A}$ is the set of all cells in the MNOs network and $s(g,a) \in [0,\infty)$ is an indication of how well the connection of cell a is in grid tile g.

Different choices for modelling s are possible, and any choice defines the *connection module* in Fig. 6. For instance, when signal strength S is used, the connection likelihood

becomes the ratio of the signal strength received from cell a to the total value of signal strength received from all cells. We propose to use signal dominance s_{dom} for the reasons explained in Section 4.

When cell plan data is lacking, but service areas are known or can be estimated $s(g, a)$ is set to be

$$s_F(g, a) := \begin{cases} 1 & \text{if } g \in F(a), \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $F(a)$ is called the *service area* which is defined as the set of grid tiles in which cell a is available for connection. We assume that the network has perfect coverage, so that for each grid tile $g \in \mathcal{G}$ there exists at least one cell a for which $g \in F(a)$. Note that these service areas may overlap, so a grid tile g may belong to multiple service areas. When using s_F instead of s, the connection likelihood defined by (7) can be rewritten as

$$\mathbb{P}_F(a \mid g) := \begin{cases} \left|\{a' \in \mathcal{A} \mid g \in F(a')\}\right|^{-1} & \text{if } g \in F(a), \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

A special case of these service areas is a tessellation where service areas do not overlap. The most common example is the Voronoi tessellation, where $\text{Vor}(a)$ is the set of grid tiles of which the centroids lie in the Voronoi region surrounding cell a. In this case, $s(g, a)$ is set to be

$$s_{\text{Vor}}(g, a) := \begin{cases} 1 & \text{if } g \in \text{Vor}(a), \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The connection likelihood can hence be simplified to

$$\mathbb{P}_{\text{Vor}}(a \mid g) = s_{\text{Vor}}(g, a). \tag{11}$$

5.2 Location prior

We define the *location prior* $\mathbb{P}(g)$ as the probability that a device is present in grid tile g, such that

$$\sum_{g \in \mathcal{G}} \mathbb{P}(g) = 1, \tag{12}$$

where $\mathcal{G}$ is the set of all possible location estimates in our model, in other words, the whole grid.

The definition of the location prior function will be based on assumptions about where a device is expected to be. In this section, we propose four options: the uniform prior, the land use prior, the network prior, and the composite priors.

5.2.1 Uniform prior

When we use the *uniform prior*, we assume the probability of a device being in any grid tile is the same value for every grid tile:

$$\mathbb{P}_{\text{uniform}}(g) := \frac{1}{|\mathcal{G}|}. \tag{13}$$

A uniform prior is sometimes viewed as uninformative. In the case of mobile phone data, however, the implicit assumption that any grid tile is as likely as the next can lead to an underestimation of devices in urban areas and an overestimation of devices in rural areas. We therefore advise against using the uniform prior as a default prior without consciously assessing the plausibility of the underlying assumption.

When the uniform prior is combined with the connection likelihood in which the Voronoi tessellation is used (or any other tessellation), see (11), the location posterior becomes

$$\mathbb{P}_{\text{Vor}}(g \mid a) = \begin{cases} \left|\{g' \in \mathcal{G} \mid g' \in \text{Vor}(a)\}\right|^{-1} & \text{if } g \in \text{Vor}(a), \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

5.2.2 Land use prior

An alternative to the assumption of uniformity is to use administrative sources on land use for the location prior. One would for example expect more devices in an urban area than in a rural area. The *land use prior* is based on a proportional expectation of the number of devices $n(g)$ in grid tile g such that:

$$n(g) \propto \mathbb{E}[\text{number of devices in } g] \quad \text{for all } g \in \mathcal{G}. \tag{15}$$

The land use prior is then defined as:

$$\mathbb{P}_{\text{land use}}(g) := \frac{n(g)}{\sum_{g' \in \mathcal{G}} n(g')}. \tag{16}$$

Due to the normalisation in its definition, the land use prior does not require an explicit estimate of the number of devices per grid tile. Any proportional measure has the same effect. One way to utilise this is when information is available on land use classes of the grid, such as levels of urbanisation. Let there be K land use classes, each with their own relative expected number of devices: $u_1, u_2, \ldots, u_K$. Let $w_1(g), w_2(g), \ldots, w_K(g) \in [0, 1]$ be the proportion of the grid tile that is covered by each class respectively, such that

$$\sum_{k=1,\ldots,K} w_k(g) = 1 \quad \text{for all } g \in \mathcal{G}. \tag{17}$$

Then $n(g)$ can be modelled as

$$n(g) := \sum_{k=1,\ldots,K} u_k \cdot w_k(g). \tag{18}$$

Before constructing such a land use prior, it is important to decide which data source to use for land use, and subsequently which classes to use. For countries in which official land use registers are unavailable, we recommend to use OpenStreetMap [15, 17, 53] (see https://osmlanduse.org).

There are are many possibilities for how to choose $u_1, u_2, \ldots, u_k$. The ideal situation is to set these numbers using auxiliary data, for instance vehicle count data [46] or public transport data [20]. However, in many applications, such data is unavailable. Therefore, the following approach can be used to set these numbers manually with general knowledge about the region of study:

- We advise to assign 1 to the class where most people are expected to be under normal conditions. Note that this depends on several factors such as time, day, and country. For estimating the nighttime population [11,12], it would make sense to set residential land as the category where most people are expected to be. However for daytime urban mobility analysis [2, 68], people are also expected to be in other urban land classes, such as retail, industry, and transport areas.
- Next, the value of 0 can be assigned to classes where one does not expect any people to be. Examples are water surfaces and wetlands, except for water sport regions and waterways.
- The value for the other land use classes could be chosen with common sense, keeping the factors such as time, date, and region into account. A small number may be assigned to forest and agricultural land, whereas higher numbers to other classes of built up land are recommended.

One of the downsides of the land use prior is that the assumptions based on administrative sources are less flexible in the case of major events. A festival in a location which would ordinarily be a quiet meadow, but suddenly contains many devices, is not accounted for in the land use prior. Such events can be recognized by the positioning and setup of the cells. For instance, extra small cells are often used to compensate for large numbers of devices [63, 65].

It can also be worthwhile to let the land use prior depend on the time and day, because the expected number of devices may vary over time. For instance, in industrial areas the expected number of devices is usually smaller during nighttime and weekends compared to daytime and working days.

5.2.3 Network prior

The following prior, which we call the *network prior*, is defined as

$$\mathbb{P}_{\text{network}}(g) := \frac{\sum_{a \in \mathcal{A}} s(g,a)}{\sum_{a \in \mathcal{A}} \sum_{g' \in \mathcal{G}} s(g',a)}. \tag{19}$$

Recall that $s(g,a)$ is a general indication of the quality of the connection of cell a is in grid tile g. When we use the modelled signal dominance s_{dom}, the expected probability $\mathbb{P}(g)$ becomes the total signal dominance in g divided by the total signal dominance in the whole grid. Basically, the network prior reflects the distribution of the total signal over all the grid tiles.

Note that we also used s in our connection likelihood, described in Section 5.1. When this network prior is used in combination with the connection likelihood $\mathbb{P}(a \mid g)$ using the same indicator s, then Eq. (6) can be simplified to

$$\mathbb{P}_{\text{network}}(g \mid a) \propto s(g,a). \tag{20}$$

The example in Section 5.4 includes calculations which illustrate this.

When using a service area indicator, defined as s_F in (8), is used, the network prior becomes

$$\mathbb{P}_{\text{network}}(g) := \frac{\left|\{a' \in \mathcal{A} \mid g \in F(a')\}\right|}{\sum_{a \in \mathcal{A}} \left|F(a)\right|}. \tag{21}$$

Observe that when a tessellation is used, such as the Voronoi tessellation, then this prior is uniform. In a tessellation, service (or Voronoi) areas do not overlap. Therefore, when assuming full network coverage (so each grid tile g is contained in exactly one service area), the nominator of (21) is 1, and the denominator is equal to the total number of grid tiles in $\mathcal{G}$.

This prior contains implicit knowledge about where an MNO is expecting people. The placement of cells is not without reason; generally, more cells are placed in crowded areas, such as city centers, than in quiet rural areas. Note that we could have defined the network prior using the cell density. However, since the network capacity also depends on the type and configuration of the cells and on the environment (buildings and trees will generally have a negative effect on the propagation of the signal) we use the signal dominance, through which these aspects are taken into account.

There are two aspects to be aware of when using the network prior. First, the placement of cells is based on estimated peak traffic rather than the average expected number of devices. MNOs normally provide better network coverage in railway stations than in residential areas, since the estimated peak traffic is higher; people typically use their phone more actively in railway stations and moreover, the expected number of devices fluctuates more over time. The second aspect to be aware of is that MNOs might place extra (partially overlapping) cells in an area for reasons other than an expected increase of the number of devices. They might do this, for example, when they detect that the quality of the network connection is insufficient on certain sections of land. In summary, the total signal strength of the network does not always reflect the estimated number of devices.

5.2.4 Composite priors

Our fourth and final proposed prior is less theoretically substantiated and more driven by practical considerations. One can combine all three priors described earlier as follows:

$$\begin{aligned}\mathbb{P}_{\text{composite}}(g) := {} & \pi_{\text{uniform}} \cdot \mathbb{P}_{\text{uniform}}(g) + \\ & \pi_{\text{land use}} \cdot \mathbb{P}_{\text{land use}}(g) + \\ & \pi_{\text{network}} \cdot \mathbb{P}_{\text{network}}(g),\end{aligned} \tag{22}$$

where $\pi := (\pi_{\text{uniform}}, \pi_{\text{land use}}, \pi_{\text{network}})$ is any vector in $\mathbb{R}^3$ such that $0 \leq \pi_{\text{uniform}} \leq 1$, $0 \leq \pi_{\text{land use}} \leq 1$, $0 \leq \pi_{\text{network}} \leq 1$, and $\pi_{\text{uniform}} + \pi_{\text{land use}} + \pi_{\text{network}} = 1$.

The components of π represent the contributions of the three previously defined priors to the final *composite prior* $\mathbb{P}_{\text{composite}}(g)$. Hence both the advantages and disadvantages of all three priors are mixed. The aim of the composite prior is to have a good balance, specified with π, between these advantages and disadvantages. What a good balance is depends on the application. A composite prior is harder to interpret from a theoretical point of view compared to the three priors discussed so far, which can be seen as an additional disadvantage.

A sensitivity analysis can be used to assess how the location posterior is affected by changes in π. The approach of such a sensitivity analysis is the following.

- For the area of interest, construct the candidate priors, e.g. the ones we discussed.
- Compute the composite prior for various values of π.
- With a fixed connection likelihood, compute the location posterior for each of the composite priors.
- Compare the location posteriors. This can be done visually or by using the a quantitative distance metric such as the Kantorovich-Wasserstein Distance [5] (KWD), which is used to compare two probability distributions.

5.3 Incorporating Timing Advance data

Some MNOs include in their signalling data a so called *Timing Advance* variable [1]. Timing Advance is also known as round-trip time [71], or trip-time bins [41]. The value of such a variable represents a time duration, and it is used to estimate and adjust for communication delays. In combination with the speed of radio waves it can alternatively be used to estimate the distance between device and cell [28].

In the case of 4G signalling data the Timing Advance variable takes on values in the discrete set $\mathcal{T} = \{0, 1, \ldots, 1282\}$. If an event then contains such a value τ for this variable, the associated device is located approximately in an annulus centered around the antenna of width $78\,\mathrm{m}$ and an inner circle of radius $\tau \cdot 78\,\mathrm{m}$. For instance, $\tau = 100$ implies that the distance between cell and device is approximated between $7800\,\mathrm{m}$ and $7878\,\mathrm{m}$. Since there are $|\mathcal{T}| = 1283$ annuli, the maximum theoretical distance is $1283 * 78\,\mathrm{m} \approx 100\,\mathrm{km}$.

When 5G signalling data are used, the distance can be estimated more accurately, because the widths of the annuli are $39\,\mathrm{m}$. Furthermore, since τ takes on values in the discrete set $\mathcal{T} = \{0, 1, \ldots, 3846\}$, there are $|\mathcal{T}| = 3847$ annuli, and therefore the maximum theoretical distance is $3847 * 39\,\mathrm{m} \approx 150\,\mathrm{km}$.

One should be aware, though, of errors present in this approximation. As the authors of [47] observed, the Timing Advance variable may take on different values while the device has the same distance to the cell, and even when the device is at the same location at different times.

Knowledge of τ may be used to improve the location posterior $\mathbb{P}(g \mid a)$ through Bayesian updating:

$$\mathbb{P}(g \mid a, \tau) \propto \mathbb{P}(g \mid a)\mathbb{P}(\tau \mid a, g), \tag{23}$$

where the *Timing Advance likelihood* $\mathbb{P}(\tau \mid a, g)$ can be modelled as the fraction of the grid tile g which lies in the annulus around the cell a specified by τ. Since τ has a certain maximum value (1282 for 4G and 3846 for 5G), this new location posterior $\mathbb{P}(g \mid a, \tau)$ equals 0 for grid tiles farther than the maximum theoretical distance ($100\,\mathrm{km}$ for 4G and $150\,\mathrm{km}$ for 5G).

Computing the likelihoods $\mathbb{P}(\tau \mid a, g)$ for all Timing Advance annuli around all cells a in the network and all tiles g in the grid that is used might prove to be too expensive computationally if calculated in the way we suggest above. One could therefore model $\mathbb{P}(\tau \mid a, g)$ more coarsely as being 1 if the centroid of g lies in the annulus specified by τ, and 0 otherwise. If the grid tiles used are substantially larger than the widths of the annuli ($78\,\mathrm{m}$ for 4G and $39\,\mathrm{m}$ for 5G), though, such as the 100 by 100 meter tiles we propose, this coarser Timing Advance likelihood model could increase the location estimation error for devices located in the smaller annuli. These errors can be mitigated by merging adjacent

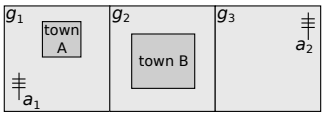


Figure 7: A schematic top view of an island of 1 by 3 kilometers.

annuli. For example, one could model

$$\mathbb{P}(\tau \mid a, g) := \begin{cases} 1 & \text{if } g \in \text{Ann}(a, \tau, b), \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

where $\text{Ann}(a, \tau, b)$ specifies the grid tiles whose centroids lie in the annuli of cell a specified by $\{\tau - b, \ldots, \tau + b\}$, where b is a globally defined integer, independent of τ, a and g, that determines how many annuli are merged on both sides to the annulus corresponding to τ.

5.4 Calculation example

To illustrate the computations involved in the model, we consider a small fictional island of 1 by 3 kilometers. All numerical values mentioned in this example can be found in Table 2. The island can be divided into three grid tiles of equal size, g_1, g_2 and g_3. Note that for a more realistic example we would use much smaller grid tiles, but for simplicity each tile in this example measures 1 by 1 kilometer.

There are two towns A and B on the island, of which the latter is about three times as large as the former. Two cells, a_1 and a_2, have been installed. These are illustrated in Fig. 7. Cell a_1 has a perfect signal in g_1 and g_2, but no signal in g_3, while cell a_2 has a perfect signal in g_3 and g_2, but no signal in g_1. In this example we express a perfect signal and no signal with values 1 and 0 respectively.

We calculate the four priors as defined in Section 5.2 based on the above information. The connection likelihood for each cell is calculated according to Eq. 7 based on the signal dominance values. Finally, the connection likelihood and priors are combined to location posteriors for all combinations of cells and priors.

The location posterior allows for all kinds of further calculations such as modelling the distribution of devices or even persons over grid tiles, when connection events are counted for each cell during a time interval. This process is complicated enough to be viewed as a separate research topic, especially when one is interested in probability distributions rather than point estimates. We consider it part of the *further statistical inference* in the framework from Fig. 6.

5.5 Statistical inference

The outcome of the modular system described in this paper is the location posterior $\mathbb{P}(g \mid a)$, which specifies the probability that a device is located in grid tile g, given that it is connected to cell a. This can be used to calculate the total number of devices that are present at a specific location during a specific time interval, or the number of devices that move from one city to another. However, many applications in official statistics are about

Table 2: The corresponding numbers of the fictional example where the composite prior is based on $\pi := (\pi_{\text{uniform}}, \pi_{\text{land use}}, \pi_{\text{network}}) = (0, 1/2, 1/2)$.

		Grid tile g		
		g_1	g_2	g_3
Signal dominance	$s(g, a_1)$	1	1	0
	$s(g, a_2)$	0	1	1
Location priors	$\mathbb{P}_{\text{uniform}}(g)$	$1/3$	$1/3$	$1/3$
	$\mathbb{P}_{\text{land use}}(g)$	$1/4$	$3/4$	0
	$\mathbb{P}_{\text{network}}(g)$	$1/4$	$2/4$	$1/4$
	$\mathbb{P}_{\text{composite}}(g)$	$1/4$	$5/8$	$1/8$
Connection likelihood	$\mathbb{P}(a_1 \mid g)$	1	$1/2$	0
	$\mathbb{P}(a_2 \mid g)$	0	$1/2$	1
Location posterior	$\mathbb{P}_{\text{uniform}}(g \mid a_1)$	$2/3$	$1/3$	0
	$\mathbb{P}_{\text{uniform}}(g \mid a_2)$	0	$1/3$	$2/3$
	$\mathbb{P}_{\text{land use}}(g \mid a_1)$	$2/5$	$3/5$	0
	$\mathbb{P}_{\text{land use}}(g \mid a_2)$	0	1	0
	$\mathbb{P}_{\text{network}}(g \mid a_1)$	$1/2$	$1/2$	0
	$\mathbb{P}_{\text{network}}(g \mid a_2)$	0	$1/2$	$1/2$
	$\mathbb{P}_{\text{composite}}(g \mid a_1)$	$4/9$	$5/9$	0
	$\mathbb{P}_{\text{composite}}(g \mid a_2)$	0	$5/7$	$2/7$

numbers of people, for instance the number of visitors of a tourist destination during holidays, or the number of people who commute between two cities. Additional methods and auxiliary data are needed to translate estimates of devices to estimates of people [52].

A generic framework has been proposed to organize the production process needed for statistical inference on MNO data [49, 52]. According to this framework, the production process runs through three distinct layers. The bottom layer is called the data- or D-layer and consists of the processing of raw mobile network data, which takes place at the MNO. The processing methods that take place in this layer are dependent on the mobile network technology used. The statistics- or S-layer is the top layer in which the processed mobile phone data is used for statistical purposes. The convergence- or C-layer connects these two layers with processing mobile network data sources into data that can be used for statistical purposes. This intermediate layer is needed since mobile network technology is complex and constantly changing. The output of the C-layer should be a stable source for the S-layer, in which this is used in combination with other data sources to produce statistics.

Our framework takes place in the D-layer, since mobile network data is processed for constructing the connection likelihood. Note that it does not matter which method is used for this process, since all described methods use mobile network data, e.g. the Voronoi method uses cell tower locations. The output of our framework, i.e. the location posterior, belongs in the C-layer, since this does not depend on technology, and hence can be used directly for statistical purposes. Using prior information could be theoretically be placed in the S-layer. Note that the process should ideally be run at the MNO due to potential privacy issues.

6 Validation methods

The methods presented in this paper have been developed within the European Statistical System. They have been used in studies with MNO data in several European countries [58]. The model parameters have been configured for each application in cooperation with radio network specialists from the partner MNOs. Several validation methods have been applied in internal studies to assess how realistic the methods are in practice. Due to confidentiality, the results cannot be shared. In this section we describe these and other validation methods and how they can be applied.

We recommend to validate the methods before applying them, since mobile networks are different from each other in terms of configuration and calibration. In other words, a valid location estimation method for a certain mobile network do not automatically imply validation for another mobile network.

Each of the validation methods described below can also be used for model calibration. After calibration we recommend to validate the methods with other ground truth data.

6.1 Ground truth data

Data that can be considered as ground truth data for the modelled signal strength values $S(g, a)$ are field measurements, which can be taken either from the center of grid tile g, or at various sample points. Note that measurements within one grid tile may have a large variation due to building and tree reflections and due to the possibility of being in- or outside of buildings.

MNOs often use field measurements in combination with their own propagation models for network analysis and optimization. These *propagation data* consist of a signal strength value $S(g, a)$ for each grid tile g and cell a given that g is located in the service area of a. A common by-product derived from these data is the *Best Service Area (BSA)* map, which visualises the cell with the strongest signal strength in each grid tile.

Field measurements can be taken with a multi-band wireless measurement receiver or a smart phone, which has the advantage that it is easier and cheaper to collect large amounts of field measurements data. Smart phone apps such as Network Cell Info [36] measure the signal strength of the connected cell and other nearby cells, as well as their geographic locations.

Face validation of the location posterior can be done with popular events, such as football matches and pop festivals [58]. Ground truth information for this task are the exact geographic areas in which the events took place, and the exact time and date of the events. Less important are the visitor counts, since our aim is to estimate the number of devices, not the number of people (which is for further statistical inference).

GPS data from smart phones can be used as ground truth data to validate the location posterior. For each measurement, the GPS coordinates of the device as well as the identification number of the connected cell are required (and the corresponding Timing Advance value in case Timing Advance data are used in the estimations). Smart phone owners have to opt in to collect field measurement data or GPS data from their devices.

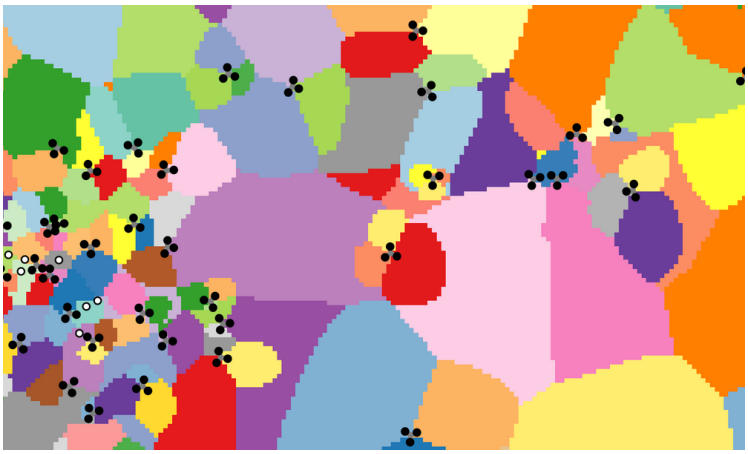


Figure 8: BSA map derived from the signal strength model using fictional cell plan data.

6.2 Signal strength validation

6.2.1 BSA map comparison

A quick and easy validation method is the BSA map comparison. A fictional example of a BSA map is depicted in Fig. 8. Fictional cell plan data is used here, where the black dots indicate the location of directional macro cells, white dots omnidirectional small cells, and the colored areas indicate which cell has the strongest signal strength. This BSA map is the result of the signal strength model. This map can be compared to the BSA map of ground truth data.

Besides a visual inspection of the differences between both maps, the comparison can be quantified, for instance with the percentage of grid tiles that are identical in both maps. Additionally, for each cell a normalized confusion matrix can be calculated from which metrics such as precision and recall can be derived [43].

6.2.2 Comparison of signal strength distributions

The BSA map only shows which cell has the strongest signal strength per grid tile. When ground truth propagation data are available, signal strength values per cell can be compared with those from the signal strength model.

This comparison can be made visually, by creating a heat map where the differences between the ground truth and the modelled signal strength values per cell are mapped to a diverging color scale.

Another approach is to apply a two-dimension distance function between two probability distributions, such as the Kantorovich-Wasserstein Distance [5] (KWD). When a probability distribution is seen as a pile of soil, the KWD is interpreted intuitively as turning one pile of soil into another by moving as little soil as possible.

In order to apply the KWD for signal strength values, these values need to be normalized per cell. This can be done with

$$\mathbb{P}_{S'}(g \mid a) = \frac{S'(g,a)}{\sum_{g' \in \mathcal{G}} S'(g',a)},$$

where $S'(g,a) = \min\big((S(g,a) + 130, -130\big)$ serves as a transformation.

For each cell a we obtain two probability distributions, namely one for the modelled signal strength values and one for the ground truth propagation data, which are compared to each other with the KWD.

The exact KWD distance value is computationally expensive to compute, but approximations exist that can be computed much faster [6]. Open source implementations exist, such as the package `kantorovich` [33] for the programming language `R`.

6.3 Signal dominance validation

In this section we describe a method to validate the signal dominance model that is equivalent to a coin fairness test. To test whether a coin is fair, the coin should be tossed a number of times, the more the better. With standard statistics, it can be stated with a certain confidence whether the coin is fair or not based on the tossed number of heads and tails.

Instead of a coin toss, we consider the situation described in Section 4 in which there are only two cells available for connection with signal strength values S_1 and S_2 and signal dominance values s_{dom1} and s_{dom2}. Recall that we have calculated the probability of connection to the first cell as $s_{\text{dom1}}/(s_{\text{dom1}} + s_{\text{dom2}})$ for a range of signal strength values. These probabilities are shown in Fig. 5 as percentages, which can also be interpreted as the expected number of cases the network chooses the first cell, out of 100 similar cases.

These probabilities can be validated with field measurement data from a set of smart phones during a certain period. For each measurement observation, let $S_1, \ldots S_k$ with $k \geq 2$ be the number of measured cells, where S_1 is the signal strength measured from the connected cell, and the other values are signal strengths from the other nearby cells.

These measurement observations over time are not independent, because in case the network connects a device to a certain cell, it will probably remain connected to this cell if the device does not move and if capacity of this and nearby cells remains the same. In order to keep the observations as independent as possible, we recommend to keep some time between measurements, say one hour. Furthermore, since a device will usually not move during night time, it will likely remain connected to the same cell. Therefore, we recommend to ignore measurements taken during night time.

An observation for which $k > 2$ can be split into $k - 1$ observations, namely $(S_1, S_2), \ldots (S_1, S_k)$. We interpret each observation (S_1, S_i) as if there are only two available cells of which the first one is chosen. In order to obtain enough observations of the same signal strength values should be discretized, e.g. rounded to the values used in Fig. 5: e pairs (S_1, S_2), the signal strength v$-50, -60, \ldots, -130$.

For each pair of discretized signal strength values (S_1, S_2) we count the number of observations in our field measurement data, and compare them to the expected number of observations. This comparison can be done visually, i.e. by creating charts like the one depicted in Fig. 5 for the actual observations and for the expected observations. Another approach is to apply a binomial test for each pair.

6.4 Location posterior validation

We recommend to validate the location posterior after the signal strength and signal dominance model have been validated, because these are used in the calculation of the location posterior.

6.4.1 Face validation with popular events

Popular events, such as football matches or pop festivals, can be used to validate the location posterior [58]. The area in which the events took place are known beforehand, e.g. a football stadium or a pop festival terrain. This area can be compared to the spatial distribution of devices that follows from the location posterior distributions of the logged devices.

For this purpose we require c_a, which is the number of unique devices of which an event has been logged with cell a within a given time interval (e.g. during a football match). Hence we derive a straightforward formula for the estimated number of devices per grid tile:

$$N(g) = \sum_{a \in \mathcal{A}} c_a \mathbb{P}(g \mid a) \tag{25}$$

This spatial density of devices can be plotted on a map in order to see to which extent the peak density areas correspond to the areas in which the events took place. The absolute estimated numbers of devices cannot be used for this analysis, because the number of devices is estimated, not the number of people. What matters in this analysis are the areas in which the peaks are estimated.

When this validation method is used, it is important to know that MNOs often place extra small cells to guarantee mobile communication during crowded events [63, 65]. Therefore it is recommended to make sure that information about these extra small cells is included in the used cell plan data.

6.4.2 Validation with GPS data

The location posterior can be validated with GPS data. For each measurement, the GPS coordinates and the identification of the connected cell are logged. In order to validate the uncertainty of the location posterior it is important to collect as many measurements per cell as possible.

Let $m(g, a)$ be the number of measurements where a device is located in grid tile g (according to GPS) and connected to cell a. The corresponding probability distribution per cell is defined as

$$\mathbb{P}_m(g \mid a) = \frac{m(g, a)}{\sum_{g' \in \mathcal{G}} m(g', a)},$$

We can use the KWD metric to compare this distribution to the posterior distribution $\mathbb{P}(g \mid a)$.

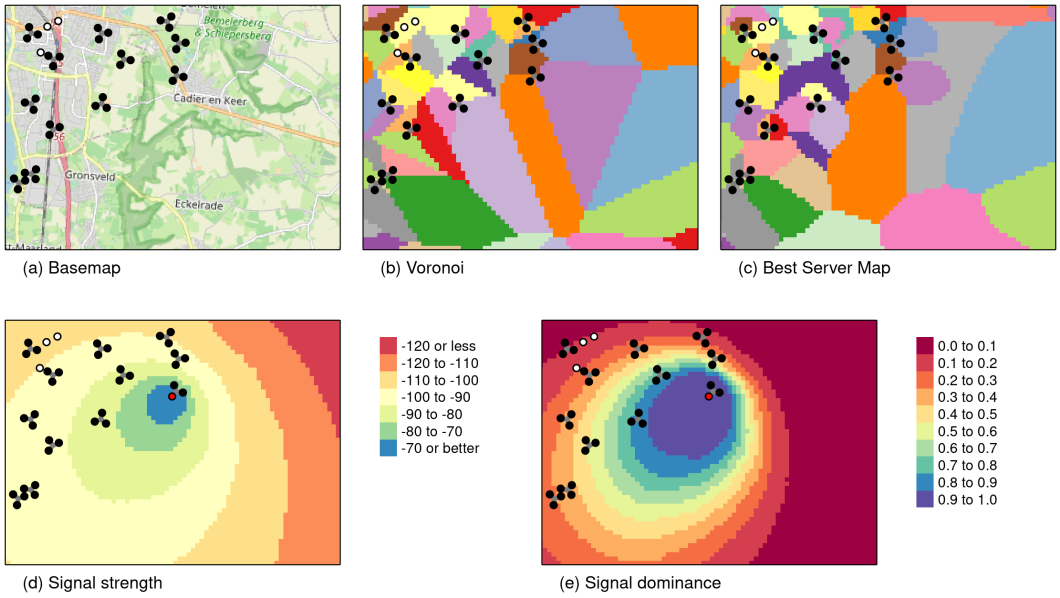


Figure 9: The application of our signal strength model.

7 Application

A fictional example that resembles a real-world situation is illustrated in Fig. 9(a), where a base map is shown with artificially placed cells. The northwestern part of this area is urbanised, while the rest is mostly rural with some small villages. The black dots represent directional macro cells, which are deployed in triplets, with propagation angles about 120 degrees apart from each other. The white dots represent omnidirectional small cells.

The Voronoi map for this example, which we will use as a reference for our signal strength model, is shown in Fig. 9(b). We apply the Voronoi algorithm only to macro cells, where we shift the locations of the Voronoi points towards the propagation directions. Subsequently, the grid tile areas in which the small cells are located are subtracted from the Voronoi tessellation, and assigned to the small cells.

Fig. 9(c) shows the BSA map based on our signal strength model. This map can be used to validate the signal strength model, as discussed in Section 6.2.1

Figs. 9(d) and (e) show the signal strength (in dBm) and signal dominance (s_{dom}) of one specific cell respectively, which is colored red. The propagation direction of this is south-west. Validation of the shown data has been described in Sections 6.2.2 and 6.3.

Fig. 10 shows the results of our modular framework for the same region, shown in (a). The rest of this figure is arranged as a cross table where the location priors are shown as rows in (c), the connection likelihoods as columns in (b), and the location posteriors as table cells in (d). For instance, the bottom right map of this figure is the result of using the composite prior (which consists of half the land use prior and half the network prior) and the signal dominance likelihood. For the network prior we have used signal dominance, i.e. s is set to s_{dom} in Eq. (19).

In the maps shown in Fig. 10(b), (c), and (d), sequential color palettes have been applied where yellow corresponds to relatively low values and respectively dark green, blue, and brown correspond to relatively high values. For instance, the dark blue areas in the land use prior map correspond to the areas with buildings and roads, and therefore, where devices are expected to be. The blue areas in the network prior map highlight the areas that have a high network coverage.

Two different connection likelihoods for the selected cell are shown in Fig. 10(b). Observe that the Voronoi likelihood has the same shape as the Voronoi area of this cell shown in Fig. 9(b) in orange, which follows directly from Eq. (11) which states that a likelihood value equals 1 if and only if the corresponding grid tile is located in the Voronoi area. The high values for the signal dominance likelihood are located in the south/south-west direction of the cell, whereas the propagation direction is south-west. This can be explained by the fact that there is less overlap with other cells in the south direction in comparison to the west direction.

The combination of the location priors and the connection likelihoods results in the location posteriors shown in Fig. 9(d). The land use prior is seen to have a strong effect on the posterior distributions. The network prior places more weight on the areas with better network coverage. Note that, by applying Eq. (20), the combination of network prior and signal dominance likelihood results in a posterior distribution that is a rescaled version of the signal dominance, shown in Fig. 9(e). Finally, the combination of the composite prior and signal dominance likelihood takes into account both land use and network coverage, and would therefore be a good candidate to be used for further statistical inference.

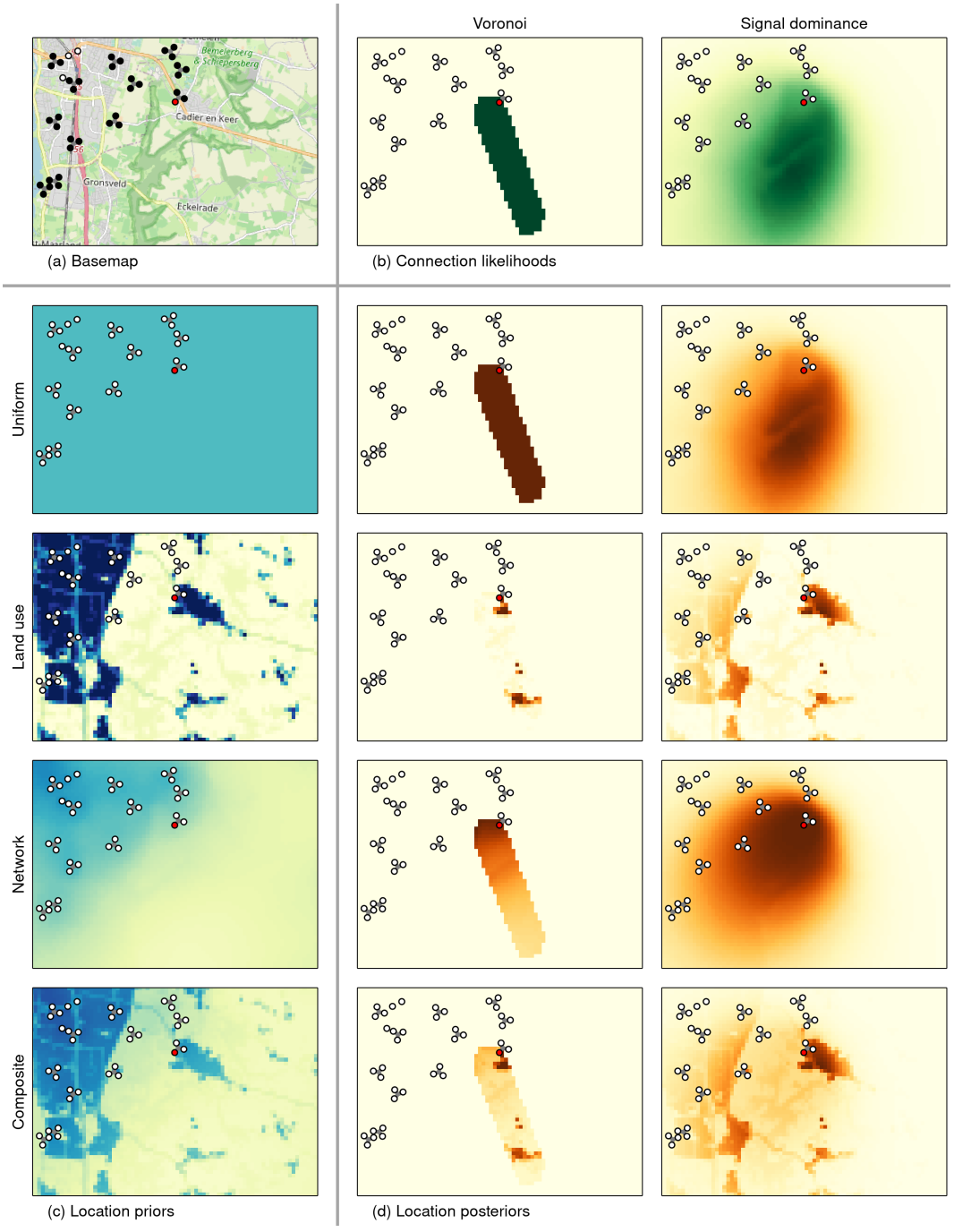
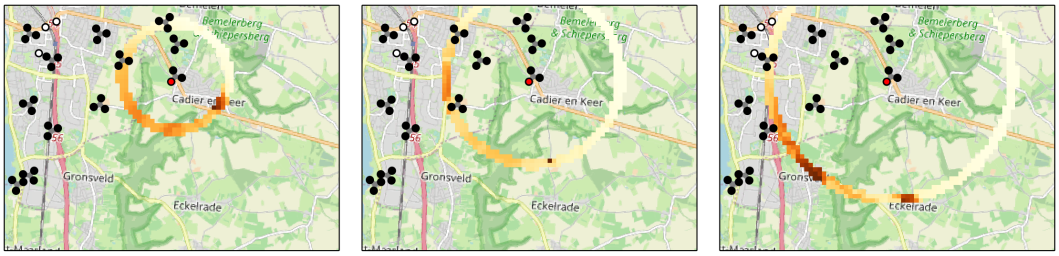


Figure 10: Example showing how three location priors and two connection likelihoods are combined into six location posteriors.



(a) TA value 15 (b) TA value 25 (c) TA value 35

Figure 11: Updated location posterior with Timing Advance.

The final step in our modular framework is to update the location posterior by Timing Advance. Recall from Section 5.3 that Timing Advance is a variable in MNO data from which the distance between the device and the cell can be estimated. Fig. 11 illustrates the location posterior shown in the bottom right map of Fig. 10 (composite prior, signal dominance likelihood) updated with τ values of 15, 25, and 35. The Timing Advance likelihood defined by Eq. (24) has been used with parameter b set to 1, which means three annuli have been merged, in Fig. 11(a) the annuli that correspond to τ values of 14, 15, and 16. In this application, the width of each annulus is 78.12 meter, so this means that only grid tiles whose centroids are between 1094 to 1328 meters away from the selected cell have a positive probability in the updated posterior. For Fig. 11(b) and (c) those distance intervals are [1878, 2109] and [2656, 2890] respectively.

8 Implementation

The methods described in this paper have been implemented in `mobloc` [60] and `mobvis` [61], packages for the programming language `R`. The former package is used for the calculations, the latter for the graphical user interfaces and visualizations. In this section we present a general work flow for `mobloc` and illustrate how `mobvis` is used in this work flow. Details and reproducible examples can be found in the documentation of the packages.

The general work flow of `mobloc` is depicted in Fig. 12. The first step in the work flow is to collect all relevant datasets. The most important dataset is is the cell plan. The geographical locations of the cells is required, whereas other physical properties, such as height, propagation direction, and power, are recommended in order to run the signal strength model. Other datasets that can be used as input are elevation data, land use data, and administrative region boundaries.

The collected data need to be in the correct format in order to be processed. Information about which object classes are supported are described in the package documentation [60]. The preprocessing stage is also used to make sure that the spatial objects have one common Coordinate Reference System (CRS). It is strongly recommended to use a CRS that preserves areas and distances for the region of interest [37]. The function `validate_cellplan` is used to validate the cell plan and to check whether it is consistent with the other input data objects.

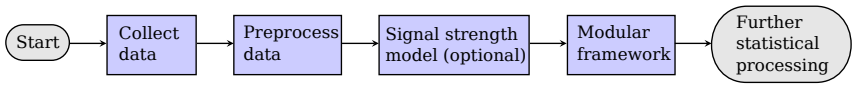


Figure 12: Work flow of `mobloc`.

The third step in the work flow is to run the signal strength model. This step is only required if the network prior is used or if the signal strength model is used in the connection likelihood module. The parameters for the signal strength model can be configured with a graphical user interface started with the function `setup_sig_strength_model` from the `mobvis` package. A screenshot of this tool is shown in Fig. 13. The function `compute_sig_strength` from the `mobloc` package is used to compute the signal strength and the corresponding signal dominance.

[!htb]

Table 3: An overview of `mobloc` functions.

Module	Option	Function
Location prior	Uniform prior	`create_uniform_prior`
	Network prior	`create_network_prior`
	Auxiliary data (e.g. land use) prior	`create_prior`
Connection likelihood	Signal strength	`create_strength_llh`
	Voronoi	`create_voronoi_llh`
Bayesian update	Posterior $\mathbb{P}(g \mid a)$, see Eq. (6)	`calculate_posterior`
	Posterior $\mathbb{P}(g \mid a, \tau)$, updated with Timing Advance, see Eq. (23)	`update_posterior_TA`

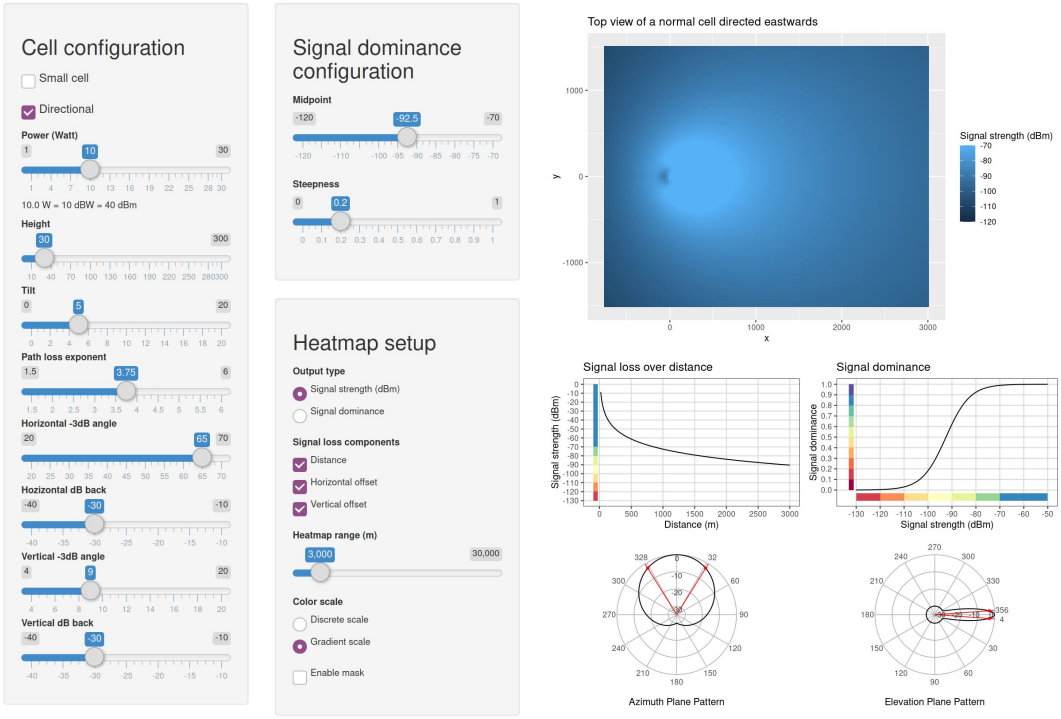


Figure 13: Screenshot of the signal strength model setup tool.

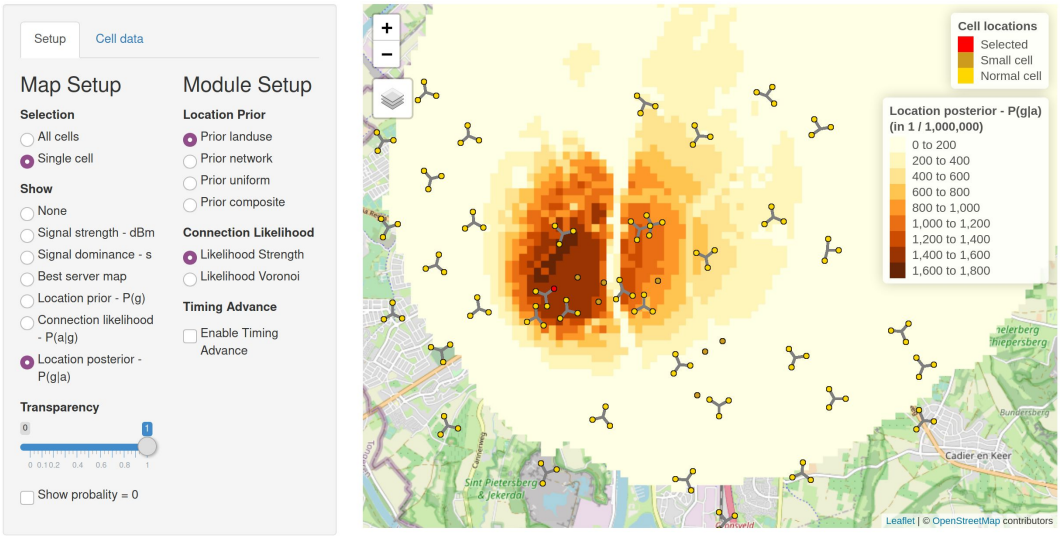


Figure 14: Screenshot of the `mobloc` exploration tool.

The fourth and final step is to run the whole modular framework that is depicted in Fig. 6. The functions that correspond to the modules are listed in Table 3.

The `mobvis` package contain functions to visualize the results. All maps from Figs. 9, 10, and 11 have been created with `mobvis`. This package also contains an interactive tool, started with `explore_mobloc`, in which the results are visualized. Via this tool, it is possible to select different cells, and compare different module options. A screenshot is shown in Fig. 14, in which the same fictional cell plan data is used as described in Section 7.

The output of `mobloc` is a data table with the following variables: cell identification number, tile identification number, optionally a Timing Advance value, and the corresponding posterior probability. This table can be used for further statistical processing and analysis with MNO data [49, 52].

9 Conclusion and discussion

With the work presented in this paper, we have contributed to the emerging research on estimating geographic locations of mobile devices using MNO data. We proposed a signal strength and a signal dominance model to estimate the likelihood probabilities that a device is connected to a cell given its geographic location. These likelihood probabilities have already been applied in further statistical inference [31, 48, 52]. Calibration and validation of the signal strength and dominance model is recommended (see Section 6). This is not a one-off task, but a task that has to be performed for each study or application, since each mobile network is configured differently.

Propagation models have been developed that are much more sophisticated than the model we propose [30, 38, 51], but also require complex technical specifications of the cells

and network measurement data. Although we were unable to test these models due to a lack of data, we assume they are more realistic, because they take the 3D environment into account for modeling the reflection, absorption, and scattering of propagation. In contrast, our signal strength model only takes the type of environment into account via land use. This has the disadvantage that the propagation around specific unusual features of the environment, such as very large buildings, may not be modeled accurately. However, when our signal strength model is used for statistical applications on a larger level than street level, e.g. mobility on neighbourhood level, the accuracy of the signal strength model may be sufficient. This should be confirmed with validation, using the methods described in Section 6.

The advantage of the signal strength model in comparison to the more complex propagation models is that it is relatively simple, and therefore easier to apply, not only because of lower data requirement but also because of lower computational complexity. This is especially important for large scale applications in developing countries [35, 67], where data and computational resources are often limited [58].

We encourage further development of propagation models with lower data requirements. Other cell plan data variables can be helpful to improve our model, in particular the radio frequency bands at which cells are operating, which have a great impact on phenomena such as reflection, absorption, and scattering of the signal [56]. So far, our signal strength model only uses land use to approximate the path loss exponent. 3D models of the environment can developed to take these phenomena into account [38].

We developed the signal dominance model on assumptions about load balancing which we described in Section 4. As with the signal strength model, calibration and validation of the signal dominance model is recommended for each new study or application, since each cellular network is configured differently.

Further research is needed on how MNOs manage handovers between networks of different generations and operating frequency bands. We currently assume that a mobile device only connects to cells from one network. When this assumption holds, the methods can be applied independently for each network. In reality however, this assumption might not hold for reasons such as coverage gaps, capacity limits, and network issues. More research is also needed on how MNOs handle roaming in cross-border regions and how this can be modelled.

We have explored several priors. We argued in Section 5.2 that each of the non-uniform priors has predictive value about where devices are located. Therefore we would recommend each of them over the uniform prior when the likelihood does not use any empirical data; the likelihood proposed by [41] uses empirical data, namely aggregated Timing Advance data about the distribution of past connections. Such data already contain information about where devices are expected to be, and therefore we agree with [41] that a uniform prior is recommended in that case.

When a non-uniform prior is used, it is important to assess the quality of the used prior data. Land use data may not always be of good quality or up-to-date, which may cause biased estimates. The same holds for the network prior; if an indicator is used that does not reflect how well the connection is, for instance when a wrongly calibrated signal dominance model is used, then estimates may be biased.

We have illustrated that the Bayesian framework can be extended with the use of additional network measurement data in Section 5.3, in this case Timing Advance data. Since these data are accurate measurements required for mobile communication, it is expected

that the updated posterior is an improvement of the original posterior. Fig. 11 illustrates how substantial this improvement can be in practice. Validation is still recommended to make sure there are no errors in the estimation process.

The Bayesian estimation method is used to estimate the location of a single device. However, the method can also be used to estimate the spatial distribution of all connected devices, as shown in Eq. (25). Two alternative methods are MLE and DF as described in Section 2.4).

The main advantage of the MLE and DF methods is that they extract extra information that becomes available from the number of devices that have been counted per cell. According a simulation study these methods perform better than the Bayesian approach when using a uniform priorËIJ [48]. More specifically, the population density estimates from these methods are more similar to the synthetic ground truth population than the estimated densities from the Bayesian approach using Eq. (25).

The main advantage of the Bayesian approach is that it is easier to understand (i.e. better explainable) and much faster to compute. Furthermore, the MLE and DF rely more on the likelihood probabilities than the Bayesian approach does. In situations where prior data are available that have good predictive value, or when empirical data from past connections [41] are available, the Bayesian approach may be preferable. More research is required to compare the approaches in different scenarios.

The Bayesian approach for spatial density estimation, as well as the MLE and DF methods, only use event data logged at a certain time. However, when we consider the dynamic behaviour of a device and take into account the sequence of logged events, much more information can be extracted by interrelating the corresponding posterior probabilities, for instance by using a Hidden Markov Model [52]. Such methods can be used to estimate trajectories, which is a key step in statistical inference on dynamic applications, such as urban mobility, but also to improve spatial density estimations.

More research is needed to validate and compare the location estimation methods with real-world data. However, access to MNO data is still a major bottleneck for statistical inference [57]. As an alternative, we encourage further development of simulated event data [40].

Funding

This work has been carried out within the projects ESSnet Big Data I and II of the European statistical system (ESS). The Grant Agreement Numbers for those projects are 11104.2016.010-2016.756 and 847375-2018-NL-BIGDATA and respectively.

Acknowledgments

First we would like to thank Sander Scholtus, Marco Puts, and Shan Shah, who contributed to the foundations of the methodology described in this paper. Also many thanks to all ESSnet Big Data project members, in particular Fabio Ricciato, David Salgado, Benjamin Sakarovitch, Roberta Radini, Tiziana Tuoto, and Sandra Hadam. Finally, we would like to thank those who helped improving the manuscript: Harm Jan Boonstra, Ralph Meijers, Sigrid van Hoek, May Offermans, Edwin de Jonge, Jan van der Laan, Marc Ponsen, Jacob Wilkins, and Christine Gutekunst.

References

[1] 3GPP. TS Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, 3 2019.

[2] AHAS, R., AASA, A., YUAN, Y., RAUBAL, M., SMOREDA, Z., LIU, Y., ZIEMLICKI, C., TIRU, M., AND 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 (2015), 2017–2039.

[3] ALEXANDER, L., JIANG, S., MURGA, M., AND GONZALEZ, M. C. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies 58* (2015), 240 –250. doi:10.1016/j.trc.2015.02.018.

[4] BADR, H., DU, H., MARSHALL, M., DONG, E., SQUIRE, M., AND GARDNER, L. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases 20*, 11 (Nov. 2020).

[5] BASSETTI, F., GUALANDI, S., AND VENERONI, M. On the computation of Kantorovich-Wasserstein distances between 2D-histograms by uncapacitated minimum cost flows, 2020.

[6] BASSETTI, F., GUALANDI, S., AND VENERONI, M. On the computation of kantorovich-wasserstein distances between two-dimensional histograms by uncapacitated minimum cost flows. *SIAM J. Optim. 30*, 3 (2020), 2441–2469. doi:10.1137/19M1261195.

[7] BISWAS, S., GUPTA, A., AND CHAKRABORTY, S. Load-balanced user associations in dense lte networks. *Computer Networks 189* (2021), 107928. doi:10.1016/j.comnet.2021.107928.

[8] CACERES, N., ROMERO, L. M., AND BENITEZ, F. G. Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys. *Transportmetrica A: Transport Science 16*, 3 (2020), 574–601. doi:10.1080/23249935.2020.1720857.

[9] CALABRESE, F., FERRARI, L., AND BLONDEL, V. Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys 47*, 2, 1–20.

[10] CHEN, B. Y., WANG, Y., WANG, D., LI, Q., LAM, W. H. K., AND SHAW, S.-L. Understanding the impacts of human mobility on accessibility using massive mobile phone tracking data. *Annals of the American Association of Geographers 108*, 4 (2018), 1115–1133. doi:10.1080/24694452.2017.1411244.

[11] DE MEERSMAN, F., SEYNAEVE, G., DEBUSSCHERE, M., LUSYNE, P., DEWITTE, P., BAEYENS, Y., WIRTHMANN, A., DEMUNTER, C., REIS, F., AND REUTER, H. Assessing the quality of mobile phone data as a source of statistics. In *European Conference on Quality in Official Statistics* (2016), Eurostat.

[12] DEVILLE, P., LINARD, C., MARTIN, S., GILBERT, M., STEVENS, F. R., GAUGHAN, A. E., BLONDEL, V. D., AND TATEM, A. J. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*, 45 (2014), 15888–15893. doi:10.1073/pnas.1408439111.

[13] DIAO, M., ZHU, Y., JOSEPH FERREIRA, J., AND RATTI, C. Inferring individual daily activities from mobile phone traces: A boston example. *Environment and Planning B: Planning and Design 43*, 5 (2016), 920–940. doi:10.1177/0265813515600896.

[14] DUJARDIN, S., JACQUES, D., STEELE, J., AND LINARD, C. Mobile phone data for urban climate change adaptation: Reviewing applications, opportunities and key challenges. *Sustainability 12*, 4 (02 2020), 1501.

[15] ESTIMA, J., AND PAINHO, M. *Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal*. 03 2015, pp. 273–293.

[16] FIGUEIRAS, J. A., AND FRATTASI, S. *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques*. John Wiley and Sons, Ltd., 2010.

[17] FONTE, C., PATRIARCA, J., MINGHINI, M., ANTONIOU, V., SEE, L., AND BROVELLI, M. *Using OpenStreetMap to Create Land Use and Land Cover Maps: Development of an Application*. 2017, p. 25.

[18] GRAELLS-GARRIDO, E., PEREDO, O. F., AND GARCÍA, J. Sensing urban patterns with antenna mappings: The case of Santiago, Chile. *Sensors 16*, 7 (2016), 1098.

[19] GRANTZ, K., MEREDITH, H., CUMMINGS, D., METCALF, C., GRENFELL, B., GILES, J., MEHTA, S., SOLOMON, S., LABRIQUE, A., KISHORE, N., BUCKEE, C., AND WESOLOWSKI, A. The use of mobile phone data to inform analysis of covid-19 pandemic epidemiology. *Nature Communications 11*, 1 (Dec. 2020).

[20] GU, Z., ZHANG, Y., CHEN, Y., AND CHANG, X. Analysis of attraction features of tourism destinations in a mega-city based on check-in data mining—a case study of Shenzhen, China. *ISPRS International Journal of Geo-Information 5*, 11 (2016).

[21] IQBAL, M. S., CHOUDHURY, C., WANG, P., AND GONZALEZ, M. C. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies 40* (03 2014), 63–74. doi:10.1016/j.trc.2014.01.002.

[22] JÄRV, O., TENKANEN, H., AND TOIVONEN, T. Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science 31*, 8 (2017), 1630–1651.

[23] JIANG, S., YANG, Y., GUPTA, S., VENEZIANO, D., ATHAVALE, S., AND GONZÁLEZ, M. C. The timegeo modeling framework for urban motility without travel surveys. *Proceedings of the National Academy of Sciences 113*, 37 (2016).

[24] KHALAF-ALLAH, M., AND KYAMAKYA, K. Bayesian mobile location in cellular networks. In *2006 14th European Signal Processing Conference* (2006), pp. 1–5.

[25] KHALAF-ALLAH, M., AND KYAMAKYA, K. Mobile location in gsm networks using database correlation with bayesian estimation. In *11th IEEE Symposium on Computers and Communications (ISCC'06)* (2006), pp. 289–293. doi:10.1109/ISCC.2006.103.

[26] KONDOR, D., GRAUWIN, S., KALLUS, Z., GÓDOR, I., SOBOLEVSKY, S., AND RATTI, C. Prediction limits of mobile phone activity modelling. *Royal Society open science 4*, 2 (2017).

[27] KORA, A. D., ELONO ONGBWA, B. A., CANCES, J.-P., AND MEGHDADI, V. Accurate radio coverage assessment methods investigation for 3G/4G networks. *Computer Networks 107*, P2 (Oct. 2016), 246–257.

[28] KREHER, R., AND GAENGER, K. *LTE Signaling, Troubleshooting and Optimization*, 1 ed. John Wiley and Sons, Ltd., 2011.

[29] KUNG, K., SOBOLEVSKY, S., AND RATTI, C. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one 9* (11 2013).

[30] KYÖSTI, P., LEHTOMÄKI, J., MEDBO, J., AND LATVA-AHO, M. Map-based channel model for evaluation of 5G wireless communication systems. *IEEE Transactions on Antennas and Propagation 65*, 12 (2017), 6491–6504. doi:10.1109/TAP.2017.2754443.

[31] LAAN, D. V. D., AND JONGE, E. D. Maximum likelihood reconstruction of population densities from mobile signalling data. In *Proceedings of the NetMob 2019 Conference* (2019).

[32] LAI, S., ERBACH-SCHOENBERG, E., PEZZULO, C., RUKTANONCHAI, N., SORICHETTA, A., STEELE, J., LI, T., DOOLEY, C., AND TATEM, A. Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications 5*, 34 (2019).

[33] LAURENT, S. *kantorovich: Kantorovich Distance Between Probability Measures*, 2020. R package version 3.0.0.

[34] LIAO, YUAN, YEH, SONIA, AND JEUKEN, GUSTAVO S. From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. *EPJ Data Sci. 8*, 1 (2019), 34. doi:10.1140/epjds/s13688-019-0212-x.

[35] LU, X., WRATHALL, D. J., SUNDSŽY, P. R., NADIRUZZAMAN, M., WETTER, E., IQBAL, A., QURESHI, T., TATEM, A., CANRIGHT, G., ENGA-MONSEN, K., AND BENGTSSON, L. Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in bangladesh. *Global Environmental Change 38* (2016), 1–7. doi:10.1016/j.gloenvcha.2016.02.002.

[36] M2CATALYST, LLC. Network Cell Info app, 2020.

[37] MALING, D. H. *Coordinate Systems and Map Projections.* Oxford: Pergamon Press., 1992.

[38] MEDBO, J., KYOSTI, P., KUSUME, K., RASCHKOWSKI, L., HANEDA, K., JAMSA, T., NURMELA, V., ROIVAINEN, A., AND MEINILA, J. Radio propagation modeling for 5G mobile and wireless communications. *IEEE Communications Magazine 54*, 6 (2016), 144–151. doi:10.1109/MCOM.2016.7498102.

[39] NABOULSI, D., FIORE, M., RIBOT, S., AND STANICA, R. Large-scale mobile traffic analysis: A survey. *IEEE Commun. Surv. Tutorials 18*, 1 (2016), 124–161. doi:10.1109/COMST.2015.2491361.

[40] OANCEA, B., NECULA, M., SANGUIAO, L., SALGADO, D., AND BARRAGÁN, S. A simulator for network event data, December 2019. ESSnet Big Data II - Deliverable I.2. Available online from https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPI_Milestones_and_deliverables.

[41] OGULENKO, A., BENENSON, I., OMER, I., AND ALON, B. Probabilistic positioning in mobile phone network and its consequences for the privacy of mobility data. *Computers Environment and Urban Systems 85* (10 2020), 101550. doi:10.1016/j.compenvurbsys.2020.101550.

[42] OLIVER, N., LEPRI, B., STERLY, H., LAMBIOTTE, R., DELETAILLE, S., DE NADAI, M., LETOUZÉ, E., SALAH, A. A., BENJAMINS, R., CATTUTO, C., COLIZZA, V., DE CORDES, N., FRAIBERGER, S. P., KOEBE, T., LEHMANN, S., MURILLO, J., PENTLAND, A., PHAM, P. N., PIVETTA, F., SARAMÄKI, J., SCARPINO, S. V., TIZZONI, M., VERHULST, S., AND VINCK, P. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Science Advances 6*, 23 (2020).

[43] OLSON, D. L., AND DELEN, D. *Advanced Data Mining Techniques.* Springer, 2008.

[44] PANWAR, N., SHARMA, S., AND SINGH, A. K. A survey on 5G: The next generation of mobile communication. *Physical Communication 18* (2016), 64–84. Special Issue on Radio Access Network Architectures and Resource Management for 5G.

[45] PUCCI, P., MANFREDINI, F., AND TAGLIOLATO, P. *Mapping urban practices through mobile phone data*. PoliMI SpringerBriefs Series, 02 2015.

[46] PUTS, M., DAAS, P., TENNEKES, M., AND DE BLOIS, C. Using huge amounts of road sensor data for official statistics. *AIMS Mathematics 4*, 1 (2019), 12–25.

[47] RAITOHARJU, M., ALI-LÖYTTY, S., AND WIROLA, L. Estimation of base station position using timing advance measurements. In *Proceedings of SPIE – The International Society for Optical Engineering* (12 2010), vol. 8285.

[48] RICCIATO, F., AND COLUCCIA, A. On the estimation of spatial density from mobile network operator data. *IEEE Transactions on Mobile Computing* (2021), 1. doi:10.1109/TMC.2021.3134561.

[49] RICCIATO, F., LANZIERI, G., WIRTHMANN, A., AND SEYNAEVE, G. Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing 6* (2020), 101263.

[50] RICCIATO, F., WIDHALM, P., CRAGLIA, M., AND PANTISANO, F. 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*, February 2017 (2017), 65–82.

[51] SALEM, Y., AND IVANEK, L. Propagation modelling of path loss models for wireless communication in urban and rural environments at 1800 gsm frequency band. *Advances in Electrical and Electronic Engineering 14* (06 2016). doi:10.15598/aeee.v14i2.1586.

[52] SALGADO, D., SANGUIAO, L., OANCEA, B., BARRAGÁN, S., AND NECULA, M. An end-to-end statistical process with mobile network data for official statistics. *EPJ Data Science 10*, 20 (2021).

[53] SCHULTZ, M., VOSS, J., AUER, M., CARTER, S., AND ZIPF, A. Open land cover from openstreetmap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation 63* (07 2017), 206–213.

[54] SHARMA, A., ROY, A., GHOSAL, S., CHAKI, R., AND BHATTACHARYA, U. Load balancing in cellular network: A review. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT 2012)* (Los Alamitos, CA, USA, 2012), IEEE Computer Society, pp. 1–5. doi:10.1109/ICCCNT.2012.6395927.

[55] SRINIVASA, S., AND HAENGGI, M. Path loss exponent estimation in large wireless networks. In *2009 Information Theory and Applications Workshop* (Feb 2009), pp. 124–129.

[56] STUTZMAN, W., AND THIELE, G. *Antenna Theory and Design, 3rd Edition*. Antenna Theory and Design. Wiley, 2012.

[57] SUAREZ-CASTILLO, M., AND AL. Access to mobile network data: an updated review, April 2021. ESSnet Big Data II - Deliverable I.1. Available online from https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPI_Milestones_and_deliverables.

[58] SUAREZ-CASTILLO, M., AND AL. Experience with real data - some experimental results with mobile network data, January 2021. ESSnet Big Data II - Deliverable I.7. Available online from https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPI_Milestones_and_deliverables.

[59] TENNEKES, M. Statistical inference on mobile phone network data. Presentation at European Forum for Geography and Statistics (EFGS 2018) https://tinyurl.com/he4a95vn, 2018.

[60] TENNEKES, M. mobloc: Mobile phone location algorithms and tools. R package. Available at https://github.com/MobilePhoneESSnetBigData/mobloc, 2020.

[61] TENNEKES, M. mobvis: Visualization of mobile phone location algorithm results. R package. Available at https://github.com/MobilePhoneESSnetBigData/mobvis, 2020.

[62] TENNEKES, M., GOOTZEN, Y., AND SHAH, S. H. A Bayesian approach to location estimation of mobile devices from mobile network operator data. CBDS Working paper 06-20 https://tinyurl.com/8684j6du, 2020.

[63] TOLSTRUP, M. *Indoor Radio Planning: A Practical Guide for 2G, 3G and 4G.* Wiley, 2015.

[64] TU, W., CAO, J., YUE, Y., SHAW, S.-L., ZHOU, M., WANG, Z., CHANG, X., XU, Y., AND LI, Q. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science 31*, 12 (2017), 2331–2358. doi:10.1080/13658816.2017.1356464.

[65] WANG, S., ZHAO, W., AND WANG, C. Budgeted cell planning for cellular networks with small cells. *Vehicular Technology, IEEE Transactions on 64* (10 2015), 4797–4806.

[66] WIDHALM, P., YANG, Y., ULM, M., ATHAVALE, S., AND GONZÁLEZ, M. C. Discovering urban activity patterns in cell phone data. *Transportation 42*, 4 (2015), 597–623.

[67] WILSON, R., ZU ERBACH-SCHOENBERG, E., ALBERT, M., POWER, D., TUDGE, S., GONZALEZ, M., GUTHRIE, S., CHAMBERLAIN, H., BROOKS, C., HUGHES, C., PITONAKOVA, L., BUCKEE, C., LU, X., WETTER, E., TATEM, A., AND BENGTSSON, L. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake. *PLOS Currents Disasters* (Feb 2016).

[68] XU, Y., BELYI, A., BOJIC, I., AND RATTI, C. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems 72* (2018), 51–67.

[69] XU, Y., LI, X., SHAW, S.-L., LU, F., YIN, L., AND CHEN, B. Y. Effects of data preprocessing methods on addressing location uncertainty in mobile signaling data. *Annals of the American Association of Geographers 111*, 2 (2021), 515–539. doi:10.1080/24694452.2020.1773232.

[70] ZAGATTI, G. A., GONZALEZ, M., AVNER, P., LOZANO-GRACIA, N., BROOKS, C. J., ALBERT, M., GRAY, J., ANTOS, S. E., BURCI, P., ZU ERBACH-SCHOENBERG, E., TATEM, A. J., WETTER, E., AND BENGTSSON, L. A trip to work: estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. *Development Engineering 3* (2018), 133–165.

[71] ZANG, H., BACCELLI, F., AND BOLOT, J. Bayesian inference for localization in cellular networks. In *Proceedings of the 29th Conference on Information Communications* (2010), INFOCOM'10, IEEE Press, pp. 1963–1971.

[72] ZHAO, Z., SHAW, S.-L., XU, Y., LU, F., CHEN, J., AND YIN, L. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science 30*, 9 (2016), 1738–1762. doi:10.1080/13658816.2015.1137298.