A Bayesian approach to location estimation of mobile devices from mobile network operator data

Martijn Tennekes, Yvonne A.P.M. Gootzen, Shan H. Shah*

Version d.d. 2019-06-27 (draft)

Contents

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

1 Introduction

Mobile phone network data have shown to be a rich potential source for official statistics, in particular on daytime population (Deville et al. 2014; Ahas et al. 2015; De Meersman et al. 2016; Kondor et al. 2017; Xu et al. 2018; Salgado et al. 2018), mobility (Jonge, Pelt, and Roos 2012; S. Kung, Sobolevsky, and Ratti 2013; Iqbal et al. 2014; Alexander et al. 2015; Diao et al. 2016; Pucci, Manfredini, and Tagliolato 2015; Widhalm et al. 2015; Jiang et al. 2016; Zagatti et al. 2018), migration (Lu et al. 2016; Wilson et al. 2016), and tourism (Deville et al. 2014; Tennekes, Offermans, and Heerschap 2017). These data are generated by the cellular network owned by a mobile network operator (MNO), and are primarily used for billing customers and for network analysis.

Mobile phone network data that are used to calculate the costs in order to bill customers are called *Call Detail Records* (CDR). A record in these data corresponds to active mobile phone usage by initiating or receiving a call, or by sending or receiving an SMS. Data that includes events on mobile data usage are often called *Data Detail Records* (DDR). Mobile phone network data that does not only include events triggered by active mobile phone use, but also passive events, such as location updates, is called *signalling data*.

A mobile communication network is also called a *cellular network*, where each *cell* enables mobile communication for a specific land area. Two types of cells can be distinguished (Panwar, Sharma, and Singh 2016): a cell that placed in a cell tower or on top of a roof, which has a theoretical range of 30 kilometers, and a cell that is used to enable mobile communication inside buildings and in dense urban areas, and has a theoretical range of two kilometers[1]. Furthermore, a cell can be directional or omnidirectional. In the networks we studied, cells with a large range tend to be directional,

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

often covering an angle of about 120 degrees, while small cells tend to be omnidirectional.

For statistical inference on mobile phone network data, geographic location is one of the most important variables of the data. However, in many cases the exact geographic location is neither measured nor stored. Only the identification number of the serving cell is required for billing costumers. There exist advanced geographic pinpointing techniques such as the usage of timing advance and the received signal strength (Calabrese, Ferrari, and Blondel 2014), but the necessary data to apply these techniques are not always available.

The majority of studies on mobile network data use Voronoi tessellation (Deville et al. 2014) to distribute the geographic location of logged events. The geographic area is divided into Voronoi regions such that each Voronoi region corresponds to the geographic location of a cell tower and each point in that region is closer to that cell tower than to any other cell tower.

There are a couple of downsides of using Voronoi tessellation to estimate the geographic location of devices. First of all, it assumes that all cells are omnidirectional. As described above, large range cells are often directional. The second downside of Voronoi tessellation is that it does not take other cell properties into account, such power, height, and tilt. Third, the coverage areas of cells overlap in reality, especially in urban areas. This is because of load balancing; if a cell has reached full capacity, neighbouring cells that also have coverage are able to take over communication with mobile devices. This means that a mobile phone is not always connected to the nearest cell nor to the cell with best signal. In urban areas, a mobile phone switches frequently between cells[2]. The fourth and last downside of using Voronoi tessellation is that it does not take into account where people are expected to be.

A couple of variations of the Voronoi algorithm have been proposed to overcome some of these limitations (**ricciato2017**; Graells-Garrido, Peredo, and García 2016; De Meersman et al. 2016). One improvement is to shift the locations of the Voronoi points from the cell tower locations towards the direction of propagation. Alternatively, when the *coverage area* is known for each cell, i.e. the area which is served by the cell, the location of the centroids of these coverage areas can be used as Voronoi points. Another improvement is to create a Voronoi tessellation for cells with a large range,

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

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 (Järv, Tenkanen, and Toivonen 2017).

We propose a modular Bayesian framework to estimate the geographic location of devices, which consists of two main modules, namely the *location prior* and the *connection likelihood*. The former consists of a priori information about where devices are expected to be, and the latter uses network information to estimate the location of the devices. The model is generic in the sense that various data sources and methods can be used for both of these modules.

In this paper, we propose several options for the location prior. The most important one is the usage of land use data. This is arguably the most straightforward option, since more devices are expected to be in certain land use categories, such as urban areas, than in other land use categories, for instance grasslands. However, any data source that contains information about where devices are expected to be can be used.

The connection likelihood describes the estimated probability that a device is connected to a certain cell given its actual location and given a connection to some cell, taking potential overlap between cell coverage areas into account. For this component, we propose a signal strength model which models the propagation per cell using physical properties of the cells, such as height, direction, and tilt. However, the Voronoi method and each of the aforementioned variations can also be used here.

The Bayesian framework can be extended iteratively, which we illustrate with the incorporation of timing advance, a variable from which the distance between a mobile device and its serving cell can be estimated.

The methods in this paper are implemented in the `R` package **mobloc**, which offers interactive dashboard tools to model the propagation and to explore all the results of our approach. These tools can also be used without knowledge of the `R` programming language.

The outline of the paper is as follows. In Section 2 we describe our modular Bayesian framework and illustrate it with a small example. We introduce the signal strength model that can be used for the connection likelihood in Section 3. In Section 5 we describe the implementation of the **mobloc** package. We conclude with a couple of remarks in Section 6

This section describes the propagation of signal strength originating from a single antenna. We distinguish two types of antennas: omnidirectional

and directional, resulting in two propagation models. Omnidirectional antennas have no aimed beam and their coverage can be thought of as a circle. Directional antennas point in a certain direction and their coverage can be thought of as an oval with one axis of symmetry. In practice, small cells are omnidirectional and normal antennas (i.e. attached to cell towers or placed on rooftops) are directional (Kora et al. 2016).

2 A modular Bayesian framework

In the proposed method, 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 is easier since all grid tiles have the same area. The following three aspects are important for the specification of the grid.

The first aspect for specifying a grid is that it is important to use a map projection in which area measurements are preserved (Tyner 2010) in order the ensure a constant grid tile area. For instance, appropriate map projections for the Netherlands are EPSG 28992 and 3035, respectively known as the Dutch National Grid and the Lambert Azimuthal Equal-Area projections (*EPSG* 2018).

Second, it is important to specify the size of grid tiles. Generally, the smaller the better, but this may introduce computational issues. We recommend 100 by 100 meter grid tiles for applications on city or country level. The optimal choice finds a balance between computational costs and the size of the region of interest. An advantage of the proposed approach is that most computations can be parallelized.

The third and last aspect that is important for the specification of a grid is that the grid should be large enough such that it contains the whole area served by the mobile phone network. Note that this may exceed the region of interest. For instance, the cells in the Netherlands also have coverage a couple of kilometers across the borders in Belgium and Germany. In this case, we would use a bounding box around the Netherlands and extend it by 10 kilometers in all four directions.

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

$$\mathbb{P}(g \mid a) \propto \mathbb{P}(g)\mathbb{P}(a \mid g), \tag{1}$$

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.

On a technical note, the result of Equation (1) can be interpreted both as a likelihood and as a probability distribution. When the cell a is given (reading left to right), we see $\mathbb{P}(g \mid a)$ as a likelihood. Whereas when a is merely the notion of an unknown cell (reading from right to left), we view $\mathbb{P}(g \mid a)$ as a distribution. Though this distinction is not critical for the understanding of our model, the intended interpretation should be clear from context. For simplicity, we will refer to $\mathbb{P}(g \mid a)$ as the connection likelihood henceforth.

2.1 Location prior

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

$$\sum_{g \in \mathcal{G}} \mathbb{P}(g) = 1, \tag{2}$$

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 prior.

2.1.1 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:

$$\mathbb{P}_{\text{uniform}}(g) := \frac{1}{|\mathcal{G}|}. \tag{3}$$

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

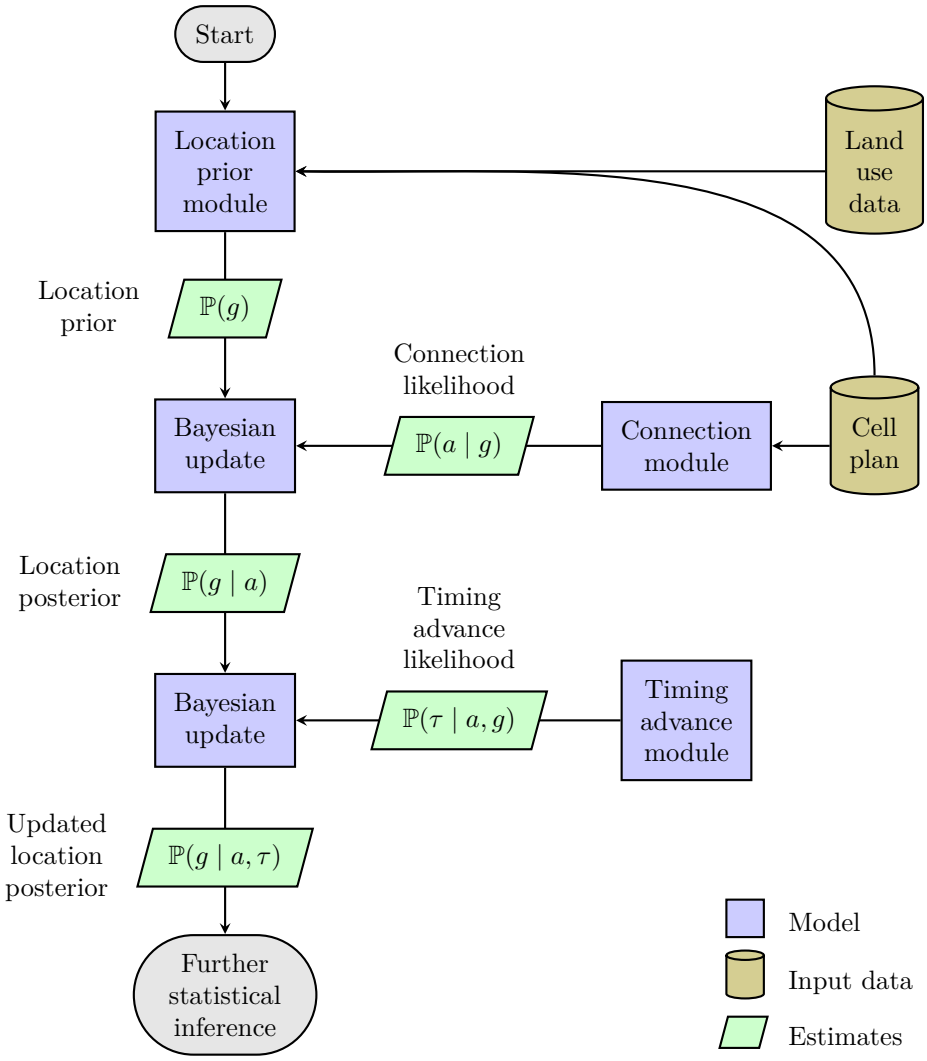

Figure 1: A modular framework for modelling location posteriors.

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.

2.1.2 Land use prior

An alternative is to use administrative sources on land use for the location prior. One would expect more devices in an urban area than in a meadow. Hence, our second prior is called the land use prior and it is based on a proportional expectation of the number of devices $n(g)$ in grid tile g such that:

$$n(g) \propto \mathbb{E}[\text{number of devices in } g] \quad \text{for all } g \in \mathcal{G}. \tag{4}$$

The land use prior is then defined as:

$$\mathbb{P}_{\text{land use}}(g) := \frac{n(g)}{\sum_{g' \in \mathcal{G}} n(g')}. \tag{5}$$

Due to the normalisation in its definition, the land use prior does 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

$$\sum_{k=1,\ldots,K} w_k(g) = 1 \quad \text{for all } g \in \mathcal{G}. \tag{6}$$

Then $n(g)$ can be modelled as

$$n(g) := \sum_{k=1,\ldots,K} u_k \cdot w_k(g). \tag{7}$$

An example of a simple land use classification and the relative expected numbers is shown in Table 1.

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 that serves most of its time as a quiet meadow and suddenly contains more devices than usual is not included in land use data.

Land use class	u_k
Urban	1.0
Main roads	0.5
Other land	0.1
Water	0.0

Table 1: An example of land use classes and their relative expected number of devices.

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 amounts of devices (S. Wang, Zhao, and C. Wang 2015; Tolstrup 2015).

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

2.1.3 Network prior

The third proposed location prior uses the total signal strength across the network and is called the network prior:

$$\mathbb{P}_{\text{network}}(g) := \frac{\sum_{a \in \mathcal{A}} s_{\text{strength}}(g, a)}{\sum_{a \in \mathcal{A}} \sum_{g' \in \mathcal{G}} s_{\text{strength}}(g', a)}, \tag{8}$$

where $s_{\text{strength}}(g, a)$ is a measure that reflects the signal strength received from cell a in grid tile g. More specifically, we will introduce the term *signal dominance* for $s(g, a)$ in Section 2.2, where we will use it for the connection likelihood. For Equation (8), we use the specific instance of $s(g, a)$, namely s_{strength}, resulting from our signal strength model, which we will describe in Section 3. Basically, the network prior reflects the distribution of the total signal over all the grid tiles.

We propose the network prior, because it 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) we use the signal dominance, in 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 also may place extra overlapping cells in order to provide network coverage for specific patches of land, which implies that some parts of land with already good coverage will have an improved network coverage, whereas the expected number of devices does not change. In summary, the total signal strength of the network does not always reflect the estimated number of devices.

Note that when the network prior is based on Voronoi signal strengths, i.e. as we will define in Section 2.2, the prior probability of a device being located in a grid tile is negatively proportional to the total coverage area of the cell that is covering the grid tile.

From a Bayesian perspective, it may seem odd to use the same input, i.e. the signal strength, in both the connection likelihood distribution and the location prior. However, we use it in two different, complementary, ways. This will be illustrated in the example in Section 2.2.2.

Substituting $\mathbb{P}(g)$ in Equation (1) by Equation (8), simplifies to:

$$\mathbb{P}(g \mid a) = \frac{s(g,a)}{\sum_{g' \in \mathcal{G}} s(g',a)}. \tag{9}$$

2.1.4 Composite prior

Our fourth and final prior is less theoretically substantiated and more of a practical approach. It is a combination of any of the above priors:

$$\begin{aligned} \mathbb{P}_{\text{composite}}(g) := \; & \pi_{\text{uniform}} \cdot \mathbb{P}_{\text{uniform}}(g) + \\ & \pi_{\text{land use}} \cdot \mathbb{P}_{\text{land use}}(g) + \\ & \pi_{\text{network}} \cdot \mathbb{P}_{\text{network}}(g), \end{aligned} \tag{10}$$

where π represents the contribution of a prior to the final prior such that:

$$\begin{cases} 0 \leq \pi_{\text{uniform}} \leq 1 \\ 0 \leq \pi_{\text{land use}} \leq 1 \\ 0 \leq \pi_{\text{network}} \leq 1 \\ \pi_{\text{uniform}} + \pi_{\text{land use}} + \pi_{\text{network}} = 1. \end{cases} \tag{11}$$

This approach comes with a mix of advantages from all priors, as well as a mix of disadvantages from all priors. The loss of theoretical understanding in the combination prior can be seen as an added disadvantage.

2.2 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

$$\mathbb{P}(a \mid g) := \frac{s(g,a)}{\sum_{a' \in \mathcal{A}} s(g,a')}, \tag{12}$$

where $\mathcal{A}$ is the set of all cells in the MNOs network and $s(g,a) \in [0,\infty)$ stands for the *signal dominance* (an umbrella term introduced by ourselves) received in grid tile g from cell a. That is, the connection likelihood is the ratio of the signal dominance received from cell a to the total value of signal dominance received from all cells. Different choices for modelling the signal dominance are possible. This choice defines the *connection module* in Figure 1. Note that $\mathbb{P}(a \mid g)$ is independent of rescaling $s(g,a)$ by a constant, and our convention is that $s(g,a)$ should be defined so as to take on values in the interval $[0,1]$.

A simple method to define the connection module is via Voronoi tessellation. In this case $s(g,a)$ is set to be

$$s_{\text{Vor}}(g,a) := \begin{cases} 1 & \text{if } g \in \text{Vor}(a), \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\text{Vor}(a)$ is the set of grid tiles of which the centroids lie in the Voronoi region surrounding cell a.

In Section 3 we propose a different, more advanced definition of the connection module by first approximating the signal strength $S(g,a)$ measured in dBm, and then applying a transformation to it to obtain the signal dominance $s_{\text{strength}}(g,a)$.

2.2.1 Incorporating timing advance data

Some MNOs include in their signalling data a so called *timing advance* variable (3GPP 2019). When a device is connecting with a cell, the combination of communication delay and signal velocity can be used to estimate the distance between device and cell. The original purpose of timing advance data is to estimate and adjust for communication delays. However, it may alternatively be interpreted as a measure of distance between the device and the cell (Kreher and Gaenger 2011). The radius around the cell is split into annuli. For 4G signalling data, each annulus has a width of 78 m. The estimated annulus of a device is detailed in the variable τ, such that the distance between cell and device is between $\tau \cdot 78\,\text{m}$ and $(\tau + 1) \cdot 78\,\text{m}$, where $\tau \in \{0, 1, \ldots, 1282\}$.

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

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

and 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 the maximum value of τ is 1282, this new location posterior $\mathbb{P}(g \mid a, \tau)$ equals 0 for grid tiles further than approximately 100 km from the cell.

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 78 m, 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 annuli. For example, one could model $\mathbb{P}(\tau \mid a, g)$ as being 1 if the centroid of g lies in the annuli 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.

2.2.2 Example

To illustrate the computations involved in the model, we consider a small fictive island of 1 by 3 kilometer. The island can be divided into three grid

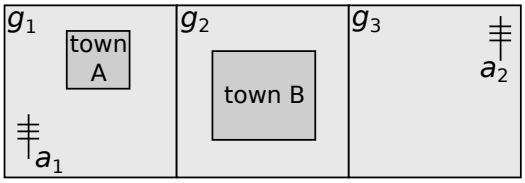

Figure 2: Fictive example: top view (schematic) of an island of 1 by 3 kilometers.

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 is 1 by 1 kilometer.

There are two small towns A and B. Town B is about three times as large. Two cells are placed on the island, a_1 and a_2. See Figure 2. Cell a_1 has perfect signal dominance in g_1 and g_2, but no signal in g_3. Cell a_2 has perfect signal dominance in g_3 and g_2, but no signal in g_1.

Table 2 shows the numbers of this example. In this example, our aim is to estimate the number of devices per tile in a specific time interval, say between 12:00 and 12:15 afternoon. Suppose that both cells have 150 connections during this period. The signal dominance, location priors, connection likelihood and resulting location posterior are listed in Table 2.

The estimated number of devices is indicated by the function $x(g)$ in Table 3.

The example illustrates the assumption in the uniform prior that the probability of a device being in a grid tile is equal for all grid tiles. When the location posterior values are added over both cells for each grid cell, it becomes apparent that the estimated number of devices is equal for all three tiles, whereas it is more likely that there are less people in g_3. These estimates are 100 devices per grid tile, even though g_2 has a better network coverage, since the signal dominance from both cells in g_2 is perfect. In other words, by using our likelihood function, the quality of the signal across the network is not reflected in the estimations of devices when using a uniform prior.

Note that no devices are assigned to g_3 for the land use prior, since $\mathbb{P}_{\text{land use}}(g_3) = 0$. Even if all land in g_3 is meadow without any roads, an number of estimated devices of 0 seems rather extreme. When using this land use prior, all devices connected to a_2 have been assigned to g_2. Note that the ratio of devices connected to a_1 that have been assigned to g_1 is higher than the prior value of $1/4$. This is because g_1 only served by a_1, whereas g_2 is also

Grid tile	g	1	2	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	$5/7$	$2/7$

Table 2: The corresponding numbers of the fictive example where the composite prior is based on $\pi = (0, 1/2, 1/2)$.

Grid tile	g	1	2	3
Estimation	$x_{\text{uniform}}(g)$	100	100	100
	$x_{\text{land use}}(g)$	60	240	0
	$x_{\text{network}}(g)$	75	150	75
	$x_{\text{composite}}(g)$	67	190	43

Table 3: The corresponding estimated number of devices of the fictive example where the composite prior is based on $\pi = (0, 1/2, 1/2)$.

served by a_2. This is reflected by the connection likelihood.

The signal dominance is used in both the connection likelihood and the network prior, but in different ways. In the connection likelihood, the signal dominance is used to compensate for overlapping cell areas whereas in the network prior, it is used to add more weight to tiles that have a good total signal dominance.

The aim of the network prior is to take the signal dominance into account in the final estimations. In the example, the total signal dominance in g_2 is twice as high as in the other tiles. Therefore, the estimated number of devices is 75, 150, and 75 for g_1, g_2, and g_3 respectively.

In this example, we have configured the composite prior such that it is the average between the land use prior and the network prior. Although the result seems plausible for this example, validation methods are needed to assess the quality of the used prior. Note that the estimates per grid tile of the composite prior are not the average between the land use and network estimates.

2.3 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 numbers of people, for instance the number of visitors of a touristic city during holidays, or the number of people who commute between two cities. Additional methods and auxiliary data are needed to translate estimations of devices to estimations of people.

A generic framework has been proposed to organize the production process needed for statistical inference on mobile phone network data (Ricciato 2018). 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 used mobile network technology. 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. However, since the whole process should ideally be run at the MNO due to potential privacy issues, the location prior will also be placed along with the other modules in the D layer.

3 Signal strength model

This section describes the propagation of signal strength originating from a single cell. We distinguish two types of cells: omnidirectional and directional, resulting in two different propagation models. Omnidirectional cells have no aimed beam and their coverage can be thought of as a circular disk. Directional cells point in a certain direction and their coverage can be thought of as an oval with one axis of symmetry. In practice, small cells are omnidirectional and normal cells (i.e. attached to cell towers or placed on rooftops) are directional (Kora et al. 2016).

3.1 Omnidirectional cells

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

$$S(g, a) := S_0 - S_{\text{dist}}(r_{g,a}), \tag{14}$$

where S_0 is the signal strength at $r_0 = 1$ meter distance from the cell in dBm 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 information, it is common to list the power P of a cell in Watt, rather than the signal strength in dBm. The

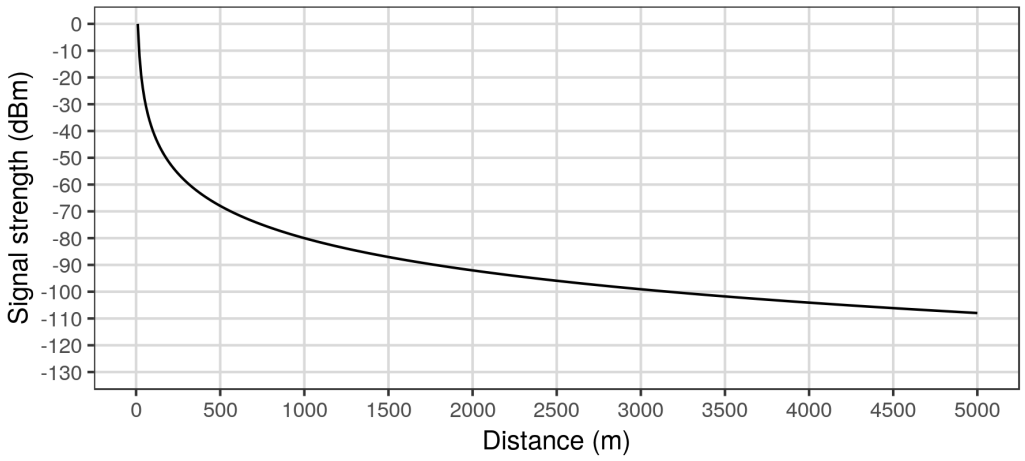

Figure 3: Signal loss as a function of the distance for a specific cell.

value of S_0 can be calculated from P using the conversion between Watt and dBm (**figueiras2010**):

$$S_0 = 30 + 10 \log_{10}(P). \tag{15}$$

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

$$S_{\text{dist}}(r) := 10 \log_{10}(r^\gamma) = 10\gamma \log_{10}(r), \tag{16}$$

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 (Srinivasa and Haenggi 2009). 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 environments. Special situations, such as tunnels, could improve the propagation such that a value of less than 2 is applicable. In our implementation, we approximate the path loss exponent by using the land use register, as we will describe in Section 5.

In Figure 3, the signal loss as a function of the distance is shown for a cell with 10 W power that is standing in an urban environment ($\gamma = 4$).

3.2 Directional cells

A directional cell has 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 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 dB or less. At 3 dB, the power of the signal is halved. The angles in the elevation plane for which the signal loss is 3 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 dB. The angles in which the signal loss is 3 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

$$S(g, a) := S_0 - S_{\text{dist}}(r_{g,a}) - S_{\text{azi}}(\delta_{g,a}, \alpha_a) - S_{\text{elev}}(\varepsilon_{g,a}, \beta_a), \tag{17}$$

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 Equation (16).

see: Additions:

4 Numerical solving for parameter calculation

Each cell type has its own signal strength pattern for both the azimuth and elevation angles. 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 both S_{azi} and S_{inc} by a linear transformation of the Gaussian formula, each with different values for parameters c and σ. Let

$$f(\varphi) = c - ce^{-\frac{\varphi^2}{2\sigma^2}}, \tag{18}$$

where c and σ^2 are constants, whose value is determined by numerically solving equations for a set of constraints. These constraints are different for S_{azi} and S_{elev} and depend on cell properties.

The resulting patterns are shown in Figure 4. 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 dB loss (which is only achieved in the main direction), the next circle corresponds to 5 dB loss, an so forth. The red lines denote the angles corresponding to 3 dB loss. The angle between the red lines is $2\alpha_a$ in the azimuth plane and $2\beta_a$ 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.

Figure 5 (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 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

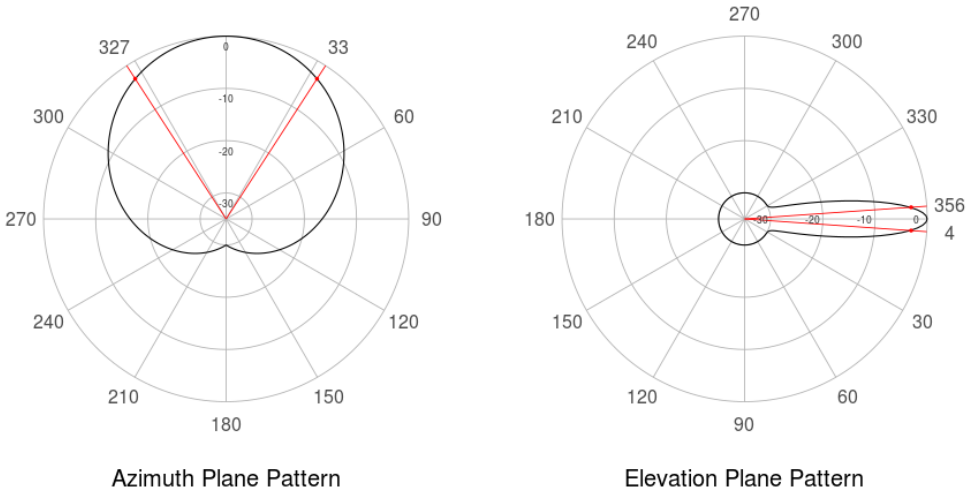

Figure 4: Radiation patterns for the azimuth and elevation planes.

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

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

meters distance. This is caused by relatively large ε angles at grid tiles nearby the cell.

4.1 Signal dominance

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 also called *load balancing*.

Our model allows for two phenomena that we feel should not be overlooked. The first is the switching of a device when it is receiving a bad signal to a cell with a better signal. Table 4 describes how the signal strength can be interpreted in terms of quality for 4G networks (Kora et al. 2016). The second

Figure 5: Signal strength (top row) and signal dominance (bottom row) at ground level.

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. The specifics of this decision making system are considered unknown.

We assume that a better signal leads to a higher chance of connection. When a device has multiple cells available with a signal strength above a certain threshold, say $-90\,\text{dBm}$, the signal strengths are both more than good enough and the cell with the highest capacity is selected rather than the cell with the best signal strength. When the choice is between cells with a lower signal strength, one can imagine that their relative differences play a more important role in the connection process. However, when there are multiple cells available with a poor signal strength, it can be assumed that the signal strength value is less important than having capacity. In short, we assume that signal strength plays a more important role in load balancing when it is in the middle range instead of in the high quality or low quality ranges.

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 $s_{\text{strength}}(g, a)$, which is then used to define the connection likelihood (13). Let us define

$$s_{\text{strength}}(g, a) := \frac{1}{1 + \exp\Big(-S_{\text{steep}}\left(S(g, a) - S_{\text{mid}}\right)\Big)}, \tag{19}$$

where S_{mid} and S_{steep} are parameters that define the midpoint and the steepness of the curve respectively. Figure 6 shows an example of Equation (19).

The signal dominance at ground level is shown in Figure 5 (bottom row). The values that are shown are normalized by the sum of all values over all grid tiles, such that the normalized values form a probability distribution. Compared to the signal strength shown in Figure 5 (top row), the signal dominance puts more emphasis on the geographic area that is in the range of the cell. Whether these signal dominance values resemble reality, should be validated by field tests.

5 Implementation

The methods described in this paper have been implemented in `mobloc` (Tennekes 2018), a package for the programming language `R`. In this section, we provide a general description of `mobloc` and its work flow. Details and

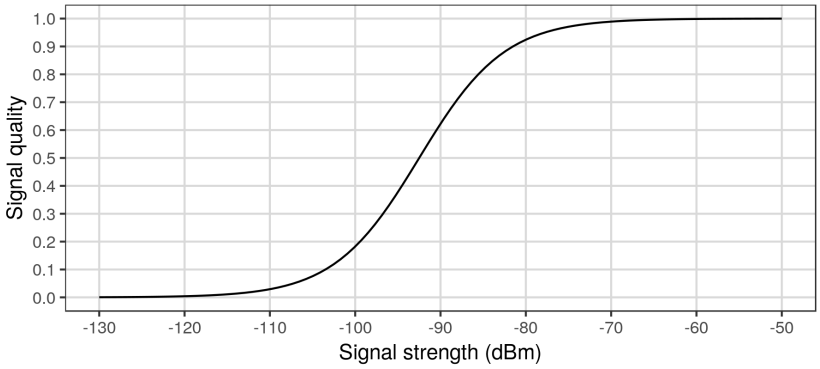

Figure 6: Logistic relation between signal strength (dBm) and signal dominance, where S_{mid} and S_{steep} are set to $-92.5\,\text{dBm}$ and $0.2\,\text{dBm}$ respectively to resemble Table 4.

instructions for use are described in the package documentation. A good starting point for new users is the vignette, in which the whole process is explained by means of a working example. After installation of the package, the vignette can be opened by running the code `vignette("mobloc")`.

The first step in the work flow is to collect all relevant datasets. The most important dataset is called the *cell plan*, which contains data about the cells. Other datasets that can be used as input are elevation data, land use data, and administrative region boundaries.

The next step is to set up to the propagation model with an interactive tool. A screenshot of this tool is shown in Figure 7. With this tool, default values for missing cell plan data can be specified. The parameters regarding the signal dominance, namely S_{mid} and S_{steep} in (19), can be specified also.

When the parameters have been specified, the cell plan data should be validated. In this process, the input format is checked and missing data are imputed. The variables of the cell plan are listed in Table 5. The `cell id`, and the `x` and `y` coordinates are the only required variables. The `direction` (azimuth angle) is strongly recommended, since it has great influence on the propagation. If omitted, the corresponding cell is assumed to be omnidirectional.

The elevation data is needed to determine the `z` coordinate, which is the elevation (i.e. meters above sea level) plus the height of the cell. If the height is not known per cell, a default value can be used.

In order to derive a path loss exponent value per cell, land use data can

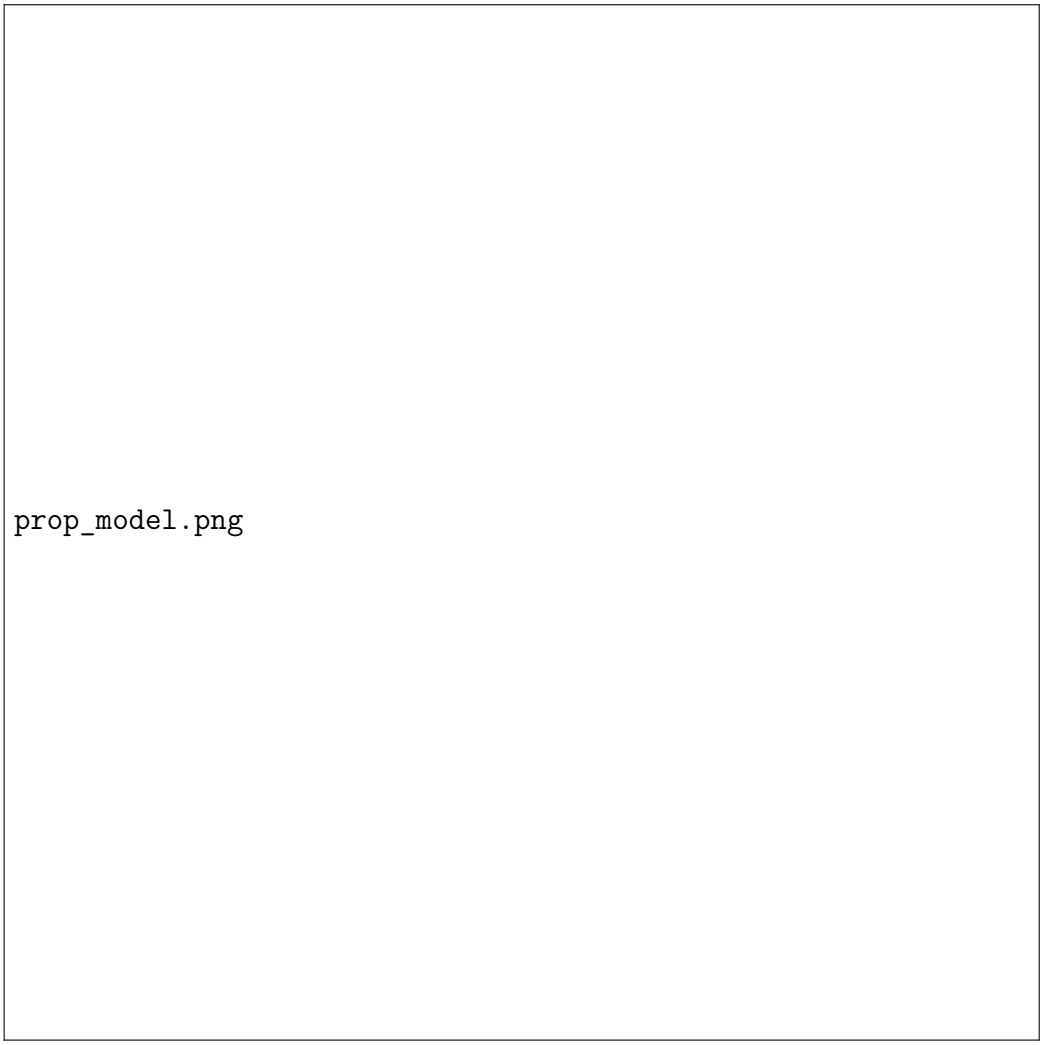

Figure 7: Propagation model setup tool.

Parameter	Symbol	Name in `mobloc`
cell id		`cell`
x coordinate		`x`
y coordinate		`y`
z coordinate		`z`
cell height		`height`
azimuth angle	φ	`direction`
elevation angle	θ	`tilt`
horizontal beam width	α	`beam_h`
vertical beam width	β	`beam_v`
power	P	`W`
path loss exponent	γ	`ple`

Table 5: A list of parameters.

be used as follows. For each grid tile, a measure can be derived that reflects the amount of objects that may reflect the propagation, such as buildings and trees. Next, a couple of grid tiles around the cell are sampled. From those grid tiles, the mean measurement value is computed and used to determine the path loss exponent value. The information about this method can be found in the software documentation (Tennekes 2018).

The next step is to compute the propagation and the connection likelihood. Due to computational limitations, the signal strength is not calculated for all cell and grid tile combinations. Instead, the signal strength is first calculated for grid tiles in the propagation direction of the cell, as well as in the opposite direction. Between the two points where the signal dominance drops below a certain threshold value (so one in front of the cell and one at the back) a circle is created. For each grid tile in this circle, the signal strength, the signal dominance, and the connection likelihood values are calculated.

The work flow is proceeded with deriving prior distributions. All priors suggested in Section 2.1 can be created by functions implemented in `mobloc`. Any other prior distribution can be used as well.

Finally, the location posterior distributions can be calculated. When timing advance data is available, these distributions can be updated with timing advance values. This results in location posterior distributions per cell and timing advance value. These results can be analysed interactively as shown in Figure 8.

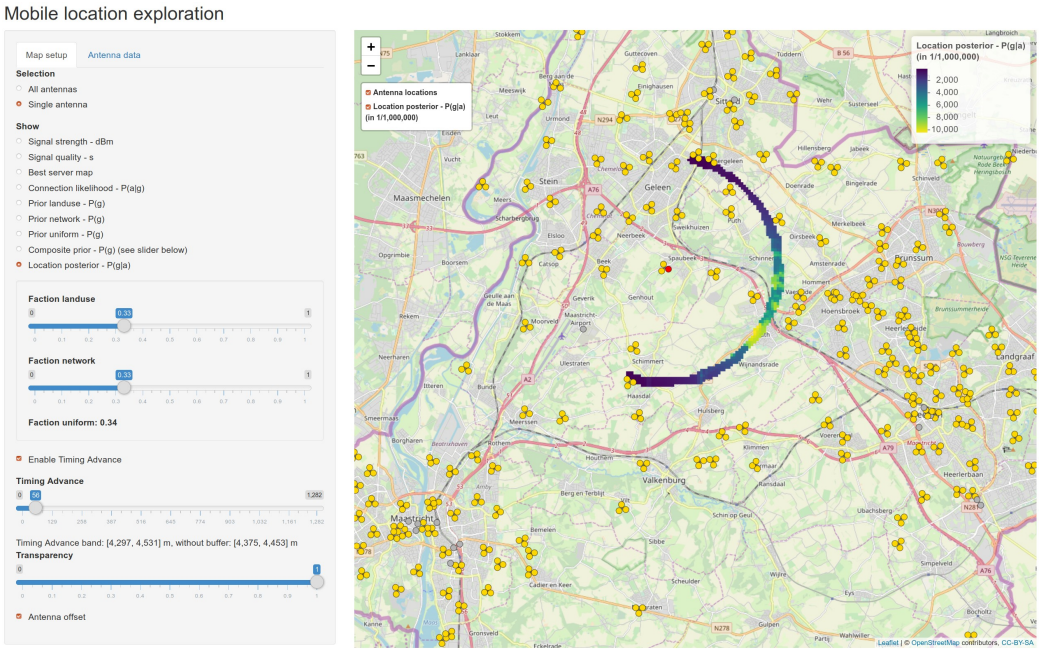

Figure 8: Tool to explore the results.

6 Concluding remarks

We proposed a method for location estimation of mobile devices using mobile network data. Per cell, we estimated a location posterior using a signal propagation model and prior location information. The location posterior can be used as an estimate of where a device is located.

An MNO often facilitates mobile communication via multiple generations of networks (e.g. 3G and 4G). For each generation, an MNO maintains a cellular network. We currently assume that a mobile device only connects to cells from one network generation, that is, the latest generation which the device supports. When this assumption holds, the methods can be independently applied for each generation. The networks can be viewed as independent because a device will only connect to cells from one generation. In reality however, this assumption might not hold for reasons such as coverage gaps, capacity and network-specific optimal communication mode such as text, voice and internet. More research on switching between cells from different generations is needed.

For each generation, cells will serve at different frequencies. For instance

an MNO may have a network of 4G cells that serve at 900, 1800 and 2100 Hz. As with handling of cells from multiple generations, more research is needed on the process of switching between frequencies. In the current implementation, cells with the same values for the variables listed in Table 5 that serve at different frequencies are considered as one cell.

Note that the methods described in this paper are modular, in the sense that one method can easily be replaced by another. If for instance a better propagation model exists, for instance by using 3d modelling of the environment, this can be used together with the other methods. The same applies for the signal dominance function, connection likelihood, and location prior.

Our propagation model can be calibrated using other data sources. Field measurements of received signal strength could provide insights in several parameters, for instance the power of the cells and the path loss exponent. The calibration process should ideally be executed for each mobile phone network, since they may be configured in different ways.

MNOs also use propagation models for network planning. Although it is not always clear what sources they use, the results can be used for validation. However, beware that these sources may not be the ground truth. A validation task that is easy to do is to compare coverage maps and best server maps from the MNO, which is often available, to those resulting from our methods, created with `mobloc`. More research is needed on how to quantify the comparison between such maps. Since these maps have probably been derived from propagation models, note that they may not represent the ground truth.

Small-scale validation is also an important part of the process. One method is to log GPS location data for a sample of devices and to compare these location data with the results of our methods based on mobile network data. Consent is required from the MNO as well as from the device owners.

The location estimation method presented in this article can be used to convert an mobile network event to a probability distribution of likely grid tiles for the location of the mobile device. This is useful for various applications, such as daytime population statistics and research on mobility patterns such as commuting and tourism.

References

3GPP (2019). *TS Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*.

Ahas, R. et al. (2015). "Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn". In: *International Journal of Geographical Information Science* 29.11, pp. 2017–2039.

Alexander, Lauren et al. (2015). "Origin–destination trips by purpose and time of day inferred from mobile phone data". In: *Transportation Research Part C: Emerging Technologies* 58, pp. 240–250. DOI: https://doi.org/10.1016/j.trc.2015.02.018.

Calabrese, Francesco, Laura Ferrari, and Vincent Blondel (2014). "Urban sensing using mobile phone network data: A survey of research". In: *ACM Computing Surveys* 47.2, pp. 1–20.

De Meersman, Freddy et al. (2016). "Assessing the quality of mobile phone data as a source of statistics". In: *European Conference on Quality in Official Statistics*. Eurostat. Madrid.

Deville, Pierre et al. (2014). "Dynamic population mapping using mobile phone data". In: *Proceedings of the National Academy of Sciences* 111.45, pp. 15888–15893. DOI: 10.1073/pnas.1408439111.

Diao, Mi et al. (2016). "Inferring individual daily activities from mobile phone traces: A Boston example". In: *Environment and Planning B: Planning and Design* 43.5, pp. 920–940. DOI: 10.1177/0265813515600896.

EPSG (2018). *Coordinate Systems Worldwide*. URL: http://www.epsg.org.

Graells-Garrido, Eduardo, Oscar F. Peredo, and José García (2016). "Sensing Urban Patterns with Antenna Mappings: The Case of Santiago, Chile". In: *Sensors* 16.7, p. 1098.

Iqbal, Md Shahadat et al. (2014). "Development of origin–destination matrices using mobile phone call data". In: *Transportation Research Part C: Emerging Technologies* 40, pp. 63–74. DOI: 10.1016/j.trc.2014.01.002.

Järv, Olle, Henrikki Tenkanen, and Tuuli Toivonen (2017). "Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation". In: *International Journal of Geographical Information Science* 31.8, pp. 1630–1651.

Jiang, Shan et al. (2016). "The TimeGeo modeling framework for urban motility without travel surveys". In: DOI: 10.1073/pnas.1524261113. URL: http://www.pnas.org/content/early/2016/08/24/1524261113.

Jonge, E. de, M. Pelt, and M. Roos (2012). *Time patterns, geospatial clustering and mobility statistics based on mobile phone network data*. Discussion paper. Statistics Netherlands.

Kondor, Dániel et al. (2017). "Prediction limits of mobile phone activity modelling". eng. In: *Royal Society open science* 4.2.

Kora, Ahmed D. et al. (2016). "Accurate Radio Coverage Assessment Methods Investigation for 3G/4G Networks". In: *Computer Networks* 107.P2, pp. 246–257. ISSN: 1389-1286.

Kreher, Ralf and Karsten Gaenger (2011). *LTE Signaling, Troubleshooting and Optimization*. 1st ed. John Wiley and Sons, Ltd.

Lu, Xin et al. (2016). "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh". In: *Global Environmental Change* 38, pp. 1–7. DOI: 10.1016/j.gloenvcha.2016.02.002.

Panwar, Nisha, Shantanu Sharma, and Awadhesh Kumar Singh (2016). "A survey on 5G: The next generation of mobile communication". In: *Physical Communication* 18. Special Issue on Radio Access Network Architectures and Resource Management for 5G, pp. 64–84.

Pucci, Paola, Fabio Manfredini, and Paolo Tagliolato (2015). *Pucci P., Manfredini F., Tagliolato P. (2015), Mapping urban practices through mobile phone data, PoliMI SpringerBriefs Series*.

Ricciato, F. (2018). "Towards a Reference Methodological Framework for processing MNO data for Official Statistics". In: *15th Global Forum on Tourism Statistics*.

S. Kung, Kevin, Stanislav Sobolevsky, and Carlo Ratti (2013). "Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data". In: *PloS one* 9.

Salgado, D. et al. (2018). *Proposed Elements for a Methodological Framework for the Production of Official Statistics with Mobile Phone Data., ESSnet Big Data, WP5, Deliverable 5.3*. Eurostat.

Srinivasa, S. and M. Haenggi (2009). "Path loss exponent estimation in large wireless networks". In: *2009 Information Theory and Applications Workshop*, pp. 124–129.

Tennekes, M. (2018). *mobloc: Mobile phone location algorithms and tools*. R package version 0.1. URL: https://github.com/MobilePhoneESSnetBigData/mobloc.

Tennekes, M., M.P.W. Offermans, and N. Heerschap (2017). "Determining an optimal time window for roaming data for tourism statistics". In: *Proceedings of the NetMob 2017 Conference*.

Tolstrup, M. (2015). *Indoor Radio Planning: A Practical Guide for 2g, 3g and 4g*. Wiley.

Tyner, Judith A. (2010). *Principles of Map Design*. Guilford Press.

Wang, Shaowei, Wentao Zhao, and Chonggang Wang (2015). "Budgeted Cell Planning for Cellular Networks With Small Cells". In: *Vehicular Technology, IEEE Transactions on* 64, pp. 4797–4806.

Widhalm, Peter et al. (2015). "Discovering urban activity patterns in cell phone data". In: *Transportation* 42.4, pp. 597–623.

Wilson, R. et al. (2016). "Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake." In: *PLOS Currents Disasters*.

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

Xu, Y. et al. (2018). "Human mobility and socioeconomic status: Analysis of Singapore and Boston". In: *Computers, Environment and Urban Systems*.

Zagatti, Guilherme Augusto et al. (2018). "A trip to work: estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR". In: *Development Engineering* 3, pp. 133–165.