Center for
Big Data Statistics
Working paper

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

Working paper no.: 06-20

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

May 2020

Contents

1 Introduction

Mobile phone network data have shown to be a rich potential source for official statistics, in particular on daytime population (Ahas et al., 2015; De Meersman et al., 2016; Deville et al., 2014; Kondor et al., 2017; Salgado et al., 2018; Xu et al., 2018), mobility (Alexander et al., 2015; Diao et al., 2016; Iqbal et al., 2014; Jiang et al., 2016; Jonge et al., 2012; Pucci et al., 2015; S. Kung et al., 2013; Widhalm et al., 2015; Zagatti et al., 2018), migration (Lu et al., 2016; Wilson et al., 2016), and tourism (Deville et al., 2014; Tennekes et al., 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*.

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

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 et al., 2016): a *macro cell* that is placed in a cell tower or on top of a roof, which has a theoretical range of 30 kilometers, 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.[1)] Furthermore, a cell can be directional or omnidirectional. In practice, most real-world deployed macro cells are directional, often covering an angle of about 120 degrees whereas small cells are usually omnidirectional.

For statistical inference on mobile phone network data, geographic location is one of the most important variables. 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 et al., 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 few downsides to 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 we would like to mention 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 (De Meersman et al., 2016; Graells-Garrido et al., 2016; Ricciato 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, 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 et al., 2017).

We propose a modular Bayesian framework to estimate the geographic location of devices, which consists of two main modules, namely the *location prior module* and the *connection module*. The former employs a priori information about where devices are expected to be to produce a *location prior*, and the latter uses network information to estimate the probability of

[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.

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

connection to cells to produce a *connection likelihood*. 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, 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 outcome of the Bayesian framework is the *location 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. For implementation, we recommend to overlay the area of interest with a grid of square tiles, typically 100 by 100 meters. The location posterior then defines the probability that a device is located in a grid tile given that it is connected to a certain cell.

The location posterior can be used for further statistical inference. We propose a simple estimator for the number of devices for a given time interval, which will be illustrated with a simple example. Other location estimation methods have been proposed in literature, for example using maximum likelihoods (Laan & Jonge, 2019; Ricciato et al., 2019).

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 4 we illustrate the method with an fictional application. We conclude with a couple of remarks in Section 5

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

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

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.

As a technical aside, 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 Equation (1), 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 2.2, 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.

The position of Equation (1) in our modular Bayesian framework is illustrated in Figure 2.1. The location prior and connection likelihood are estimates produced from models we call the location prior module and connection module, respectively. The former module may use land use data as input, while we propose to use static network data for the latter in the form of the *cell plan*, which is a dataset that contains the locations and physical properties of all the cells in the network. 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 necessary information is available. Figure 2.1 shows how a timing advance module might be used to construct an updated location posterior.

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

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.

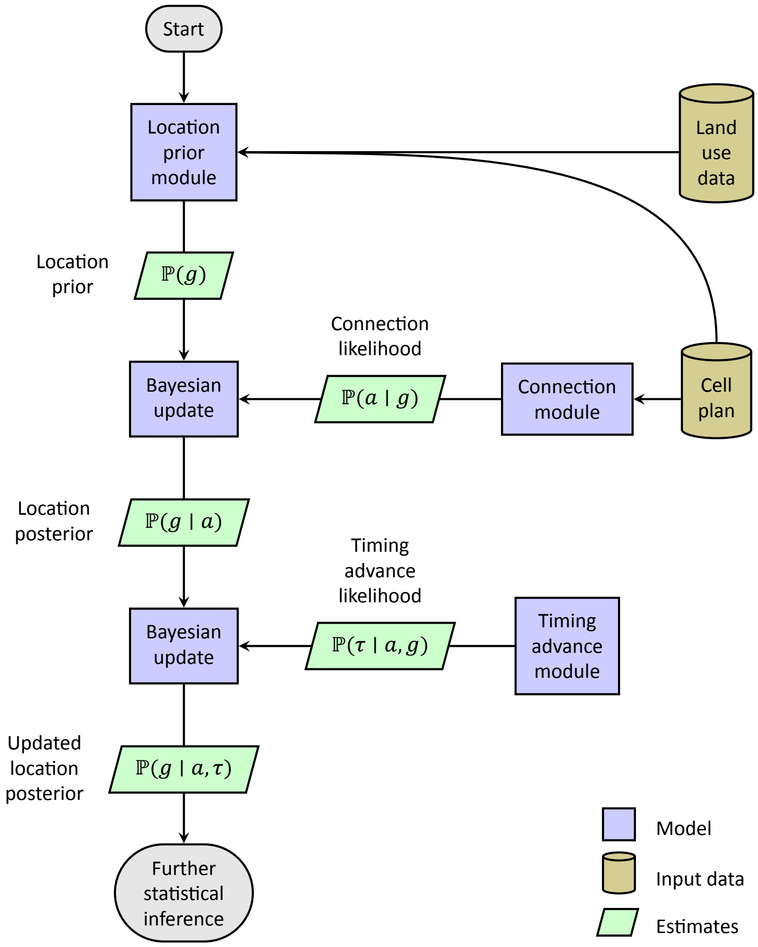
The definition of the location prior function will be based on assumptions about where a device is expected to be. In this section, we propose four options: the uniform prior, the land use prior, the network prior, and the composite priors.

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:

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

2.1 A modular framework for modelling location posteriors.



2.1 An example of land use classes and their relative expected number of devices.

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

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.

2.1.2 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 meadow. Hence, our second proposed 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: $u_1, u_2, \dots, u_K$. Let $w_1(g), w_2(g), \dots, w_K(g) \in [0, 1]$ be the proportion of the grid tile that is covered by each class respectively, such that

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

Then $n(g)$ can be modelled as

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

An example of a simple land use classification and the relative expected numbers is shown in Table 2.1.

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 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 (Tolstrup, 2015; Wang et al., 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 following prior, which we call the *network prior*, is defined as

$$\mathbb{P}(g) := \frac{\sum_{a \in \mathcal{A}} s(g,a)}{\sum_{a \in \mathcal{A}} \sum_{g' \in \mathcal{G}} s(g',a)}, \tag{8}$$

where $s(g,a)$ for a grid tile g and cell a is called *signal dominance*, and $\mathcal{A}$ denotes the set of all cells in the network. In Section 2.2, a general notion of *signal dominance* will be introduced to serve as input for our model for the connection likelihood. A specific instance of it, the *Voronoi signal dominance* $s_{\text{Vor}}(g,a)$, will be defined in that section as well, and another instance $s_{\text{strength}}(g,a)$ will be elaborated on in Section 3. Any of these instances can be used for Equation (8).

This prior, together with our model for the connection likelihood $\mathbb{P}(a \mid g)$, which is discussed in Section 2.2, simplifies Equation (1) to

$$\mathbb{P}(g \mid a) \propto s(g,a). \tag{9}$$

The interpretation of the prior (8) depends on the instance of the signal dominance $s(g,a)$ one has chosen. Let $F\colon \mathcal{G} \to \mathcal{A}$ be an arbitrary function that maps grid tiles to cells. Then, by setting

$$s_F(g,a) := \begin{cases} 1 & \text{if } a = F(g), \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

the prior in Equation (8) becomes uniform. Since s_{Vor} is a specific instance of the function s_F, the network prior with $s = s_{\text{Vor}}$ also simplifies to the uniform prior.

When using the term *network prior* from here on, we refer to Equation (8) with signal dominance based on the instance $s_{\text{strength}}(g,a)$.

Basically, the network prior reflects the distribution of the total signal over all the grid tiles. 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.

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. The example in Section 2.4 includes calculations in which the signal strength model is used for both the connection likelihood and location prior.

2.1.4 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:

$$\begin{aligned}\mathbb{P}_{\text{composite}}(g) := \pi_{\text{uniform}} \cdot \mathbb{P}_{\text{uniform}}(g) + \\ \pi_{\text{land use}} \cdot \mathbb{P}_{\text{land use}}(g) + \\ \pi_{\text{network}} \cdot \mathbb{P}_{\text{network}}(g),\end{aligned} \tag{11}$$

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

$$\begin{cases} 0 \leq \pi_{\text{uniform}} \leq 1, \\ 0 \leq \pi_{\text{land use}} \leq 1, \\ 0 \leq \pi_{\text{network}} \leq 1, \\ \pi_{\text{uniform}} + \pi_{\text{land use}} + \pi_{\text{network}} = 1. \end{cases} \tag{12}$$

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 optimal choice of π depends on the situation at hand and must be derived experimentally. 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.

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

$$\mathbb{P}(a \mid g) := \frac{s(g, a)}{\sum_{a' \in \mathcal{A}} s(g, a')}, \tag{13}$$

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, and any choice defines the *connection module* in Figure 2.1. Note that $\mathbb{P}(a \mid g)$ is independent of rescaling the function $s(g, a)$ by a constant, and our convention is that $s(g, a)$ should be defined so as to take on values in the interval $[0, 1]$.

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

$$s_{\text{Vor}}(g, a) := \begin{cases} 1 & \text{if } g \in \text{Vor}(a), \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

where Vor(a) is the set of grid tiles of which the centroids lie in the Voronoi region surrounding cell a. Denoting (13) in this case by $\mathbb{P}_{\text{Vor}}(a \mid g)$ then gives

$$\mathbb{P}_{\text{Vor}}(a \mid g) = s_{\text{Vor}}(g, a), \tag{15}$$

and combining this connection likelihood with the uniform prior $\mathbb{P}_{\text{uniform}}(g)$ results in the location posterior

$$\mathbb{P}_{\text{Vor}}(g \mid a) = \begin{cases} \left|\{g' \in \mathcal{G} \mid g' \in \text{Vor}(a)\}\right|^{-1} & \text{if } g \in \text{Vor}(a), \\ 0 & \text{otherwise}. \end{cases} \tag{16}$$

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 a signal dominance $s_{\text{strength}}(g, a)$.

2.3 Incorporating timing advance data

Some MNOs include in their signalling data a so called *timing advance* variable (3GPP, 2019). The value of such a variable represents a time duration, and it is used to estimate and adjust for communication delays. However, in combination with the speed of radio waves it can alternatively be used to estimate the distance between device and cell (Kreher & Gaenger, 2011). In the case of 4G signalling data the timing advance variable takes on values in the discrete set $\{0, 1, \dots, 1282\}$. 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 m and an inner circle of radius $\tau \cdot 78$ m.

One should be aware, though, of errors present in this approximation. As the authors of (Raitoharju et al., 2010) 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. We namely have

$$\mathbb{P}(g \mid a, \tau) \propto \mathbb{P}(g \mid a)\mathbb{P}(\tau \mid a, g), \tag{17}$$

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.

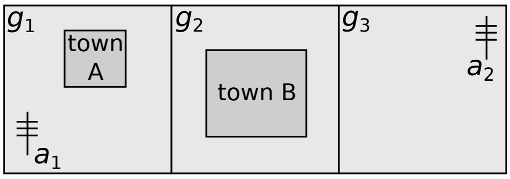
Computing the likelihoods $\mathbb{P}(\tau \mid a, g)$ for all timing advance annuli around all cells a in the network and all tiles g in the grid that is used might prove to be too expensive computationally if calculated in the way we suggest above. One could therefore model $\mathbb{P}(\tau \mid a, g)$ 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 $\{\tau - b, \dots, \tau + b\}$, 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.

2.4 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 2.2. 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 Figure 2.2. Cell a_1 has perfect signal dominance in g_1 and g_2, but no signal in g_3, while cell a_2 has perfect signal dominance in g_3 and g_2, but no signal in g_1. Perfect signal dominance and no signal

2.2 A schematic top view of an island of 1 by 3 kilometers.



2.2 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	$5/7$	$2/7$

are expressed by values 1 and 0, respectively. The signal dominance can be interpreted as the outcome of the signal strength model from Section 3.

We calculate the four priors as defined in Section 2.1 based on the above information. The connection likelihood for each cell is calculated according to Equation 13 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 Figure 2.1.

2.5 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 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.

A generic framework has been proposed to organize the production process needed for statistical inference on mobile phone network data (Ricciato, 2018; Ricciato et al., 2019). 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.

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 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 normal cells (i.e. attached to cell towers or placed on rooftops) are often directional (Kora et al., 2016).

3.1 Omnidirectional cells

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

$$S(g, a) := S_0 - S_{\text{dist}}(r_{g,a}), \tag{18}$$

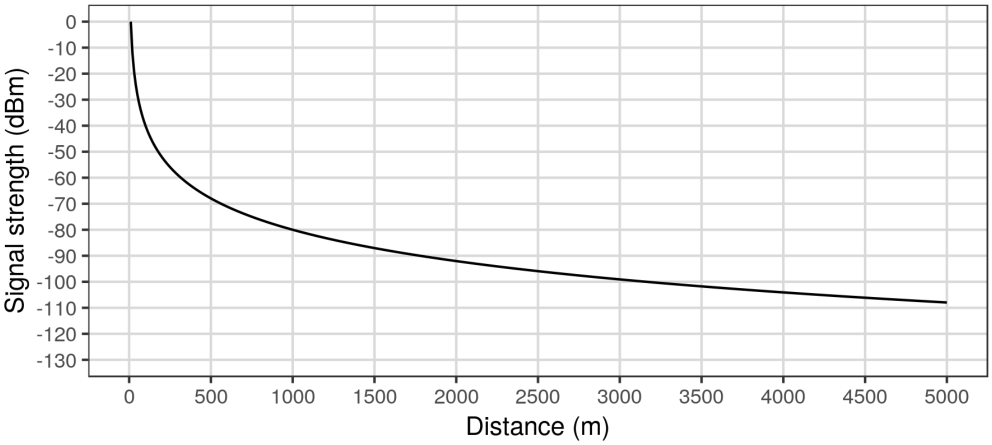
where S_0 is the signal strength at $r_0 = 1$ meter distance from the cell in dBm and $r_{g,a}$ is the distance between the middle point of grid tile g and cell a in meters. The value of S_0 can be different for every cell and is assumed to be a known property. In cell plan information, it is common to list the power P of a cell in Watt, rather than the signal strength in dBm. The value of S_0 can be calculated from P using the conversion between Watt and dBm (Figueiras & Frattasi, 2010):

$$S_0 = 30 + 10 \log_{10}(P). \tag{19}$$

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

$$S_{\text{dist}}(r) := 10 \log_{10}(r^{\gamma}) = 10\gamma \log_{10}(r), \tag{20}$$

3.1 Signal loss as a function of the distance for a specific cell.



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 & 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. The path loss exponent can be approximated by using the land use register.

In Figure 3.1, 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 ($\gamma = 4$).

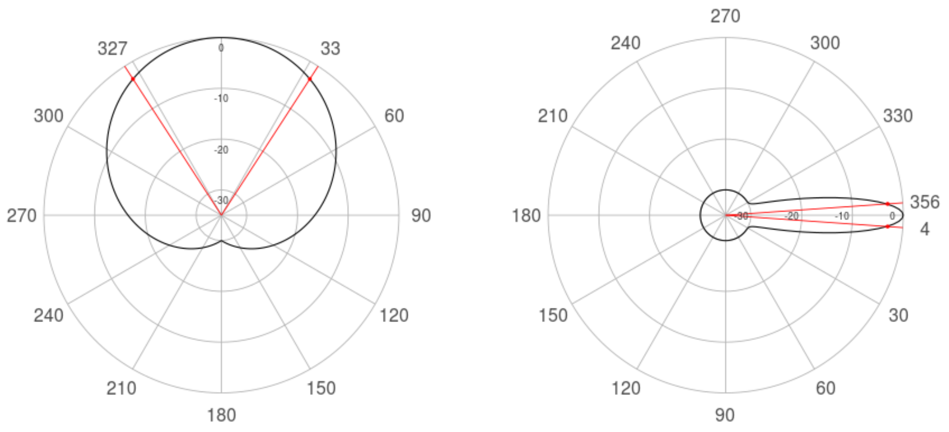
3.2 Directional cells

A directional cell 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 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

3.2 Radiation patterns for the azimuth and elevation planes.



Azimuth Plane Pattern

Elevation Plane Pattern

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

$$S(g,a) := S_0 - S_{\text{dist}}(r_{g,a}) - S_{\text{azi}}(\delta_{g,a}, \alpha_a) - S_{\text{elev}}(\varepsilon_{g,a}, \beta_a), \tag{21}$$

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 (20).

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_{elev} by a linear transformation of the Gaussian formula, each with different values for parameters c and σ. Let

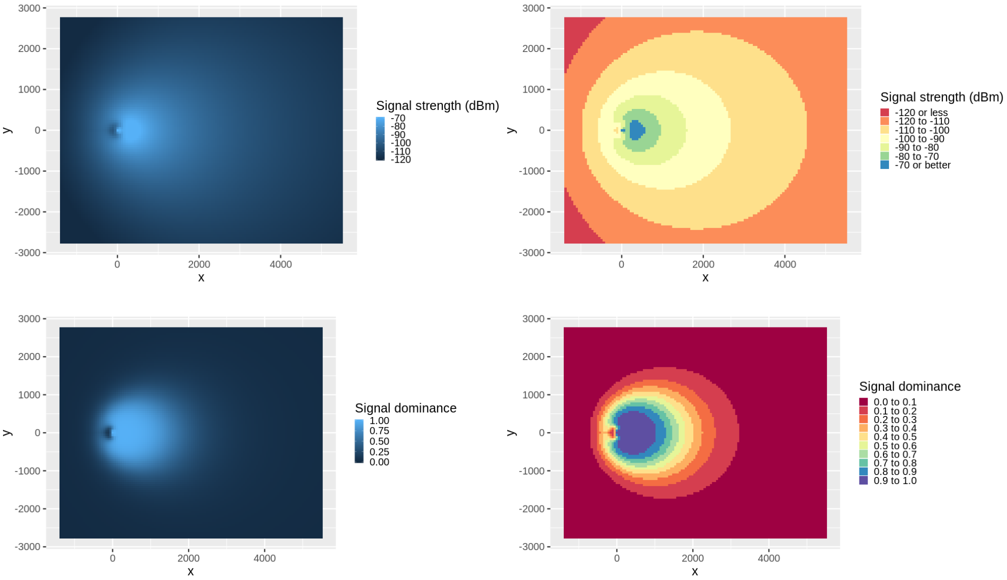
$$f(\varphi) = c - ce^{-\frac{\varphi^2}{2\sigma^2}}, \tag{22}$$

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 3.2. 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 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.

3.3 Signal strength (top row) and signal dominance (bottom row) at ground level.



3.1 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

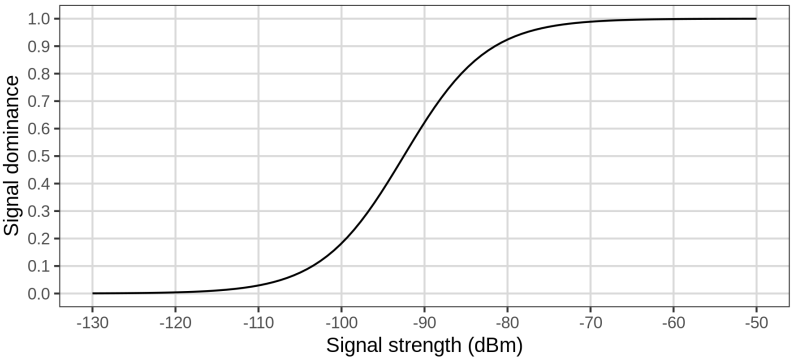
Figure 3.3 (top row) illustrates the signal strength at the ground level from above for a specific cell. In this case, the cell is placed at $x = 0$, $y = 0$ at 55 meters above ground level in an urban environment ($\gamma = 4$), has a power of 10 W, and is directed eastwards with an elevation angle (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.

3.3 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 3.1 describes how the signal strength can be interpreted in terms of quality for 4G networks (Kora et al., 2016). 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. The specifics of this system are considered unknown.

3.4 Logistic relation between signal strength (dBm) and signal dominance, where S_{mid} and S_{steep} are set to -92.5 dBm and 0.2 dBm respectively to resemble Table 3.1.



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 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

$$s_{\text{strength}}(g, a) := \frac{1}{1 + \exp\big(-S_{\text{steep}}\,(S(g, a) - S_{\text{mid}})\big)}, \tag{23}$$

where S_{mid} and S_{steep} are parameters that define the midpoint and the steepness of the curve respectively. Figure 3.4 shows an example of Equation (23).

The signal dominance at ground level is shown in Figure 3.3 (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 3.3 (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.

4 Application

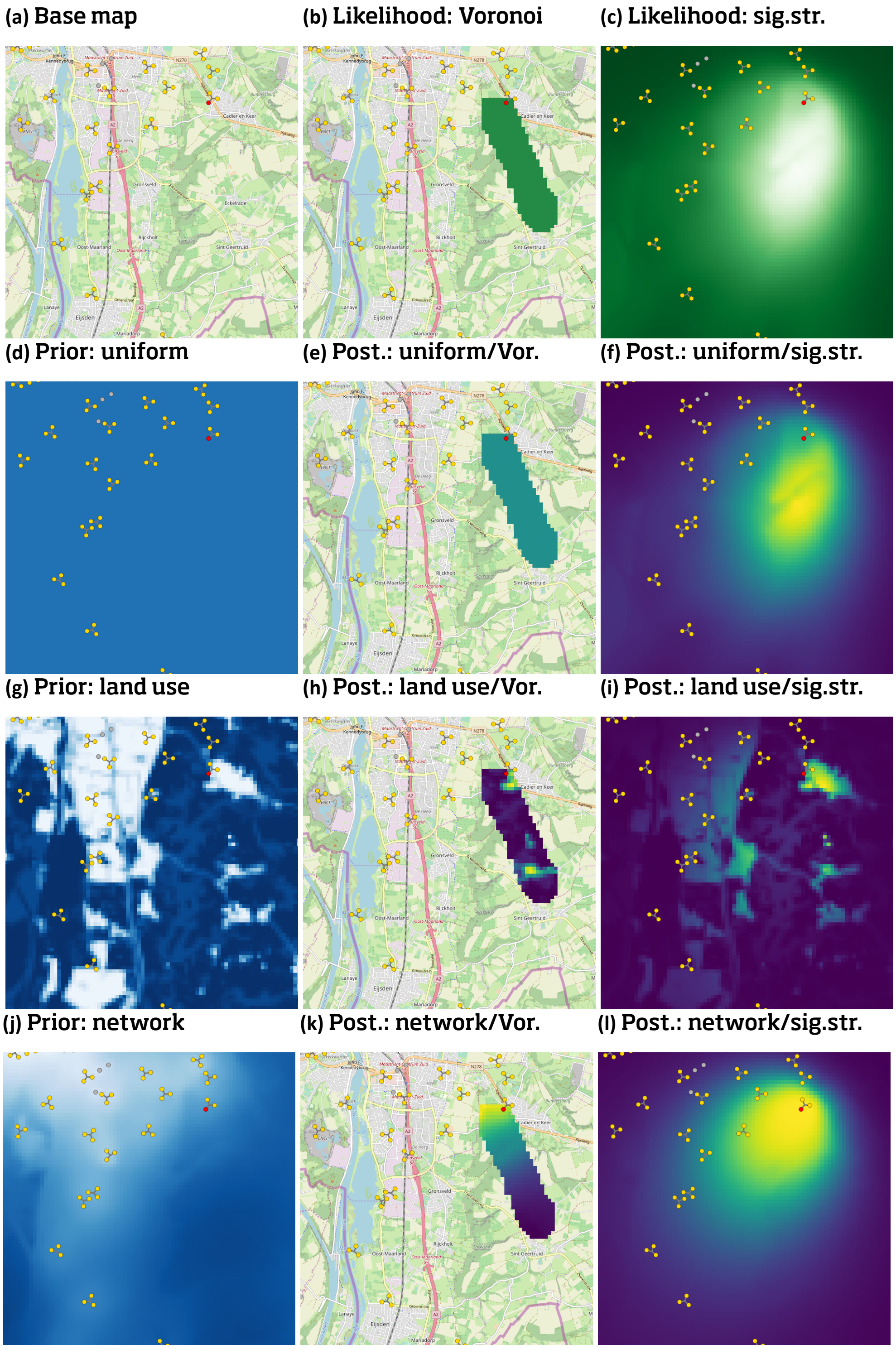
A fictional example that resembles a real-world situation is illustrated in Figure 4.1. Figure 4.1(a) shows a base map with artificially placed cells. The northwestern part of this area is urbanised, while the rest is mostly rural with some small villages. The selected single cell corresponds to the red dot in each subfigure, whereas yellow dots represent other cells. This cell is directional and points southwest.

Figures 4.1(d), (g), and (j), respectively, show the uniform, land use, and network priors for this area. Bright colors in these and the other subfigures correspond to relatively high values. The bright areas in Figures 4.1(g) correspond to the areas with buildings and roads (and therefore, where devices are expected to be). The bright areas in Figures 4.1(j) illustrate which areas have a high network coverage.

Two different connection likelihoods are shown in Figures 4.1(b) and (c). Since the cells are concentrated in the northwestern part of the area, the Voronoi region surrounding the selected cell has a particular shape in the direction of the rural area towards the southeast. Note that all values within this Voronoi region equal 1, which follows from Equation (15). The signal strength likelihood takes on high values in the south direction of the cell, where less overlap with other cells exists.

These three location priors and two connection likelihoods can be combined to form the six location posteriors shown in Figure 4.1(e), (f), (h), (i), (k), and (l). 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 Figure 4.1(l) illustrates Equation (9), that is, when the network prior is combined with the signal strength likelihood, the resulting posterior distribution is a rescaled version of the signal dominance.

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



(a) Base map

(b) Likelihood: Voronoi

(c) Likelihood: sig.str.

(d) Prior: uniform

(e) Post.: uniform/Vor.

(f) Post.: uniform/sig.str.

(g) Prior: land use

(h) Post.: land use/Vor.

(i) Post.: land use/sig.str.

(j) Prior: network

(k) Post.: network/Vor.

(l) Post.: network/sig.str.

5 Concluding remarks

We have proposed a Bayesian approach to estimate the location of mobile devices using mobile operator network data. 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, which for example uses a 3d model 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.

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.

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.

The quality of our method can be assessed by using other sources. MNOs often have coverage maps and best server maps, which are usually created with propagation models and measurements. One could compare these maps to those resulting from our methods. However, be aware that the coverage maps and best server maps may not represent the ground truth. More research is needed on how to interpret the (dis)similarities between such maps.

Small-scale validation is another approach to quality assessment. One method would be to log GPS location data for a sample of devices and to compare this data with the results of our methods based on mobile network data. This would require consent both from the MNO as well as from the device owners.

The results of our modular Bayesian approach can be useful for various applications, including official statistics. Development of statistical inference methods is required to use the results of our framework for the production of official statistics, such as daytime population statistics, commuting patterns and tourism.

Acknowledgements

First and foremost, we would like to thank our colleague Sander Scholtus from Statistics Netherlands, who contributed to the foundations of the method described in this paper. Also thanks to colleague Marco Puts, with whom we had many inspiring whiteboard discussions.

This work was supported by the ESSnet Big Data, a project within the European Statistical System. We would like to thank the project members, in particular Fabio Ricciato, David Salgado, Benjamin Sakarovitch, Roberta Radini, Tiziana Tuoto, and Sandra Hadam for the lively and insightful discussions as well as constructive feedback on earlier versions of this paper.

Also, many thanks go to several other colleagues from Statistics Netherlands who helped us to fine-tune the methods and to improve the manuscript: Harm Jan Boonstra, Ralph Meijers, Sigrid van Hoek, May Offermans, Edwin de Jonge, Jan van der Laan, and Marc Ponsen.

Finally, thanks to Jacob Wilkins from the University of Oxford for his useful feedback on the manuscript.

References

3GPP (Mar. 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, pages 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, pages 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, pages 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, pages 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, pages 920–940. DOI: `10.1177/0265813515600896`.

Figueiras, João and Simone Frattasi (2010). *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques*. John Wiley and Sons, Ltd.

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, page 1098.

Iqbal, Md Shahadat et al. (Mar. 2014). "Development of origin–destination matrices using mobile phone call data". In: *Transportation Research Part C: Emerging Technologies* 40, pages 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, pages 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. (Oct. 2016). "Accurate Radio Coverage Assessment Methods Investigation for 3G/4G Networks". In: *Computer Networks* 107.P2, pages 246–257. ISSN: 1389-1286.

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

Laan, D.J. van der and E. de Jonge (2019). "Determining an optimal time window for roaming data for tourism statistics". In: *Proceedings of the NetMob 2019 Conference*.

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, pages 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, pages 64–84.

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

Raitoharju, Matti, Simo Ali-Löytty, and Lauri Wirola (Dec. 2010). "Estimation of Base Station Position Using Timing Advance Measurements". In: *Proceedings of SPIE – The International Society for Optical Engineering*. Volume 8285.

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

Ricciato, F., G. Lanzieri, and A. Wirthmann (2019). "Towards a methodological framework for estimating present population density from mobile network operator data". In: *IUSSP Seminar on Digital Demography in the Era of Big Data*.

Ricciato, F. et al. (May 2016). "Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation". In: *Pervasive and Mobile Computing*.

S. Kung, Kevin, Stanislav Sobolevsky, and Carlo Ratti (Nov. 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 (Feb. 2009). "Path loss exponent estimation in large wireless networks". In: *2009 Information Theory and Applications Workshop*, pages 124–129.

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.

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

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

Wilson, R. et al. (Feb. 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, pages 133–165.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands

Design
Edenspiekermann

Enquiries
Telephone: +31 88 570 70 70
Via contact form: www.cbs.nl/infoservice