Report

Estimating hourly population flows in the Netherlands

Statistics Netherlands

March 2020

Contents

Summary

This report describes several methods developed in the ongoing project *Fast Population Statistics* by Statistics Netherlands (SN), as started in 2017. In collaboration with T-Mobile Netherlands SN used mobile phone network data to map population flows within the Netherlands on an hourly basis over periods of five weeks at the level of municipalities and districts. The estimation, aggregation and disclosure protection methods used for constructing these flows are explained and the results that have been obtained so far are summarised.

It is shown that by estimating the presence of people in a given region weekends, working weeks, major football matches, festivals and national holidays can be distinguished. Furthermore, business and commercial regions differ from residential ones by having lower counts on Wednesdays and Fridays and the time of day that the peak of population presence is reached. The relative changes over time of these population estimates hence are rich in information. When interpreted as absolute figures by themselves, however, the estimates can be improbable when compared to the known residential population.

Keywords

Mobile phone, device, signalling data, call detail records, cell, mobile network operator, Bayesian, location, mobility, origin-destination, population flow, inflow, outflow, dynamic population, Municipal Personal Records Database, T-Mobile Netherlands

1 Introduction

1.1 A different kind of population statistics

One of the most fundamental data sources government agencies in the Netherlands have available about population demographics is the Dutch *Municipal Personal Records Database* (PRD). Every municipality is obliged to register their residents in this database, along with their residential address and other demographic variables. Additionally, many other types of non-residents are (able to have themselves be) logged in the PRD (see [Pri17] for more details). While a faultless register of this type is impossible to create, regular quality monitoring ensures that the PRD is the most reliable and up-to-date source on, in particular, residential population counts.

However, as the complexity of society is increasing, so is the need by public policy makers for more fine-grained and high-dimensional data which could help to understand and address the issues that grow along with this trend. With regards to population counts, value has been found in not merely knowing where people reside, but also to learn their movements over time at a detailed spatial and temporal scale. Statistics on such *population flows* could help to improve urban planning, allocate public and emergency services more effectively, map exposure to environmental pollution and epidemic modelling, among other applications. An example pertaining more to less developed countries is the disaggregation and updating of out-of-date census data.

Making this idea more concrete: the figures describing such movements can be laid out in a so called *flow cube (of persons)* (also called an *origin-destination-time matrix* in the literature). It has three axes called *place of residence*, *place of presence* and *time*. The figure contained in the cube at such a triple of coordinates is the number of people present at a given place and time, grouped by their place of residence.

In this subsection or the next one some references to work by other researchers would be nice.

For example, if the smallest resolution of location used is 'districts' and the smallest resolution of time used is 'hours', then a selection of rows from the flow cube, presented in a 'long' format instead of a 'wide' format, could look as in Table 1.1. This table should be read as follows: on 4 June 2018 at 12:00 approximately, 439 people who have district A as their place of residence were also present in district A. At the same time, 19 people whose place of residence is district A were located in district B.

Table 1.1 An illustration of a selection of rows from a flow cube, presented in a 'long' format

Day	Hour	Place of residence	Place of presence	Count
20180604	12:00	district A	district A	439
20180604	12:00	district A	district B	19
20180605	17:00	district B	district C	671

Traditional surveys, such as the annual national mobility survey [SNf] carried out by Statistics Netherlands (SN), are by themselves not adequate for estimating population flows of this kind. The size of the sample necessary to produce such detailed figures with some level of reliability would prove to be too costly and place an excessive burden on the population. Moreover, surveys are highly reliant on either respondents' ability to memorise their daily movements, or their willingness to keep a detailed log, further increasing difficulties of (high quality) data collection.

1.2 Enter mobile phone network data

Since the ownership of mobile phones and the frequency and variedness of their use has expanded greatly over the last two decades, the international research community has turned to the resulting data generated by them as a potentially useful source for producing new kinds of population statistics, and enhancing existing ones. In the context of this report we do not mean the type of data collected from internal phone sensors, GPS or survey apps, but rather the network data automatically logged and stored by Mobile Network Operators (MNOs).

At this point an introduction of some terminology is in order. To enable telecommunication across a large geographic area, an MNO installs *base stations* at strategically chosen positions. Each base station contains one or more antennas, which in turn serve a relatively small area. In the literature these small land areas are sometimes called *cells*, but we will reserve that term for the antennas themselves, dropping the term *antenna*. These cells, together with the surrounding developed technology and infrastructure, comprise an MNOs *(mobile) network*.

When cellular devices of end-users, such as mobile phones, are switched on and present in the total coverage area of the MNOs network, they make regular connections to the network's cells. This holds both for devices subscribed to the particular MNO, and for ones roaming on the network. Which specific cell is connected to at a given moment depends on many factors: the geographic distance of the device to cells, the current load carried by cells caused by other nearby devices, obstructions such as buildings, trees and hills, and atmospheric conditions.

The connections made by all these devices over time are stored as records in massive tables by the MNO, known as the *signalling data*. They do not contain the actual contents of the communication. Some records are the result of deliberate activity by the user of the device, or that of the device at the other end of the communication. Examples include initiating or receiving voice calls, sending or receiving SMS messages and mobile data usage. These are converted by the MNO into *Call Detail Records* (CDR), which are then used for billing customers. Devices also generate records passively, for example, when they switch from one cell to another. Those are typically stored as well for the purpose of network analysis. One can expect a smartphone to generate hundreds of records in a signalling data table per day, of which fewer are generated at night. Besides the active usage by the user, the number of records depends on the brand and operating system of the device and the network technology (1G to 5G) used.

Its large volume, high spatiotemporal detail and low (additional) cost of collection make for attractive features of CDR and the richer signalling data to overcome the limitations of, or supplement traditional registers and surveys. However, even though this data only consists of communication metadata, it can reveal a great amount of information about device's owners. Protecting their privacy during data processing therefore imposes additional methodological challenges.

Any feedback on the last two sentences of this paragraph is appreciated.

Statistics Netherlands has had previous experiences in using mobile network data for statistical purposes. In the projects [SNc] and [SNd] SN worked with anonymised, aggregated datasets based on data obtained from the Dutch MNO Vodafone Netherlands by a mediating company. In 2017, a new pilot was initiated with the Dutch MNO T-Mobile Netherlands. Initially, a six month collaboration was started, which has been successfully continued ever since. It allowed SN a unique opportunity to study signalling data within this MNOs data centre, and export anonymised, aggregated datasets to SN's computer infrastructure for further analysis and processing. One of the direct goals was to use this data to construct flow cubes of persons for certain observation periods, while a long term ambition of SN is to develop an open and efficient methodology for deriving statistics based on signalling data, which could subsequently be implemented at other MNOs and (national statistical) institutes, both nationally and internationally.

1.3 Data sources used

The primary data source used to construct the flow cubes of persons was a table of 4G signalling data as stored in the big data environment of the MNO T-Mobile Netherlands. It should be stressed that this microdata was never transferred to SN – this report explains which datasets *were* exported to SN, and which disclosure control measures were taken before doing so. No distinction in the processing steps has been made between deliberately and passively generated records. They were treated on equal footing, based on the belief that a larger volume of records, regardless of their type, is likely to lead to better estimates.

Each record in this signalling data table contains over 300 variables. However, for our purposes, only the following three variables were used:

`imsi` standing for International Mobile Subscriber Identity (IMSI), which is an anonymous unique identifier of the device that generated the record. As a measure to protect privacy, neither researchers from the MNO nor from SN have direct access to this identifier, as the actual IMSI is partially hashed to the variable `imsi`,

`start_time` a time-stamp signifying the start of the interaction of the device with the cell which created the record,

`e_cgi` a unique identifier of the cell on the MNOs network with which the device interacted.

The rest of the variables may prove to be useful in future research, but understanding their nature will require substantial knowledge of the engineering aspects of the network.

The country of origin of a device can be extracted from the variable `imsi`, since the first three digits of the value of this variable are the country's mobile country code (MCC). This allows one to distinguish in particular records generated by Dutch devices from those generated by roaming devices.

The second data source provided by the MNO was an up-to-date *cell plan* of the network, which contains various physical properties and settings of the cells, including their geographical coordinates. The variables that were used are specified in Section 2.

The third data source is created by SN itself. Based on the PRD, SN periodically publishes figures on the number of residents at several administrative levels, such as municipalities, districts and neighbourhoods. To reduce the risk of disclosure the public figures on districts and neighbourhoods have rounding methods applied to them. We write $\mathrm{Pop}(x)$ for this publically available (at [SNe]) number of residents of administrative region x on 1 January 2017.

The fourth data source used was a rectangular grid divided into 100×100 metres square tiles which covers the Netherlands, including the Wadden Sea, the West Frisian Islands and a 25 kilometres offshore coastal zone. Specifically, the grid used the map projection known as 'Amersfoort / RD new', denoted by EPSG:28992.

The fifth data source was the digital geometry of the boundaries of all municipalities, districts and neighbourhoods in the Netherlands in 2017, as created by SN. This resource is publically available as well at [SNg] in the Esri shapefile format.

The sixth data source consisted of a *Current Dutch Elevation* file, which is a digital elevation map of the Netherlands. More precisely, the version of this map available at [PDOK] for 25×25 metres square tiles was used. Its construction is the result of a collaboration between the Dutch provinces, government and water boards.

1.4 Methodological outline

The processing pipeline which constructs flow cubes of persons starts with a Bayesian model which, given a connection of a device to a cell, estimates the geographic location of the device. The model considers each connection of a device individually – it does not take connections that are near in time into account. It is furthermore independent of the specific device. The model only uses the cell plan, the grid and the boundaries of administrative regions, as described in Section 1.3, as input data, and it does not make use of the signalling data. Therefore we consider this model as the *static* component of the pipeline. It is explained in Section 2.

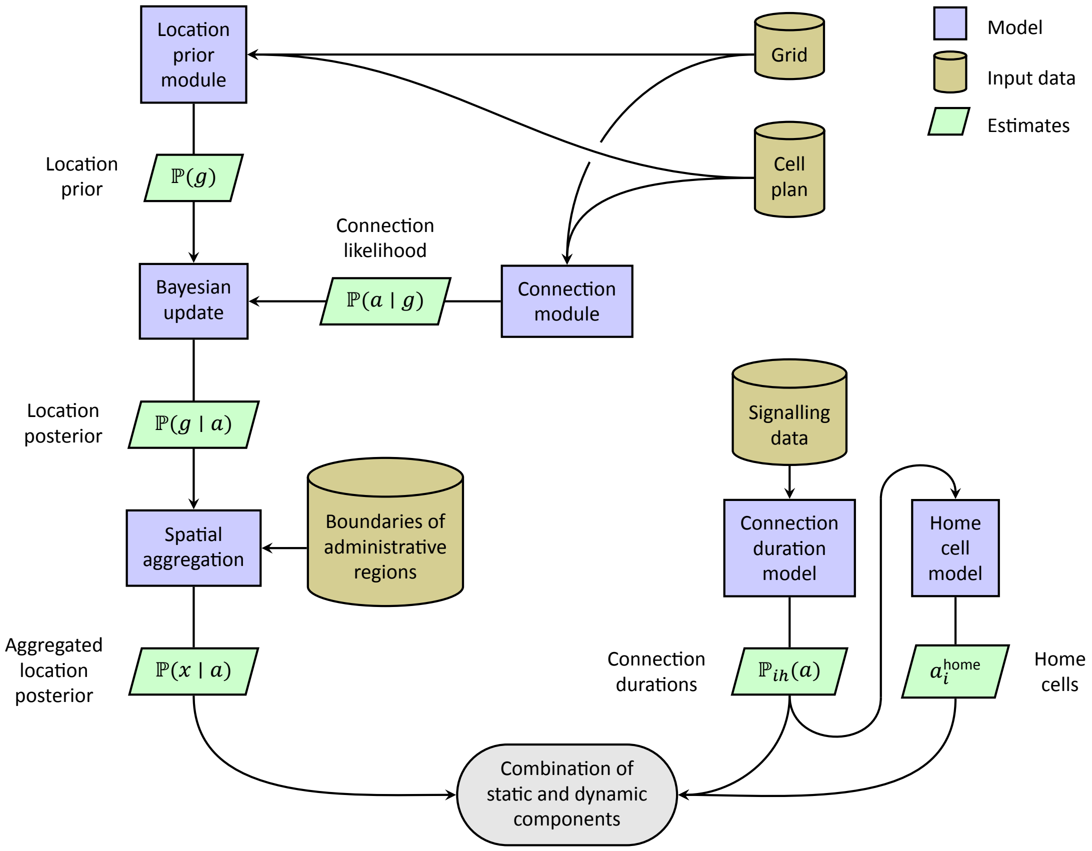
Independently of this static component a *dynamic* component has been built which is the only place in the pipeline where the signalling data *is* used. It estimates for every Dutch device and cell present in the signalling data, and every hour in the chosen observation period the fraction of the hour the device spent connected to the cell. These fractions are then used to estimate for every Dutch device its *home cell*, meaning (roughly) the cell to which it connected to the longest. This dynamic component is explained in Sections 3.1, 3.3 and 3.4.

After studying Section 2 and the aforementioned parts of Section 3 the reader might find it helpful to return to Figure 1.1 to see how the various methods fit together. The static component is outlined in the upper-left hand side of this diagram, while the dynamic component is summarised in the lower-right hand side.

The next step in the processing pipeline is the combination of the outputs of the static and dynamic components to estimate for every device in the signalling data probable places of residence and, for each hour, probable places of presence. (Here, the place of residence of a device refers to that of its owner.) These place estimates are then assembled into a *flow cube of devices*, which has the same axes as a flow cube of persons, but the figures now represent estimates of numbers of Dutch devices on the network of T-Mobile Netherlands making use of 4G technology.

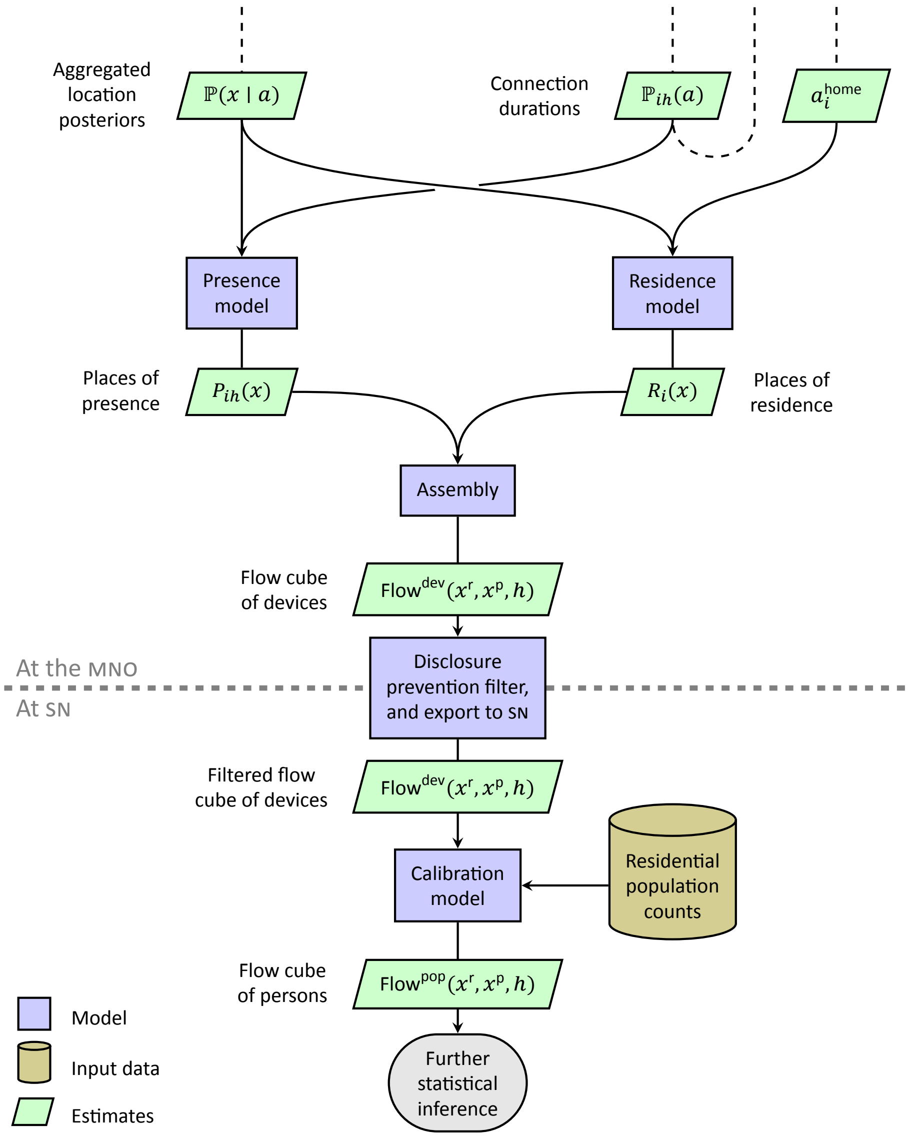
All processing steps explained so far took place inside the secured computer infrastructure of T-Mobile Netherlands. Before exporting the constructed flow cube of devices to SN a disclosure preventing filtering procedure is applied to it. The resulting filtered cube is finally *calibrated* into a flow cube of persons at SN, using the residential population counts based on the PRD. These steps are further explained in Sections 3.2, 3.3, 3.5 and 4. After studying those Sections the reader can return to Figure 1.2 for a visual summary. It should be noted that all methods involved naturally estimate numbers of devices or people as decimal numbers, instead of as integers, unlike the column 'count' in the example of Table 1.1 suggests.

Figure 1.1 The static and dynamic components of the processing pipeline



Examples of the information which can be gleaned from the flow cubes of persons we constructed are given in Section 5. Finally, Section 6 contains an outlook towards possible improvements on the developed methodology.

Figure 1.2 Combination of the static and dynamic components of the processing pipeline



2 Location estimation

Recall that two of the dimensions of a flow cube are spatial: place of residence and place of presence. However, a raw signalling data table contains no GPS data or demographic data on the owner of the device. An MNO has some demographic data on its customers – at least their name and postal address – but it is only allowed to join these with signalling data for the purpose of billing. Moreover, postal addresses of phones for business usage are frequently business addresses, and must therefore be considered unreliable. Constructing a flow cube of devices therefore requires one to approximate these places instead.

For this purpose a Bayesian model was developed which estimates the probability $\mathbb{P}(g \mid a)$ that a device is present in a grid tile g given its connection to some cell a. We have in particular that summing over all tiles g gives $\sum_g \mathbb{P}(g \mid a) = 1$, which reflects that a device (when connected to an cell) is located with probability 1 somewhere within the MNOs network range. As our notation of this probability already suggests, the model does not use or assume any information about the device, except that it is connected to a.

Our location estimation model uses Bayes' formula in the following way:

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

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 probability $\mathbb{P}(g)$ that a device is located in g without any connection knowledge represents the *location prior* about the relative frequency of connections made from g. Together they allow the calculation of $\mathbb{P}(g \mid a)$, which will henceforth be referred to as the *location posterior*.

The places of presence and residence of devices will be calculated at a coarser spatial level than that of grid tiles, namely in terms of neighbourhoods, districts and municipalities. The translation to such an administrative region x is done by defining the *aggregated location posterior* as

$$\mathbb{P}(x \mid a) := \sum_{g \in x} \mathbb{P}(g \mid a), \tag{2}$$

where we write $g \in x$ if the center of g lies in x. The reason to use the intermediary grid tiles is to allow for a more detailed modelling of location priors and connection likelihoods.

The rest of this section explains the models used for the location prior and connection likelihood. Further elaborations on, and extensions of the location estimation model can be found in [TGS20].

2.1 Connection likelihood

We model the connection likelihood $\mathbb{P}(a \mid g)$ as

$$\mathbb{P}(a \mid g) := \frac{s(g,a)}{\sum_{a' \in \mathcal{A}} s(g,a')}, \tag{3}$$

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 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 1.1. Note that $\mathbb{P}(a \mid g)$ is independent of rescaling the function $s(g,a)$ by a constant factor, and our convention is that $s(g,a)$ should be defined so as to take on values in the interval $[0,1]$. In Section 2.3 we propose a definition of the connection module by first approximating the signal strength $S(g,a)$ measured in dBm, followed by a transformation to obtain a signal dominance $s_{\text{strength}}(g,a)$. That is, this signal dominance will be an expression in terms of the approximated signal strength.

2.2 Location prior

Any model of the location prior will make assumptions on where devices are expected to be. In this section two options are discussed: the uniform prior and the network prior. More advanced models are presented in [TGS20].

Write $\mathcal{G}$ for the set of all tiles in the grid. When we use the *uniform prior*, we assume the probability of a device being in any grid tile is the same value for every tile:

$$\mathbb{P}_{\text{uniform}}(g) := \frac{1}{|\mathcal{G}|}. \tag{4}$$

In Bayesian statistics 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.

A *network prior* is defined as follows in terms of any choice for a model $s(g, a)$ for the signal dominance:

$$\mathbb{P}(g) := \frac{\sum_{a \in \mathcal{A}} s(g, a)}{\sum_{a \in \mathcal{A}} \sum_{g' \in \mathcal{G}} s(g', a)}. \tag{5}$$

This prior, together with the model (3) for the connection likelihood $\mathbb{P}(a \mid g)$ simplifies eq. (1) to

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

The interpretation of the prior (5) depends on the instance of the signal dominance $s(g, a)$ one has chosen. If the instance $s_{\text{strength}}(g, a)$, to be introduced in Section 2.3 is used, we will speak of *the* network prior and denote it by $\mathbb{P}_{\text{strength}}(g)$. Basically, $\mathbb{P}_{\text{strength}}(g)$ reflects the distribution of the total signal over the entire grid. This prior contains implicit knowledge about where an MNO is expecting people. The placement of cells namely 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. Usually, MNOs 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 keep in mind 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.

2.3 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 omnidirectional and normal cells (i.e. attached to cell towers or placed on rooftops) are directional [Kor+16].

2.3.1 Omnidirectional cells

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

$$S(g,a) := S_0 - S_{\text{dist}}(r_{g,a}), \tag{7}$$

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 center 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 [FF10]:

$$S_0 = 30 + 10\log_{10}(P). \tag{8}$$

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

$$S_{\text{dist}}(r) := 10\log_{10}(r^{\gamma}) = 10\gamma\log_{10}(r), \tag{9}$$

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 [SH09]. In free space, γ equals 2, which is what we used, but varying values would result in a more physically accurate model. Methods to vary the path loss exponent using land use data are explored in [TGS20].

2.3.2 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 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

$$S(g,a) := S_0 - S_{\text{dist}}(r_{g,a}) - S_{\text{azi}}(\delta_{g,a}, \alpha_a) - S_{\text{elev}}(\varepsilon_{g,a}, \beta_a), \tag{10}$$

where S_0 is the signal strength at $r_0 = 1$ meter distance from the cell, in the direction of the beam so that $\delta = 0$ and $\varepsilon = 0$. The signal loss due to distance to the cell, azimuth angle difference and elevation angle difference is specified by S_{dist}, S_{azi} and S_{elev}, respectively. The definition of S_{dist} is similar to the omnidirectional cell and can be found in eq. (9).

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

$$f(\varphi) := c - c\exp\left(-\frac{\varphi^2}{2\sigma^2}\right), \tag{11}$$

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 2.1. 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, and 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 2.2 (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.

2.3.3 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*.

Figure 2.1 Radiation patterns for the azimuth and elevation planes

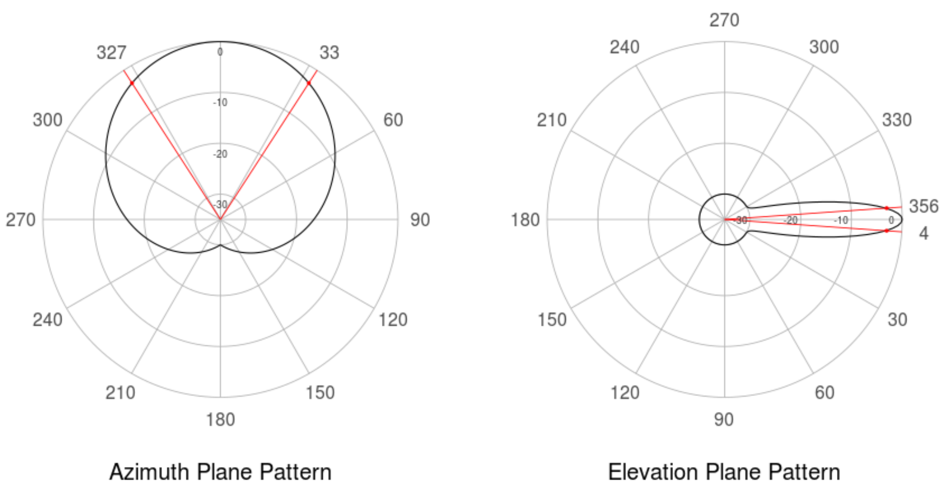
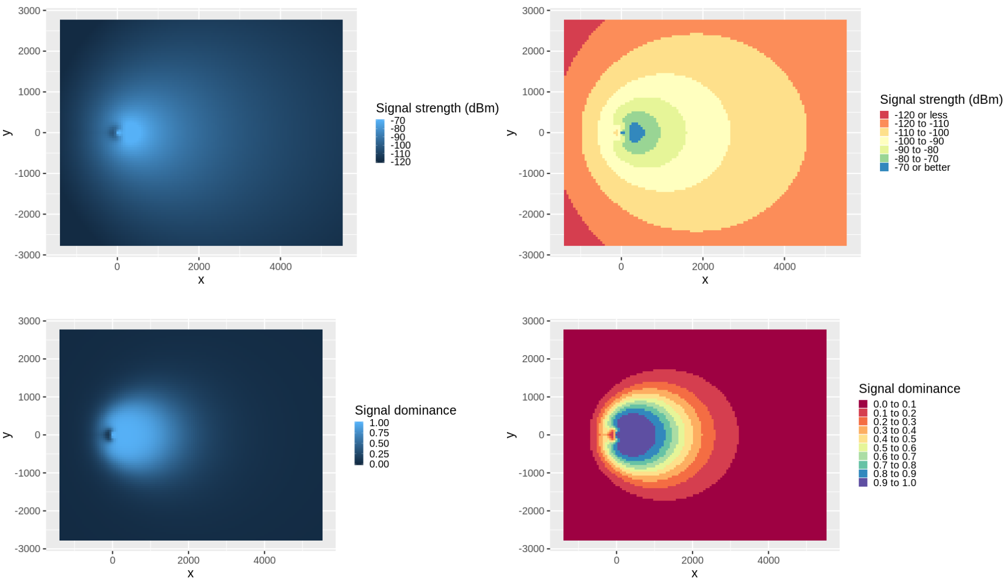


Figure 2.2 Signal strength (top row) and signal dominance (bottom row) at ground level



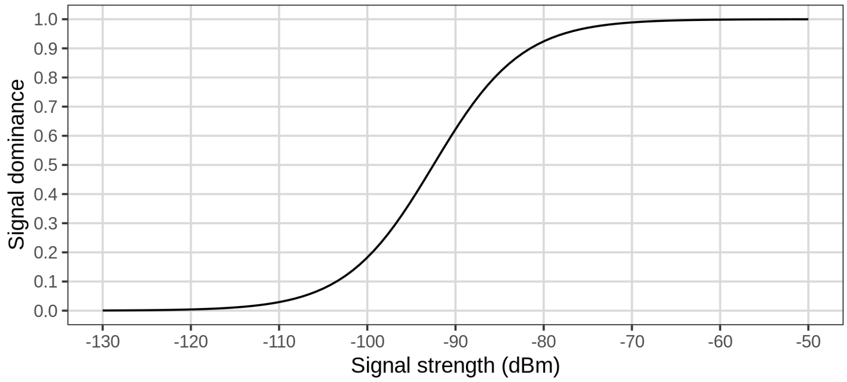
Our model allows for two phenomena that we feel should not be overlooked. The first is the switching of a device when it is receiving a bad signal to a cell with a better signal. Table 2.1 describes how the signal strength can be interpreted in terms of quality for 4G networks [Kor+16]. The second phenomenon 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.

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

Table 2.1 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 2.3 Logistic relation between signal strength and signal dominance



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 (3). Let us define

$$s_{\text{strength}}(g,a) := \frac{1}{1+\exp\left(-S_{\text{steep}}\left(S(g,a)-S_{\text{mid}}\right)\right)}, \tag{12}$$

where S_{mid} and S_{steep} are parameters that define the midpoint and the steepness of the curve respectively. Figure 2.3 shows an example of eq. (12). In that Figure S_{mid} and S_{steep} are set to −92.5 dBm and 0.2 dBm respectively to resemble Table 2.1. The signal dominance at ground level is shown in Figure 2.2 (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 2.2 (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.

3 Construction of the flow cube of devices

This Section starts by describing how the signalling data is used to estimate connection durations of observed devices. Together with a location posterior computed in Section 2 these durations serve as the input data for models which estimate places of presence (Section 3.2) and residence (Section 3.3) for all devices present in the signalling data. The assembly of these estimated places into a flow cube of devices is explained in Section 3.5, along with the privacy-protecting filtering procedure to make the cube suitable for export to SN.

3.1 Estimating connection durations

A device can make more than one connection per hour, even with the same cell. The number of connections per hour can moreover vary strongly between devices and different hours. Since we want to estimate population flows on an hourly basis, it is necessary to account in various computations for this phenomenon. More precisely: given a device i and an hour h, the estimated fraction of h that i spent connected to a cell a is an important quantity. We denote this fraction by $\mathbb{P}_{ih}(a)$ to suggest an alternative interpretation: it is the probability that i connected to a during h. It is estimated from the signalling data by

$$\mathbb{P}_{ih}(a) := \frac{\#\{\text{connections made by } i \text{ at } a \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}. \tag{13}$$

If i made no connections during h then we define $\mathbb{P}_{ih}(a)$ to be 0. If i made at least one connection during h, then obviously, $\sum_a \mathbb{P}_{ih}(a) = 1$.

3.2 Estimating a device's place of presence

To estimate the place of presence the probability distribution of the device i at hour h is defined as

$$P_{ih}(x) := \sum_a \mathbb{P}_{ih}(a) \cdot \mathbb{P}(x \mid a), \tag{14}$$

where x stands for an arbitrary administrative region and $\mathbb{P}(x \mid a)$ is a location posterior as defined in Section 2. If i made no connections during h then $\sum_x P_{ih}(x) = 0$. If i made at least one connection during h, then obviously $\sum_x P_{ih}(x) = 1$.

The definition of eq. (14) can be motivated as follows. We first approximate the fraction $P_{ih}(x)$ of h that i spent in x by

$$P_{ih}(x) \approx \frac{\#\{\text{connections made by } i \text{ from } x \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}. \tag{15}$$

This is of course a coarse approximation since it does not take the timestamps of the connections during h into account. In any case, we next bring cells into the picture by noting that the set in the numerator in eq. (15) is a disjoint union over all cells in the MNOs network:

$$\begin{aligned}&\{\text{connections made by } i \text{ from } x \text{ during } h\}\\ &\qquad = \bigsqcup_a \{\text{connections made by } i \text{ from } x \text{ at } a \text{ during } h\}.\end{aligned} \tag{16}$$

Therefore, eq. (15) can be expanded to

$$P_{ih}(x) \approx \sum_a \frac{\#\{\text{connections made by } i \text{ from } x \text{ at } a \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}. \tag{17}$$

The signalling data tells us how often i made an connection at a cell a during h, and given each such connection there is a probability $\mathbb{P}(x \mid a)$ that the connection was made from x. Hence, the numerator in each term in (17) can be estimated as

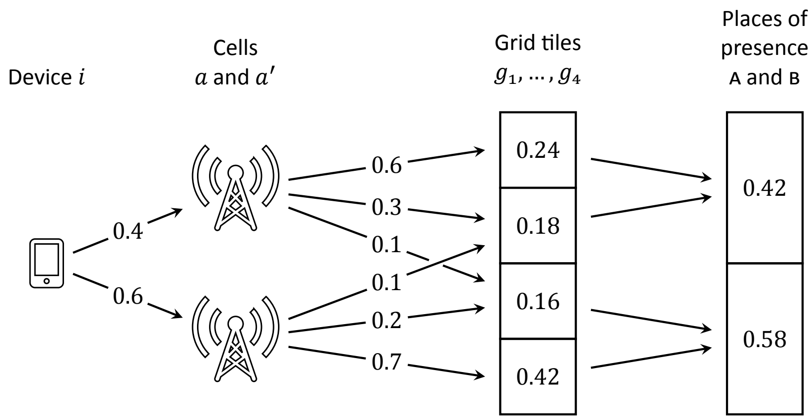
$$\begin{aligned} &\#\{\text{connections made by } i \text{ from } x \text{ at } a \text{ during } h\} \\ &\qquad \approx \#\{\text{connections made by } i \text{ at } a \text{ during } h\} \cdot \mathbb{P}(x \mid a), \end{aligned} \tag{18}$$

with which we arrive at the right-hand side of eq. (14).

Let us give an illustration of this model. Fix a device i and an hour h. Assume that the MNO's network consists of two cells a and a' and that the Netherlands is partitioned into four grid tiles $g_1, \dots, g_4$, of which g_1 and g_2 together form a place A, while B is the union of g_3 and g_4. The tiles g_1, g_2 and g_3 fall within the range of a, and g_2, g_3 and g_4 form the range of a'. All these objects are shown in Figure 3.1, ordered from top to bottom. The figures on the arrows from the cells to the grid tiles are examples of location posterior distributions.

Suppose that i connected to both a and a' during h, and that it follows from the signalling data that $P_{ih}(a) = 0.4$ and $P_{ih}(a') = 0.6$. The estimated probabilities for the place of presence during h across the grid tiles are then shown in the third column of Figure 3.1, while the fourth column lists the aggregated probabilities across the places A and B.

Figure 3.1 Illustration of the estimation of places of presence



3.3 Estimating a device's place of residence

To estimate the place of residence of a device i its *home cell* a_i^{home} is determined first, meaning (roughly) the cell to which i was connected to the longest during the observation period of five weeks. One might determine this starting by calculating for each cell the total number of hours connected to it by i over the entire period, then sorting the cells by these hour totals and finally selecting the top ranked. Due to some computational barriers, this process was implemented slightly differently, as explained next.

First, for each cell a the number of hours $h_{iw}^{\text{tot}}(a)$ connected to it by the device i was calculated per separate week w. This was done by summing the probabilities $\mathbb{P}_{ih}(a)$ associated to all hours in that week:

$$h_{iw}^{\text{tot}}(a) := \sum_{h \in w} \mathbb{P}_{ih}(a) \qquad \text{(a sum over } 24 \cdot 7 \text{ hours).} \tag{19}$$

To reduce the amount of memory needed to store all these hour counts, for each device only the top ten cells per week (in terms of connected hours) were preserved. The datasets per week were then combined into a single dataset by summing the number of hours per device and cell over all weeks in the observation period. Let us say that this period consists of, for example, five weeks:

$$h_i^{\text{tot}}(a) := \sum_w h_{iw}^{\text{tot}}(a) \qquad \text{(a sum over five weeks).} \tag{20}$$

It is assumed that the top ten per week will always include the cell the device connected to the most amount of hours during the five week period. Finally, the home cell a_i^{home} of device i is set to be the cell which maximises the number of hours $h^{\text{tot}}(i, a)$:

$$a_i^{\text{home}} := \arg\max_a h_i^{\text{tot}}(a) \qquad \text{(the top ranked of } 5 \cdot 10 \text{ non-unique cells).} \tag{21}$$

Given this home cell, the probability that the device i has a region x as its place of residence is determined using the aggregated location posterior distribution:

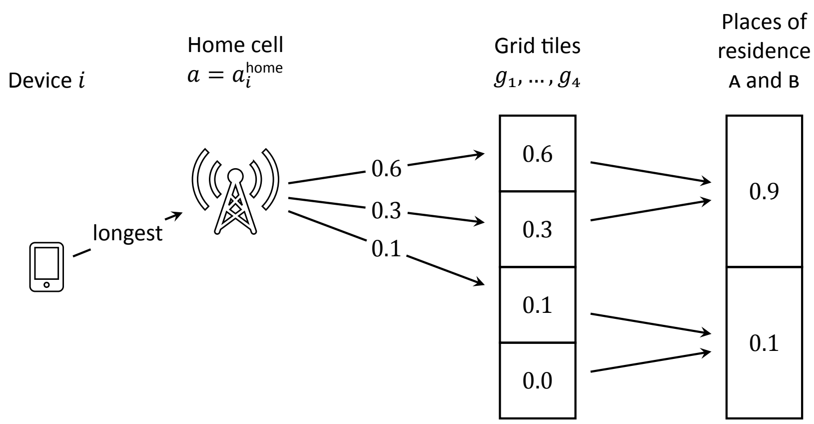
$$R_i(x) := \mathbb{P}(x \mid a_i^{\text{home}}). \tag{22}$$

Note that R_i is a probability mass function because $\mathbb{P}(\cdot \mid a_i^{\text{home}})$ is: summing over all regions x gives $\sum_x R_i(x) = 1$, which reflects that a device which has a home cell must have its place of residence somewhere within the MNOs network range.

An alternative approach might have been to consider the most frequently connected cell at nighttime. The advantage of the applied approach, however, is that it does not need a definition of 'nighttime' and therefore is unlikely to be biased against people with unusual night and day mobility patterns. A disadvantage, though, is that the devices of people who spend more time near their place of work than usual for the average fulltime job, possibly due to long working days or social activities, may receive an incorrect approximation of their place of residence.

An illustration of this model might again be helpful. We continue the example given in Section 3.2, and suppose that the home cell a_i^{home} of the device i has been determined to be the cell a. The estimated probabilities for the place of residence across the grid tiles are then shown in the third column of Figure 3.2, while the fourth column lists the aggregated probabilities across the places A and B.

Figure 3.2 Illustration of the estimation of places of residence



3.4 Removal of wandering devices

The method explained in Section 3.3 assigns to every device a home cell and hence a mass function of probable places of residence. However, it was found that about 25% of all devices present in the signalling data connected to their home cell for less than 60 hours during the observation period of five weeks. The probable places of residence derived from such an cell are considered unlikely to be near the true place of residence of the device's owner.

Give possible reasons why there are devices which don't cross this threshold: M2M, airplane mode, …

Within this project we did not study these *wandering devices* closer and decided to discard them before further processing. From this point onwards in this document 'all devices' will stand for only those which crossed this threshold of 60 hours.

3.5 Final assembly of the flow cube

Having estimates of a device's place of residence and, at every hour, place of presence now gives all the components to build the following 2-dimensional matrix of device i for hour h:

$$RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}) := R_i(x^{\mathrm{r}}) \cdot P_{ih}(x^{\mathrm{p}}). \tag{23}$$

It stores the probability that i has place of residence x^{r} (which is independent of h) and place of presence x^{p} during hour h. The multiplication in the above equation is based on the assumption that the probability mass functions of place of residence R_i and place of presence P_{ih} are independent from each other. If i made at least one connection during h, then RP_{ih} is a joint probability mass function: we have

$$\sum_{x^{\mathrm{r}}} \sum_{x^{\mathrm{p}}} RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}) = 1. \tag{24}$$

If i made no connections during h, then the above double sum equals 0.

At face value the assumption of independence might seem erroneous: surely knowledge of a device's place of residence should affect the probability of the place of presence? Our justification is based on interpreting R_i and P_{ih} as mass functions *conditioned on the signalling data of* i *during the observation period*. That is, it is assumed to be known what the home cell of i and the cells connected to during h are. Given this information, the functions R_i and P_{ih} are not spread out over the entirety of the Netherlands any longer, but are in practice each concentrated in a small number of regions. Our assumption is hence more precisely stated as independence of these two conditional probability mass functions. This implicit knowledge of the signalling data has been left out of notations for the sake of brevity.

Continuing the illustrations given in Sections 3.2 and 3.3, the result of the calculation of the matrix RP_{ih} is shown in Figure 3.3.

Figure 3.3 Illustration of the assembly of places of presence and residence

			Place of presence	
			A	B
			0.42	0.58
Place of residence	A	0.9	0.378	0.522
	B	0.1	0.042	0.058

The total flow cube of devices $\mathrm{Flow}^{\mathrm{dev}}$ is obtained by summing all matrices $RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}})$ over all devices i for each hour h:

$$\mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h) := \sum_{i} RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}). \tag{25}$$

All processing steps explained so far took place securely at the MNO T-Mobile Netherlands. Even though the signalling data was already anonymous, in the sense that records do not contain the name, address or demographic data of the owners of the devices which generated them, to further reduce a possible risk of disclosure the figures in the flow cube of devices strictly lower than 15 were deleted. The resulting filtered cube was then exported to SN.

At this point the cube can be used to estimate the *incoming* and *outgoing flow* of devices for a given region x and hour h. By this we mean the number of devices entering or leaving x from other places, possibly including x itself. They are computed respectively as follows:

$$\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{in}}(x, h) := \sum_{x^{\mathrm{r}}} \mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x, h), \tag{26}$$

$$\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{out}}(x, h) := \sum_{x^{\mathrm{p}}} \mathrm{Flow}^{\mathrm{dev}}(x, x^{\mathrm{p}}, h). \tag{27}$$

In other words, the inflow is calculated by summing over all places of residence, while the outflow is calculated by summing over all places of presence.

4 Construction of the flow cube of persons

The elements of the flow cube $\mathrm{Flow}^{\mathrm{dev}}$ of devices are estimates of numbers of devices. These figures differ from the corresponding numbers of persons for at least the following reasons:

– They represent merely counts of devices of the MNO from which the signalling data is obtained. The people who communicate via different MNOs are therefore excluded.
– Not everyone owns a mobile phone, or if they do, carry their devices everywhere with them. This holds especially for young children and the elderly.
– Some people might carry multiple devices with them. One can think of, for example, people carrying both a phone for work and one for personal use.

The cube of devices hence needs to be transformed or *calibrated* to a flow cube $\mathrm{Flow}^{\mathrm{pop}}$ of persons (with axes and dimensions equal to those of $\mathrm{Flow}^{\mathrm{dev}}$). Before proceeding to the description of the calibration method, note that from such a cube incoming and outgoing flows of persons can be calculated in the exact same way as from the flow cube of devices. They are denoted by $\mathrm{Flow}^{\mathrm{pop}}_{\mathrm{in}}(x, h)$ and $\mathrm{Flow}^{\mathrm{pop}}_{\mathrm{out}}(x, h)$, respectively, for every region x and hour h.

4.1 Calibration

Recall from Section 1.3 the definition of the residential population figures $\mathrm{Pop}(\cdot)$. Our calibration method was based on the assumption that the flow cube $\mathrm{Flow}^{\mathrm{pop}}$ of persons ought to satisfy the following combination of two equations for all places x^{r} and x^{p} and hours h:

$$\frac{\mathrm{Flow}^{\mathrm{pop}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)}{\mathrm{Flow}^{\mathrm{pop}}_{\mathrm{out}}(x^{\mathrm{r}}, h)} = \frac{\mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)}{\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{out}}(x^{\mathrm{r}}, h)}, \tag{28a}$$

$$\mathrm{Flow}^{\mathrm{pop}}_{\mathrm{out}}(x^{\mathrm{r}}, h) = \mathrm{Pop}(x^{\mathrm{r}}). \tag{28b}$$

If we momentarily leave out the reference to the hour h for brevity, eq. (28a) can be understood as the following claim:

> Suppose that a certain fraction of the residents of x^{r} is present in x^{p}. Then the same fraction of the MNOs devices with place of residence x^{r} is also present in x^{p}. The converse implication holds as well.

In other words, the first equation assumes a uniform presence of the MNOs devices in the flow of persons from location x^{r}. This homogeneity is not obvious, because, for example, the market share of the MNO in the flow from x^{r} to a location x^{p}_1 might differ from that in the flow to another location x^{p}_2. Further research is needed to quantify the bias resulting from this assumption.

The second equation (28b) results from the assumption

> The number of residents of x^{r} who are present in any of the regions x considered together equals the number of residents of x^{r}.

This assumption of course introduces a small error since the date (in our case 1 January 2017) for which the figure $\mathrm{Pop}(x^{\mathrm{r}})$ was determined is somewhat different from the observation period for which the signalling data was obtained. A larger error is introduced if the set of regions $\{x\}$ does not also include locations abroad. Residents of x^{r} might namely be abroad during (part of) the observation period. Correcting for this misestimation would involve additional tourism or holiday statistics, which we did not attempt at this stage of the project.

The two equations (28a) and (28b) are easily seen to be equivalent to the single equation

$$\mathrm{Flow}^{\mathrm{pop}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h) = \mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h) \cdot \frac{\mathrm{Pop}(x^{\mathrm{r}})}{\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{out}}(x^{\mathrm{r}}, h)}. \tag{29}$$

Written in this way all known variables are present on the right hand side, while the variable on the left hand side is the one we wish to compute. The factor calibrating the estimate $\mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)$ of a number of devices to the estimate $\mathrm{Flow}^{\mathrm{pop}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h)$ of a number of persons is hence defined to be the fraction

$$\frac{\mathrm{Pop}(x^{\mathrm{r}})}{\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{out}}(x^{\mathrm{r}}, h)} \tag{30}$$

and it is independent of the place of presence x^{p}. Note, moreover, that this calibration method does not require the actual figures in the flow cube of devices, but only the fractions on the right-hand side of eq. (28a) which are derived from this cube.

4.2 Example of calibration

This calibration method is best illustrated via an example. Suppose the Netherlands is partitioned into three regions A, B and C, having residential population figures according to the PRD Pop(A) $= 5\,000$, Pop(B) $= 750$ and Pop(C) $= 1\,000$, respectively. Fix an hour h and suppose that the corresponding 2-dimensional slice $\text{Flow}^{\text{dev}}(\cdot, \cdot, h)$ of the flow cube of devices looks as in Table 4.1. This table for example tells us that $\text{Flow}^{\text{dev}}(\text{A}, \text{B}, h) = 20$. We also added the column totals to this table, that is, the incoming flow $\text{Flow}^{\text{dev}}_{\text{in}}(\cdot, h)$, and the row totals $\text{Flow}^{\text{dev}}_{\text{out}}(\cdot, h)$ for each of the regions A, B and C.

Table 4.1 The slice $\text{Flow}^{\text{dev}}(\cdot, \cdot, h)$ at hour h of the flow cube of devices

		Place of presence			
		A	B	C	Total
Place of residence	A	900	20	80	1 000
	B	80	120	50	250
	C	70	40	140	250
	Total	1 050	180	270	

The residential population figures Pop($\cdot$) for the regions are higher than the row totals, by factors 5, 3 and 4, respectively. Correcting for this discrepancy via our method implies that the rows of the table above should be multiplied by these calibration factors. We then obtain the slice $\text{Flow}^{\text{pop}}(\cdot, \cdot, h)$ at hour h of the flow cube of persons as in Table 4.2. The row totals now equal the residential population figures Pop($\cdot$) and the column totals are the incoming flows of persons $\text{Flow}^{\text{pop}}_{\text{in}}(\cdot, h)$.

Table 4.2 The slice $\text{Flow}^{\text{pop}}(\cdot, \cdot, h)$ at hour h of the flow cube of persons

		Place of presence			
		A	B	C	Total
Place of residence	A	4 500	100	400	5 000
	B	240	360	150	750
	C	350	120	560	1 000
	Total	5 090	580	1 110	

5 Results

With the methodology explained in Sections 2 to 4 flow cubes of persons can be produced. The low technical complexity of the methods allow for efficient data processing with standard computational resources, but obvious questions about the quality of the output arise due to the simplifying assumptions that have been made. In this section several cubes produced by SN are evaluated for plausibility by checking which natural, known phenomena and events can or cannot be observed. This report does not aim to further quantify possible errors that have been found. A major hindrance for such a more detailed investigation is the lack of availability of benchmark data against which population flow estimates can be compared.

5.1 Flow cubes produced

In our discussion up to now x has stood for a generic administrative region, such as a neighbourhood, district or municipality. The choice for the level of spatial detail at which one is able to produce population flow statistics is restricted by at least the following factors:

- the spatial density of the cells belonging to the MNOs network. This density differs according to, among other factors, the level of urbanicity. Since a mobile network is optimised for the needs of its users, densely populated regions contain more cells to ensure optimal service,
- the accuracy of the model used to estimate the location posterior probabilities $\mathbb{P}(g \mid a)$,
- the mass lost from the cube by the threshold of 15 devices enforced before export to SN. A greater level of spatial detail namely implies that the cube will contain more cells, each of which is more likely to contain a lower number of devices,
- the market share of the MNO which made the signalling data available.

Consideration of the factors above led SN to construct three flow cubes, all for observation periods five weeks in length:

- one cube which used the uniform prior in the location estimation model, at the level of districts, for the period of 5 March up to and including 8 April 2018,
- two cubes which used the network prior, one at the level of districts and one for municipalities, for the period of 28 May up to and including 1 July 2018.

Because for certain durations in the former period the amount of available signalling data was insufficient, the corresponding figures in the flow cube were removed. Consequently, the analysed observation period started on 8 March.

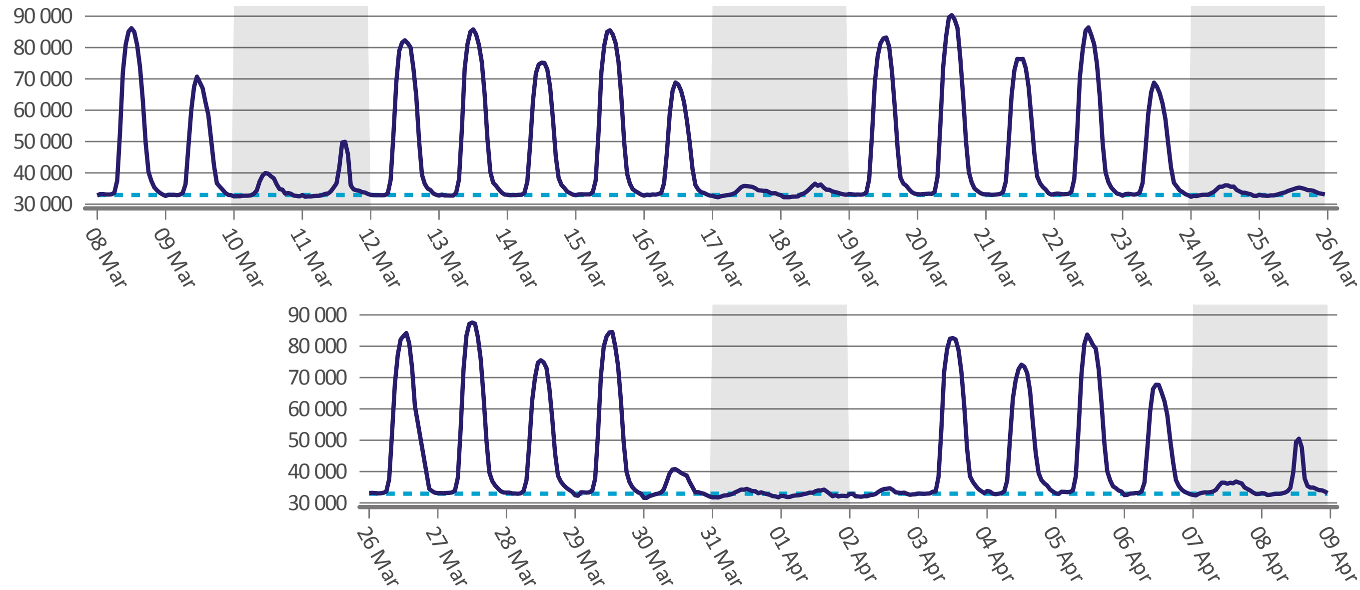
5.2 Visualisations of flows

Given a fixed region x, plotting the inflow $\text{Flow}_{\text{in}}^{\text{pop}}(x, h)$ for varying h results in what is known colloquially as the *daytime population graph* of x. We prefer to use the term *dynamic population* (DP) *graph* instead because it is defined for all hours. This graph shows how the number of people present in x (the *dynamic population*) varies over time, in contrast to the residential population which can be approximated to be constant over the short observation periods under consideration.

Figure 5.1 shows a DP graph of the district East of the municipality of Utrecht, along with the associated dashed, constant residential population graph. Shaded regions indicate weekends. This district is adjacent to the city centre, but stretches to the edge of the city, and includes many shops, the football stadium Galgenwaard and Utrecht Science Park – the largest science park of the Netherlands. It is home to university buildings, a university medical centre and student housing.

The graph shows that the dynamic population in this district doubles or triples at noon during workweek days with respect to the residential population. Far less people can be observed during weekends. Spikes can be seen, though, on the two Sundays 11 March and 8 April when the football club FC Utrecht played home matches at Galgenwaard. The graph additionally distinguishes Wednesdays and Fridays, on which employees often choose to work fewer hours and universities offer fewer classes. Good Friday fell on 30 March, while 2 April was Easter

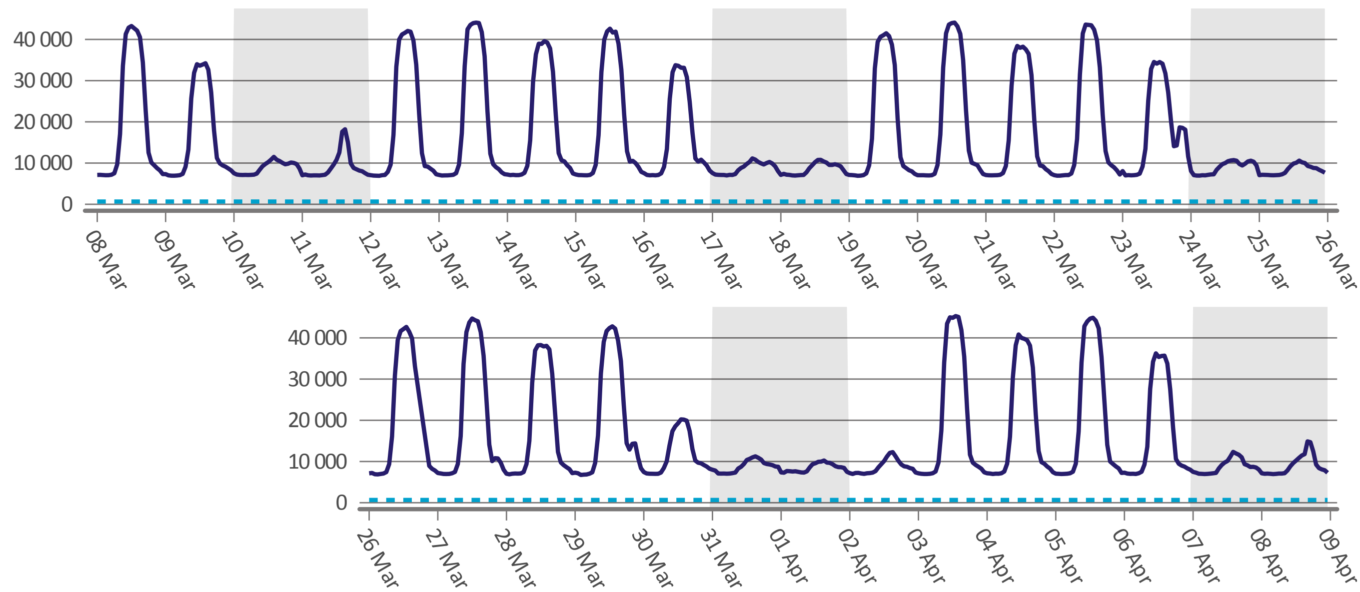
Figure 5.1   A DP graph of the district East of the municipality of Utrecht



Monday and Utrecht University and the HU University of Applied Sciences Utrecht were indeed both closed on these dates.

The estimated dynamic population at night in Utrecht East closely matches the residential population known from the PRD. Our methods do not result in such close matches for all districts, as demonstrated by the DP graph of the district Amstel III/Bullewijk of Amsterdam shown in Figure 5.2. This district lies on the far outskirts of the city and contains large retail stores, hotels, a university medical centre, the football stadium Johan Cruyff Arena, a concert hall and university examination halls.

Figure 5.2   A DP graph of the district Amstel III/Bullewijk of the municipality of Amsterdam



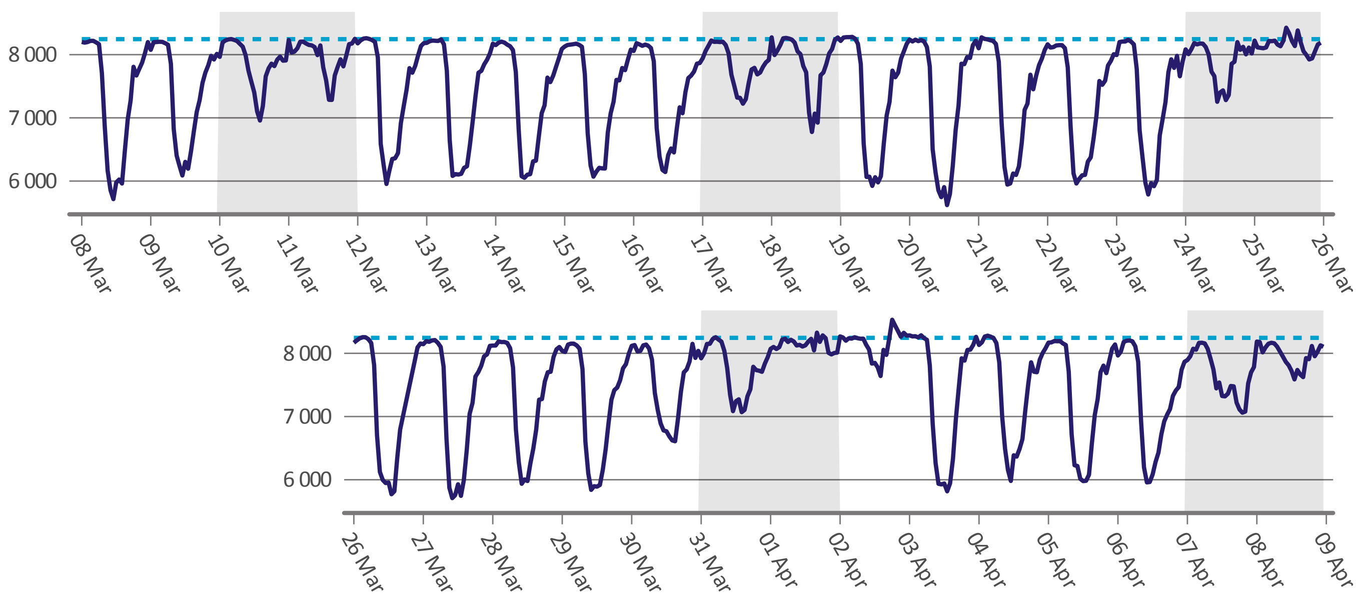
A very similar activity pattern can be observed here as for Utrecht East, including the spikes on 11 March and 8 April when the football club AFC Ajax played home matches at the Arena.

However, the residential population of this district was merely 665, while the level of the DP graph at night is higher with an difference of at least 6 000 people.

In this specific case such a difference can possibly explained by the presence of hotels, adjacent motorways and train and subway stations. There were nevertheless many districts for which we could not interpret such discrepancies. A possible explanation is that the location prior used, or the location estimation model in general, does not sufficiently respect the boundaries of districts. Indeed, while Amstel III/Bullewijk showed a surplus with respect to the residential population, the adjacent districts Holendrecht/Reigersbos and '00' of the municipalities of Amsterdam and Ouder-Amstel, respectively, show deficits. Football matches, being extremely localised in space and time and having well-documented visitor counts, provide additional evidence for this hypothesis. Stadiums in the Netherlands are namely often completely contained within a single district, but they are situated at their boundary. In our analysis it was observed that matches are sometimes not clearly visible in the DP graph of the stadium's district itself, but they *are* in the neighbouring one.

Figure 5.3 shows a typical DP graph of a residential area, as a comparison with the districts focused on commerce and education considered so far. Bavel is a village which has been absorbed as a district into the municipality of Breda, but is still slightly geographically separated from it. One observes that the lowest levels of the dynamic population are not reached at night, but at noon, as expected of districts of this type. It should be noted that the DP graphs of other commuter districts we analysed rarely matched the residential population this cleanly.

Figure 5.3 A DP graph of the district Bavel of the municipality of Breda

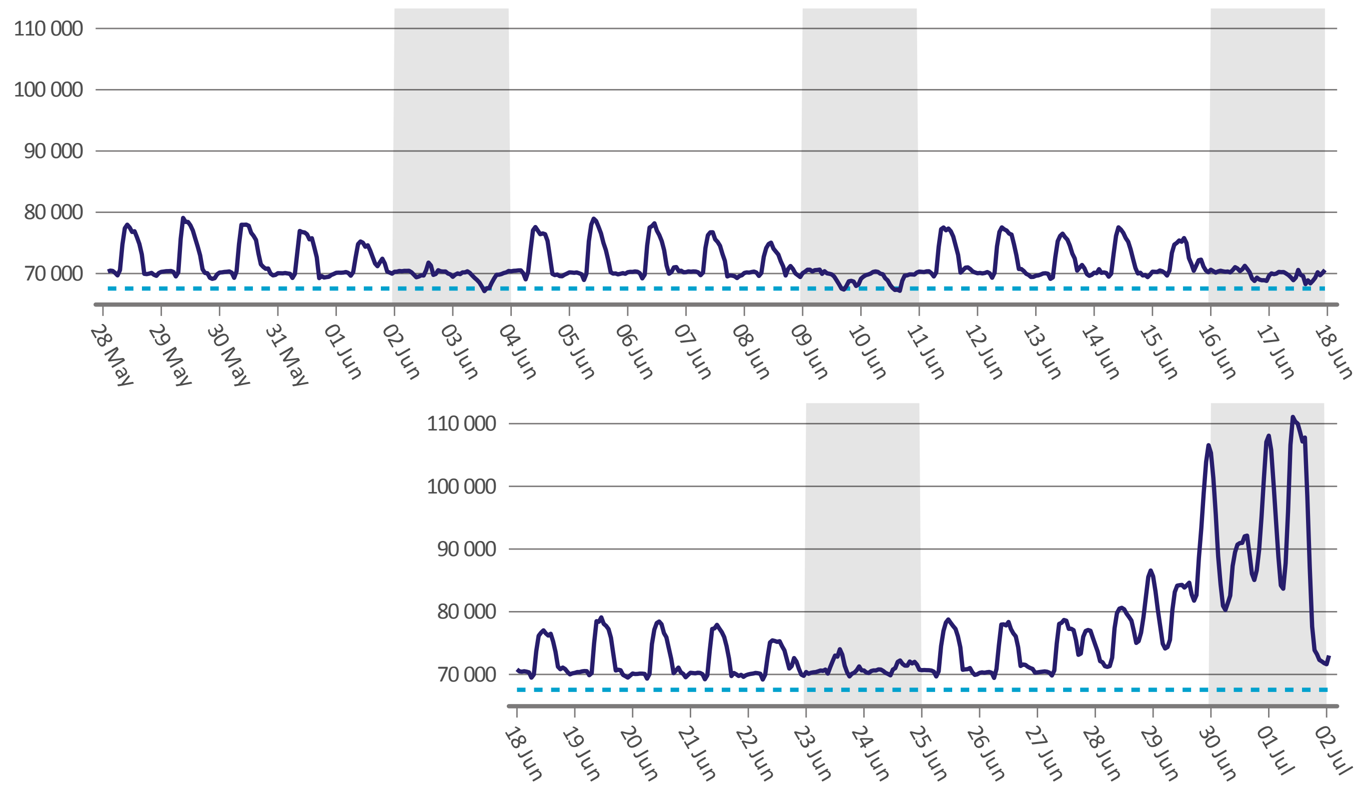


June traditionally is a month in which many summer festivals take place throughout the Netherlands. Such events are typically more spread out in space and time than football matches. In the hopes of capturing these well the flow cube for the June period was produced at the municipal level also, and it was simultaneously decided to switch from the uniform to the network prior.

The eight round of the 2018 Grand Prix motorcycle racing season took place on Sunday 1 July 2018 at the TT Circuit in the municipality of Assen. Surrounding events such as qualifying races started on Thursday 27 June, and the preceding TT Festival lasted from the evening of 27 June until the night of Saturday 30 June. These can be clearly read off from the DP graph of the municipality of Assen shown in Figure 5.4. It should be emphasised that we do not consider the (absolute) level of this graph to be reliable. That is, we do not claim to have measured the number of visitors at the events accurately.

What do you think of the last two sentences of the paragraph?

Figure 5.4 A DP graph of the municipality of Assen



Another example of an observed music festival is *Down the Rabbit Hole*, which took place in 2018 from the morning of Friday 29 June until Sunday night on 1 July, as seen in the DP graph of the municipality of Beuningen in Figure 5.5. The words of caution we gave about the events in Assen apply here as well.

A dashboard has been published by SN at [SNb] in which the flow cube for the June period at the municipal level can be explored interactively. It features DP graphs and a population density map, both of which can be animated simultaneously through control buttons. An icon in the upper left-hand corner shows hourly weather information to study the relation with population activities. Furthermore, a municipality can be selected after which the incoming and outgoing flows from and to the other Dutch municipalities are visualised as a fan of arrows. The screenshot in Figure 5.6 shows the (major) origins of the visitors of the *Boulevard Outdoor Festival* in the municipality of Wierden on Saturday 30 June 2018. The number of arrows in such fans is quite small as a consequence of the threshold of 15 devices that is enforced on the flow cube of devices before it is exported to SN.

We should actually be more precise by what we mean by 'quite small'. Small compared to what? However, this section is deliberately written in an informal style. This report makes no attempt to quantify the quality of our results. So I want to leave this as it is.

Figure 5.5 A DP graph of the municipality of Beuningen

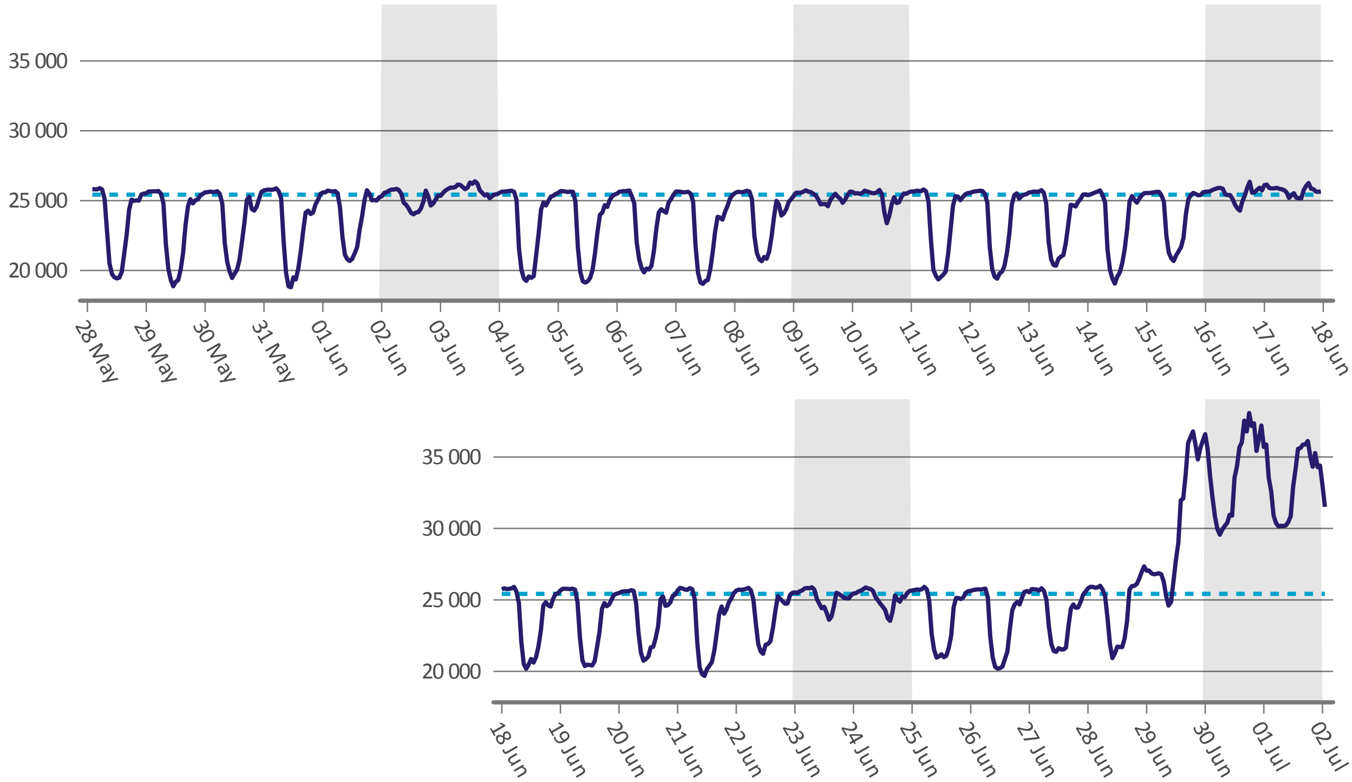
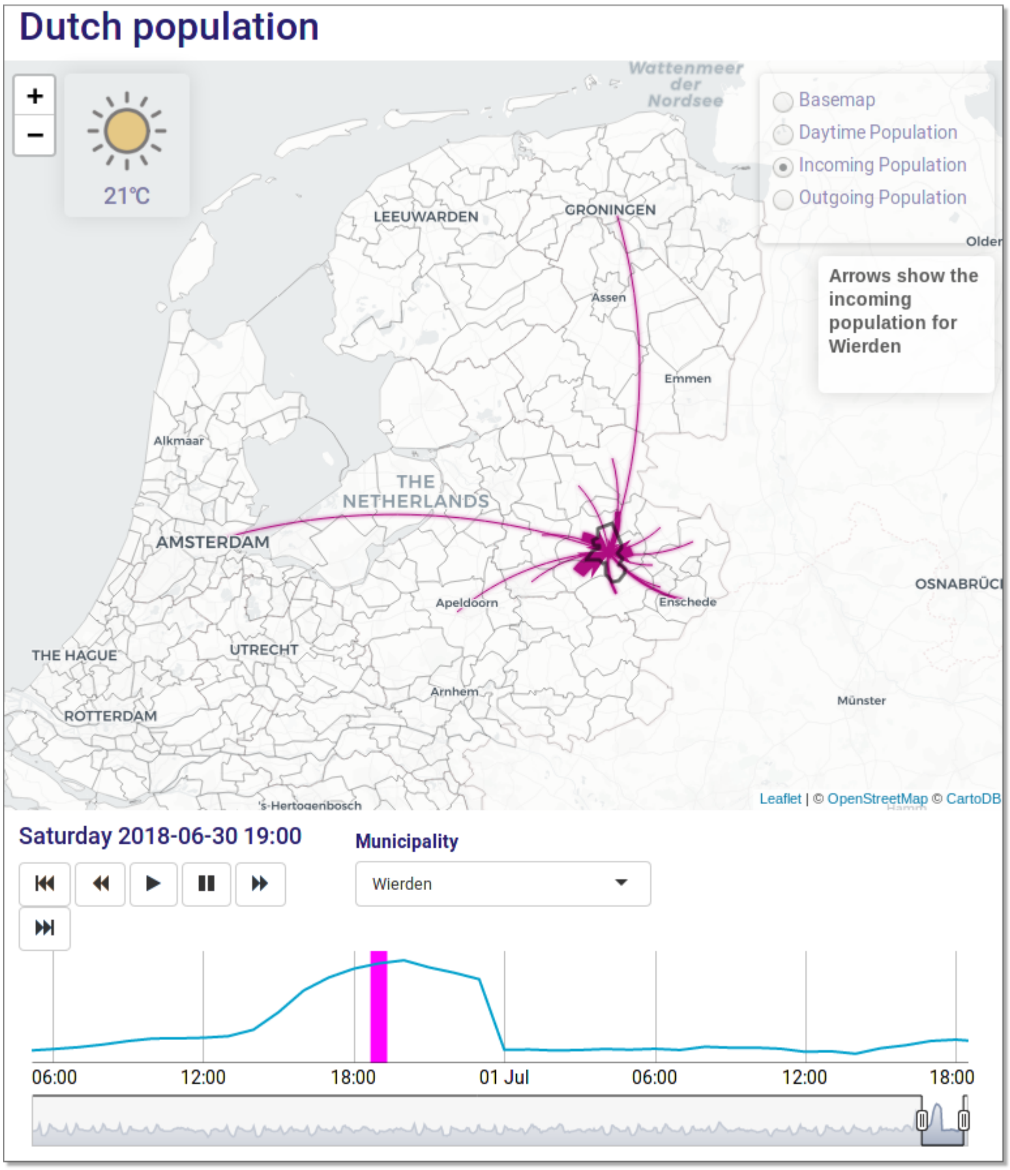


Figure 5.6 A screenshot of the population flow dashboard showing the municipality of Wierden



6 Summary and outlook

Section 5 shows that the relatively simple methods developed as part of the ongoing project *Fast Population Statistics* are already able to detect a variety of population activities at a high temporal detail resulting from

- workweeks and weekends,
- (in urban districts) Wednesday and Fridays versus the other days of the workweek,
- holidays, and
- football matches and festivals.

These activities can be read off by first producing flow cubes of persons and next observing the *relative* changes in the level of DP graphs and densities of population flow fans.

However, the currently developed technology does not seem to result in plausible *absolute* numbers in these cubes. We namely learned that DP graphs do not always match the residential population at night in cases where one does expect them to, and football matches tend to be visible not (only) in the district in which the stadium lies, but in neighbouring districts (as well). Because these discrepancies are lessened when the flow cubes are produced for municipalities instead of for districts we suspect that our estimation methods, while temporally precise, are spatially fairly 'blurry'. That is, the borders of the regions are not respected sufficiently. Further research will have to point out whether this is due to a spatial coarseness inherent in the mobile phone technology, or whether our methods can be refined by for example incorporating more information than we have done so far. For instance, replacing the uniform prior by the network prior in the location estimation model did not seem to improve the spatial blurriness, but in [TGS20] a *land use prior* is suggested which makes use of administrative data sources on land use to tune the geolocation of devices.

Other aspects of the methodology which can possibly be improved are

- obtaining access to signalling data from more than one Dutch MNO. This might reduce biases resulting from (demographic) differences in customer bases,
- investigating whether, besides 4G, also 3G and 2G signalling data can be used to increase the mass of records,
- if *timing advance data* is made available by the MNO, which is used to adjust for communication delays between base stations and devices, exploiting this to improve the geolocation of devices. Some ideas on this are discussed in [TGS20],
- developing methods which offer a measure of disclosure protection similar to the filtering threshold of 15 devices, but which result in less mass lost in the flow cube exported to SN,
- taking into account in the calibration method that a portion of the Dutch residents are abroad during the observation period. Auxiliary data, such as the *Continuous Holiday Survey* [SNa], might be used for this purpose,
- taking into account in the calibration method that not every Dutch resident owns a mobile phone and some people use multiple phones.

This report has focused on estimating the flows of Dutch residents within the Netherlands. However, signalling data may also be used for other purposes. As mentioned in Section 1.3, this data contains records generated by devices with other countries of origin as well, and the country can be read off the variable `imsi`. The techniques developed in the current report hence allow one to construct a flow cube of foreign devices, with the axis representing the place of residence replaced by the country of origin. Preliminary experiments of this type with only Belgian and German devices have been carried out at SN and showed plausible results. Calibrating such a cube to a flow cube of persons will provide a challenge, though.

The report [PGH19] describes recent research by SN on enrichening a flow cube of foreign devices with a fourth axis *cluster*. The project develops methods to group foreign devices into clusters using features derived from their signalling data via suitably chosen unsupervised learning algorithms. Identifying these clusters could help to separate commuters from tourists, further subdivide different types of tourists and, ultimately, provide Dutch policy makers and residents with new information on both the benefits and issues that arise from tourism. The methodology of the current report serves as a basis for such more detailed studies.

TiDeToSta folks should check this paragraph to see if it is correct.

Our Bayesian method inspired France. See City Data report.

References

[FF10] João Figueiras and Simone Frattasi. *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques*. John Wiley and Sons, Ltd., 2010.

[Kor+16] Ahmed D. Kora, Brice A. Elono Ongbwa, Jean-Pierre Cances and Vahid Meghdadi. 'Accurate Radio Coverage Assessment Methods Investigation for 3G/4G Networks'. In: *Computer Networks* 107 (9th Oct. 2016), pp. 246–257. ISSN: 1389-1286.

[PDOK] *Dataset: Actueel Hoogtebestand Nederland (AHN1)*. Dutch. Publieke Dienstverlening Op de Kaart (PDOK). URL: https://www.pdok.nl/introductie/-/article/actueel-hoogtebestand-nederland-ahn1- (visited on 12th Dec. 2019).

[PGH19] Marc Ponsen, Yvonne A.P.M. Gootzen and Sigrid van Hoek. *Towards Timely and Detailed Tourism Statistics via Clustering of Mobile Phone Data*. Statistics Netherlands, 31st Oct. 2019.

[Pri17] Kees Prins. *Population register data, basis for the Netherlands Population Statistics*. Statistics Netherlands, 2017. URL: https://www.cbs.nl/-/media/_pdf/2017/38/population-register-data.pdf (visited on 17th Oct. 2019).

[SH09] S. Srinivasa and M. Haenggi. 'Path loss exponent estimation in large wireless networks'. In: *2009 Information Theory and Applications Workshop*. Feb. 2009, pp. 124–129.

[SNa] *Continu Vakantie Onderzoek (CVO), vanaf 2017*. Dutch. Statistics Netherlands. URL: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/continu-vakantie-onderzoek--cvo---vanaf-2017 (visited on 17th Oct. 2019).

[SNb] *Dutch population*. Statistics Netherlands. URL: https://dashboards.cbs.nl/v1/dtp/ (visited on 13th Dec. 2019).

[SNc] *Hoeveel mensen zijn hier nu?* Dutch. Statistics Netherlands. URL: https://www.cbs.nl/nl-nl/onze-diensten/innovatie/project/hoeveel-mensen-zijn-hier-nu- (visited on 17th Oct. 2019).

[SNd] *How busy is it in Amsterdam on King's Day?* Statistics Netherlands. URL: https://www.cbs.nl/en-gb/our-services/innovation/project/how-busy-is-it-in-amsterdam-on-king-s-day- (visited on 17th Oct. 2019).

[SNe] *Kerncijfers wijken en buurten 2017*. Dutch. Statistics Netherlands. URL: https://www.cbs.nl/nl-nl/maatwerk/2017/31/kerncijfers-wijken-en-buurten-2017 (visited on 14th Oct. 2019).

[SNf] *Onderzoek Verplaatsingen in Nederland (OViN)*. Dutch. Statistics Netherlands. URL: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/onderzoek-verplaatsingen-in-nederland--ovin-- (visited on 17th Oct. 2019).

[SNg] *Wijk- en buurtkaart 2017*. Dutch. Statistics Netherlands. URL: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2017 (visited on 18th Oct. 2019).

[TGS20] Martijn Tennekes, Yvonne A.P.M. Gootzen and Shan H. Shah. 'A Bayesian approach to location estimation of mobile devices from mobile network operator data'. 2020. To be published.

[Val+19] Johan van der Valk, Martijn Souren, Martijn Tennekes, Shan Shah, May Offermans, Edwin de Jonge, Jan van der Laan, Yvonne Gootzen, Sander Scholtus and Anna Mitriaieva. 'Experiences of using anonymized aggregated mobile phone data in The Netherlands'. In: *City data from LFS and Big Data*. European Commission, 28th Feb. 2019. URL: https://ec.europa.eu/regional_policy/en/information/publications/studies/2019/city-data-from-lfs-and-big-data (visited on 17th Oct. 2019).

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