ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Workpackage I

Mobile Network Data

Deliverable I.7 (Experience with Real Data)

Some experimental results with mobile network data

Final Version, January 31, 2021

Prepared by:

- M. Suarez-Castillo (INSEE, France)
- M. Poulhes (INSEE, France)
- E. Coudin (INSEE, France)
- F. Sémécurbe (INSEE, France)
- R. Radini (ISTAT, Italy)
- T. Tuoto (ISTAT, Italy)
- S. Hadam (Destatis, Germany)
- N. Rosenski (Destatis, Germany)
- M. Tennekes (CBS, The Netherlands)
- S.H. Shah (CBS, The Netherlands)
- Y.A.P.M. Gootzen (CBS, The Netherlands)
- D. Salgado (INE, Spain)

Workpackage Leader:

David Salgado (INE, Spain)
david.salgado.fernandez@ine.es
telephone : +34 91 5813151
mobile phone : N/A

Contents

1

Introduction

The ultimate goal of this research project is the construction of a production framework to incorporate mobile network data into the production of official statistics. Thus, empirical evidence with real data is a key element in the whole research and any methodological, technological, or organizational proposal must be ultimately tested on real mobile network data. This document gathers the experience of four National Statistics Institutes (NSIs) with real data.

As described in deliverable I.1 of the present work package, the conditions to access real data are remarkably restricting and limit the course of action of NSIs even for research conditions. Any data processing proposal must be agreed in due advance with MNOs and changes in the statistical methodology and the software and hardware tools are very hard to introduce. The scope of the research is thus limited and severely framed by these conditions.

Indeed, as of this writing, even after initial contacts and experiences with encouraging results, access to data has been blocked and the proposed end-to-end statistical process depicted in the deliverables of this work package has not been possibly tested on real data. However, the experiences shared in the following chapters already contains different elements of this proposed end-to-end process. Hereafter, we present studies with real data conducted by CBS in The Netherlands, Destatis in Germany, ISTAT in Italy, and INSEE in France. They all contain invaluable feedback of diverse aspects to build a statistical process on real data. More research is needed in the future in parallel to searching a solution to access data in optimal privacy-preserving and research conditions.

2

Dutch population flows on municipality level

2.1. Introduction

2.1.1. 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 Prins (2017) 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 be useful for the analysis of commuting, tourism, and mobility in general. In the light of the COVID-19 pandemic, these statistics can potentially be used to monitor the impact of social restrictions.

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

The resolution of location to be used is 'municipality' and the resolution of time to be used is 'hours', such that a selection of rows from the flow cube, presented in a 'long' format instead of a 'wide' format, could look as in Table 2.1. This table should be read as follows: on 4 December 2020 at 12:00 approximately, 450 people who have municipality A as their municipality of residence were also present in municipality A. At the same time, 50 people whose municipality of residence is municipality A were located in municipality B.

Traditional surveys, such as the annual national mobility survey Statistics Netherlands (2017b) 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 both the respondents' ability to memorise their daily movements, and their willingness to keep a detailed log, further increasing difficulties of (high quality) data collection.

Table 2.1: An illustration of a selection of rows from a flow cube, presented in a 'long' format

Day	Hour	Municipality of residence	Municipality of presence	Count
20201204	12:00	municipality A	municipality A	450
20201204	12:00	municipality A	municipality B	50
20201205	17:00	municipality B	municipality C	650

Mobile phone data comes in various shapes and sizes. We distinguish app based data and network based data. The former might be based on GPS coordinates while the latter is data that is collected by MNO's. This report only discusses network based data for the purpose of constructing population flow cubes.

2.1.2. Enter mobile phone network data

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 locations. Each base station contains one or more cells, which in turn serve a relatively small area each. These cells, together with the surrounding developed technology and infrastructure, comprise an MNO's *(mobile) network*.

Two types of data are needed from the MNO for statistical inference on the municipality of devices: *cell plan* data, which contains information about the mobile communication network infrastructure, and *event* data. There are generally two types of event data, depending on which transaction events are contained. They do not contain the actual contents of the communication. Event data that are used to calculate the costs in order to bill customers are called *Call Detail Records* (CDR), which contain events related to active mobile phone use [1]. Event data that also contains passive events such as location updates, are called *signalling data*. These data are used by the MNO for network analysis and optimization. Signalling data are usually much richer than CDR data and are therefore recommended for statistical inference. 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 (2G to 5G) used. Though the pilot research was performed on signalling data, the methodologies presented in this report are applicable to CDR as well. To avoid confusion, we will use the term network event data henceforth, which can be interpreted as either CDR or signalling data. Any type of network event data contributes to a more detailed population flow cube.

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. Therefore additional disclosure control measures during data processing are necessary to ensure that their privacy is preserved.

Statistics Netherlands has previously collaborated with Dutch MNOs on several exploratory studies on mobile phone data. One of which was a pilot study that allowed SN a unique opportunity to study signalling data within an MNO's data centre, and export anonymised, aggregated datasets to SN's computer infrastructure for further analysis and processing (see van der Valk et al. (2019)). 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 network event data, which could subsequently be implemented at other MNOs and (national statistical) institutes,

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

both nationally and internationally.

Since the turn of the century many other NSIs, universities and research labs across the globe have started investigating how mobile phone network data can be used to help study human mobility. We refer to Chen et al. (2016) for a reasonably recent overview of the literature. It should be noted, though, that the notions of *origin-destination* and *flow* used in those papers, such as the early, influential Calabrese et al. (2011), typically do not agree exactly with the type of flow cubes that the current report is focused on.

2.1.3. Data sources used

The primary data source used to construct the flow cubes of persons was a table of 4Gnetwork event (signalling) data stored in the big data environment of one of the MNOs of the 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.

The following three variables were used from the network event data:

`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 MNO's network with which the device interacted.

Variables other than those mentioned above are not intended to be used in the estimation of the flow cube.

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 to be provided by each MNO is an up-to-date *cell plan* of the network, which contains various physical properties and settings of the cells, including their geographical coordinates. This data source is used in Section 2.3.

The third data source is created by SN itself. Based on the PRD, SN periodically publishes figures on the number of residents at the administrative level of municipality. To reduce the risk of disclosure the public figures on municipalities have rounding methods applied to them. We write $\text{Pop}(x)$ for this publicly available (at Statistics Netherlands (2017c)) number of residents of municipality x.

The fourth data source to be used is 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 uses the map projection known as 'Amersfoort / RD new', denoted by EPSG:28992.

The fifth data source is the digital geometry of the boundaries of all municipalities in the Netherlands, as created by SN. This resource is publicly available as well at Statistics Netherlands (2017d) in the Esri shapefile format.

The sixth data source consists 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 (2019) for 25×25 metres square tiles are used. Its construction is the result of a collaboration between the Dutch provinces, government and water boards.

2.2. Applied methodology

The processing pipeline to create the population flows is summarized in Figure 2.1. The underlying methodology is in many ways simpler than the methodology presented in Deliverable I3 of this work package WPI of the ESSNet Big Data II project. However, we have successfully illustrated how real MNO data can be used for an end-to-end statistical production process.

The notation used in this report follow Tennekes et al. (2020). Note that this notation differs from the notation used in the other reports of this project (in particular deliverable I3). It is often much simpler, because we only used the static approach for the geolocation algorithm. A conversion of notation is provided in Table 2.2. The terms and used notations are explained in the remainder of this section.

Table 2.2: Conversion of notation

Term	**This section**	**Deliverable I3**
Grid tile	g	T_i
Cell (record from event data)	a	E_d (for device d)
Auxiliary data	no general notation	$I(t)$ (for time t
Prior probability	$\mathbb{P}(g)$	$\mathbb{P}(T_{di}(t) \mid I(t))$
Likelihood prob. / event location	$\mathbb{P}(a \mid g)$	$\mathbb{P}(E_d \mid T_{di}, I(t))$
Posterior probability	$\mathbb{P}(g \mid a)$	$\mathbb{P}(T_{di} \mid E_d, I(t))$

The processing pipeline consists of a *static* and a *dynamic* component, depicted in Figure 2.1 with the two pink blocks.

In the static component, the location of a device is estimated. More precisely, for each cell a, the posterior probabilities $\mathbb{P}(g \mid a)$, i.e. the probability that a device is located in grid tile a when an event is generated at cell a, are calculated. Next, these probabilities are aggregated to municipalities. The applied model only uses the cell plan, the grid and the boundaries of administrative regions, as described in Section 2.1.3, as input data, and it does not make use of network event data. The output of the static component is a table with probabilities of presence in municipalities given a connection to a cell. It is further explained in Section 2.3.

Independently of this static component a *dynamic* component is the only place in the pipeline where the network event data *is* used. It estimates for every Dutch device and cell present in the network event data, and every hour in the chosen observation period of 30 days 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. The output of the dynamic component is a table with connection fractions of devices and cells. This dynamic component is explained in Sections 2.4.1, 2.4.3 and 2.4.4.

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 network event data probable municipalities of residence and, for each hour, probable municipalities of presence. (Here, the municipality of residence of a device refers to that of its owner.) These municipality estimates are then assembled into a *flow cube of devices*, where figures represent estimates of numbers of Dutch devices on the network of the MNO making use of the cellular network.

All processing steps explained so far take place inside the secured computer infrastructure of the MNO. 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 2.4.2, 2.4.3, 2.4.5 and 2.5.

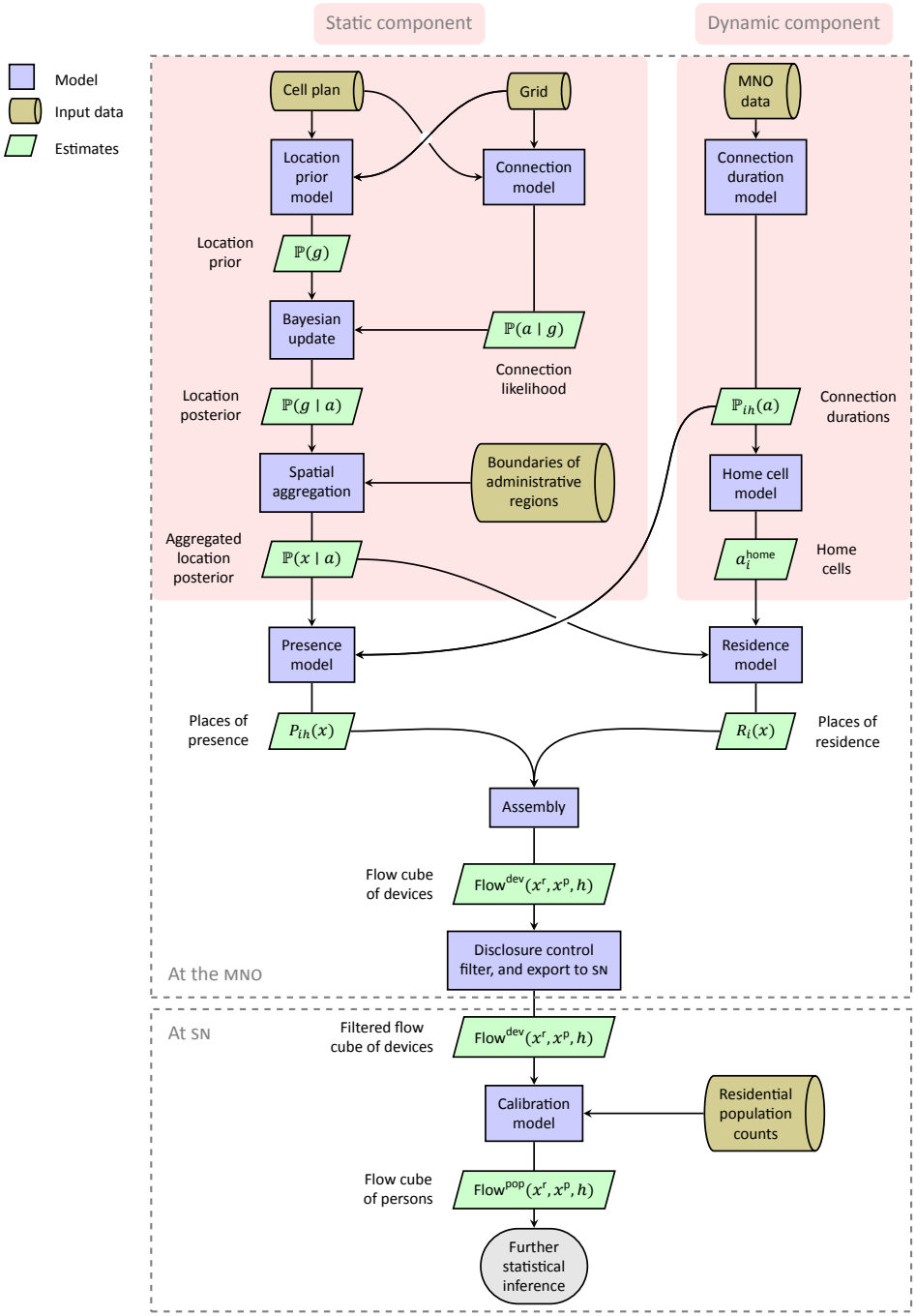


Figure 2.1: The processing pipeline

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 2.1 suggests.

2.3. Location estimation

Recall that two of the dimensions of a flow cube are spatial: municipality of residence and municipality of presence. However, a raw network event 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 network event 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

municipalities 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 a cell) is located with probability 1 somewhere within the MNO's 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:

$$\mathbb{P}(g \mid a) \propto \mathbb{P}(a \mid g)\mathbb{P}(g). \tag{2.1}$$

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 different methods that can be used for the connection priors and connection likelihoods are described in detail in Tennekes et al. (2020). In this work, a physical model of signal strength has been proposed for the connection likelihood, as an alternative to the Voronoi tessellation.

The municipalities of presence and residence of devices are calculated at a coarser spatial level than that of grid tiles, namely in terms of municipalities. The translation to a municipality x is done by defining the *aggregated location posterior* as

$$\mathbb{P}(x \mid a) := \sum_{g \in x} \mathbb{P}(g \mid a), \tag{2.2}$$

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.

2.4. Construction of the flow cube of devices

This Section starts by describing how the network event data are used to estimate connection fractions of devices. Together with a location posterior computed in Section 2.3 these fractions serve as the input data for models which estimate municipalities of presence (Section 2.4.2) and residence (Section 2.4.3) for all devices present in the network event data. The assembly of these estimated municipalities into a flow cube of devices is explained in Section 2.4.5, along with the privacy-protecting filtering procedure to make the cube suitable for export to SN.

2.4.1. Estimation of connection fractions

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 value. 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 network event data by

$$\mathbb{P}_{ih}(a) := \frac{\#\{\text{connections made by } i \text{ at } a \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}. \tag{2.3}$$

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 $\sum_a \mathbb{P}_{ih}(a) = 1$.

2.4.2. Estimation of municipality of presence of a device

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

$$P_{ih}(x) := \sum_a \mathbb{P}_{ih}(a) \cdot \mathbb{P}(x \mid a), \tag{2.4}$$

where x stands for an arbitrary municipality and $\mathbb{P}(x \mid a)$ is a location posterior as defined in Section 2.3. 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 Equation (2.4) can be motivated as follows. We first approximate the fraction $P_{ih}(x)$ of h that i spent in x by

$$P_{ih}(x) \approx \frac{\#\{\text{connections made by } i \text{ from } x \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}. \tag{2.5}$$

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 Equation (2.5) is a disjoint union over all cells in the MNO's network:

$$\begin{aligned} &\{\text{connections made by } i \text{ from } x \text{ during } h\} \\ &\quad = \bigsqcup_a \{\text{connections made by } i \text{ from } x \text{ at } a \text{ during } h\}. \end{aligned} \tag{2.6}$$

Therefore, Equation (2.5) can be expanded to

$$P_{ih}(x) \approx \sum_a \frac{\#\{\text{connections made by } i \text{ from } x \text{ at } a \text{ during } h\}}{\#\{\text{connections made by } i \text{ during } h\}}. \tag{2.7}$$

The network event 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 (2.7) can be estimated as

$$\begin{aligned} &\#\{\text{connections made by } i \text{ from } x \text{ at } a \text{ during } h\} \\ &\quad \approx \#\{\text{connections made by } i \text{ at } a \text{ during } h\} \cdot \mathbb{P}(x \mid a), \end{aligned} \tag{2.8}$$

with which we arrive at the right-hand side of Equation (2.4).

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 $g_1, \ldots, g_4$, of which g_1 and g_2 together form a municipality 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 2.2, 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 network event data that $P_{ih}(a) = 0.4$ and $P_{ih}(a') = 0.6$. The estimated probabilities for the municipality of presence during h across the grid tiles are then shown in the third column of Figure 2.2, while the fourth column lists the aggregated probabilities across the municipalities A and B.

2.4.3. Estimation of municipality of residence of a device

To estimate the municipality 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 30 days. One

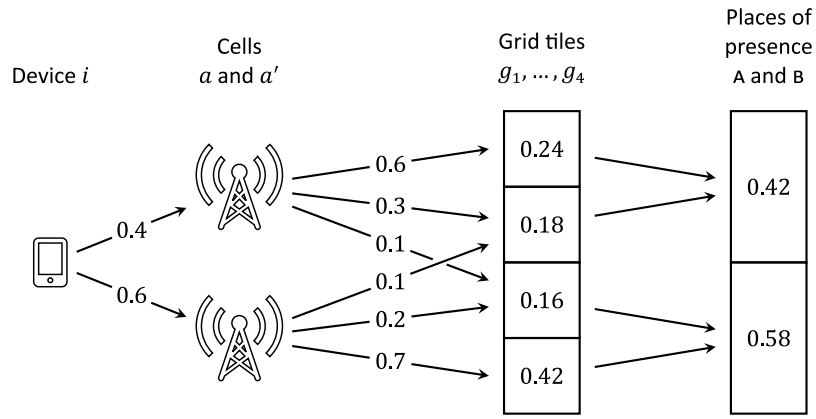


Figure 2.2: Illustration of the estimation of municipalities of presence

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 is 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 is calculated per subperiod of a number of days, depending on the available computational capacities. For simplicity's sake, let us say this period consists of 7 days so the number of connected hours is calculated per separate week w. This is done by summing the connection fractions $\mathbb{P}_{ih}(a)$ associated to all hours in that week:

$$h_{iw}^{\text{tot}}(a) := \sum_{h \in w} \mathbb{P}_{ih}(a) \qquad \text{(a sum over } 24 \cdot 7 \text{ hours)}. \tag{2.9}$$

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:

$$h_{i}^{\text{tot}}(a) := \sum_{w} h_{iw}^{\text{tot}}(a) \qquad \text{(a sum over 30 days)}. \tag{2.10}$$

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 30 day period. Finally, the home cell a_i^{home} of device i is set to be the cell which maximises the number of hours $h^{\text{tot}}(i, a)$:

$$a_i^{\text{home}} := \arg\max_a h_i^{\text{tot}}(a) \qquad \text{(the top ranked of combined non-unique cells)}. \tag{2.11}$$

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

$$R_i(x) := \mathbb{P}(x \mid a_i^{\text{home}}). \tag{2.12}$$

Note that R_i is a probability mass function because $\mathbb{P}(\cdot \mid a_i^{\text{home}})$ is: summing over all municipalities x gives $\sum_x R_i(x) = 1$, which reflects that a device which has a home cell must have its municipality of residence somewhere within the MNO's network range. Technically, all non-municipality regions (such as Belgium or Germany) must be viewed as one big municipality for $\sum_x R_i(x) = 1$ to hold. From this step, a municipality of residence will remain per device per 20 day period. All other intermediate data, such as lists of top 10 cell connections are no longer needed and are not saved.

An alternative approach might be to consider the most frequently connected cell at nighttime. The advantage of the proposed 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 municipality of residence.

An illustration of this model might again be helpful. We continue the example given in Section 2.4.2, 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 municipality of residence across the grid tiles are then shown in the third column of Figure 2.3, while the fourth column lists the aggregated probabilities across the municipalities A and B.

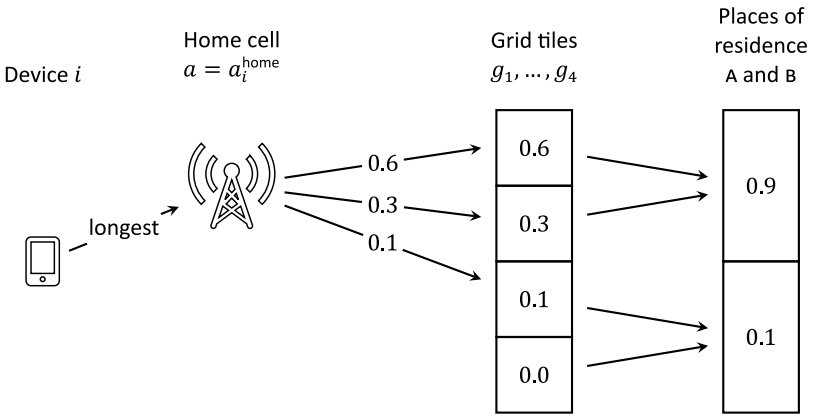


Figure 2.3: Illustration of the estimation of municipalities of residence

2.4.4. Data cleaning

The relation between observed devices and a flow cube of persons is not one-to-one. Investigation is needed during this data cleaning step.

For example, the method explained in Section 2.4.3 assigns to every device a home cell and hence a mass function of probable municipalities of residence. However, it was found in the pilot study that the distribution across all Dutch devices of the fraction of the observation period (which was five weeks instead of 30 days) that devices spent connected to their home cell was bimodal. The two peaks were best separated by dividing this distribution at 60 hours of connection. About 25% of all devices present in the network event data then connected to their home cell for less than 60 hours during the observation period. The probable municipalities of residence derived from such a cell are considered unlikely to be near the true municipality of residence of the device's owner. We have dubbed these devices in the pilot as *wandering devices* and have not studied them further and decided to discard them before further processing. From this point onwards in this document 'all devices' will stand for only those which crossed the threshold of hours. This threshold should be determined again since the observation period will be 30 days instead of five weeks.

2.4.5. Final assembly of the flow cube

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

$$RP_{ih}(x^{\text{r}}, x^{\text{p}}) := R_i(x^{\text{r}}) \cdot P_{ih}(x^{\text{p}}). \tag{2.13}$$

It stores the probability that i has municipality of residence x^{r} (which is independent of h) and municipality of presence x^{p} during hour h. The multiplication in the above equation is based on the assumption that the probability mass functions of municipality of residence R_i and municipality 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

$$\sum_{x^{\mathrm{r}}} \sum_{x^{\mathrm{p}}} RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}) = 1. \tag{2.14}$$

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 municipality of residence should affect the probability of the municipality of presence? Our justification is based on interpreting R_i and P_{ih} as mass functions *conditioned on the network event 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 network event data has been left out of notations for the sake of brevity.

Continuing the illustrations given in Sections 2.4.2 and 2.4.3, the result of the calculation of the matrix RP_{ih} is shown in Figure 2.4.

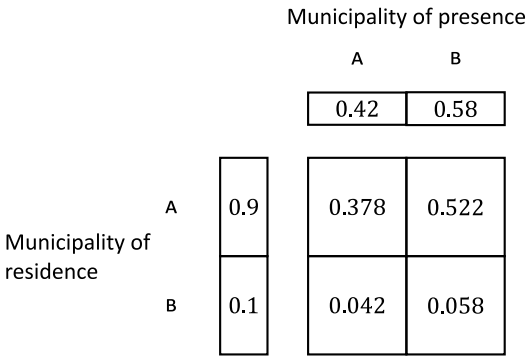


Figure 2.4: Illustration of the assembly of municipalities of presence and residence

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:

$$\mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x^{\mathrm{p}}, h) := \sum_i RP_{ih}(x^{\mathrm{r}}, x^{\mathrm{p}}). \tag{2.15}$$

All processing steps explained so far will take place securely at the MNO. Even though the network event data now only consists of the total number of devices counted per municipality, and does not contain any individual record, to further reduce a possible risk of disclosure the figures in the flow cube of devices strictly lower than 15 are to be deleted. The resulting filtered cube is then to be exported by the MNO via a secured connection to SN, once every 24 hours. The MNO deletes this data after exporting it to SN.

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

$$\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{in}}(x, h) := \sum_{x^{\mathrm{r}}} \mathrm{Flow}^{\mathrm{dev}}(x^{\mathrm{r}}, x, h), \tag{2.16}$$

$$\mathrm{Flow}^{\mathrm{dev}}_{\mathrm{out}}(x, h) := \sum_{x^{\mathrm{p}}} \mathrm{Flow}^{\mathrm{dev}}(x, x^{\mathrm{p}}, h). \tag{2.17}$$

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

2.5. Construction of the flow cube of persons

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

- The MNO that delivered the data does not have a 100% market share, in other words, only the devices from this MNO have been registered.
- 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 Flow^{pop} of persons (with axes and dimensions equal to those of Flow^{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 $\text{Flow}^{\text{pop}}_{\text{in}}(x, h)$ and $\text{Flow}^{\text{pop}}_{\text{out}}(x, h)$, respectively, for every municipality x and hour h.

2.5.1. Calibration

Recall from Section 2.1.3 the definition of the residential population figures $\text{Pop}(\cdot)$, based on the PRD. Our calibration method is based on the assumption that the flow cube Flow^{pop} of persons ought to satisfy the following combination of two equations for all municipalities x^{r} and x^{p} and hours h:

$$\frac{\text{Flow}^{\text{pop}}(x^{\text{r}}, x^{\text{p}}, h)}{\text{Flow}^{\text{pop}}_{\text{out}}(x^{\text{r}}, h)} = \frac{\text{Flow}^{\text{dev}}(x^{\text{r}}, x^{\text{p}}, h)}{\text{Flow}^{\text{dev}}_{\text{out}}(x^{\text{r}}, h)}, \tag{2.18a}$$

$$\text{Flow}^{\text{pop}}_{\text{out}}(x^{\text{r}}, h) = \text{Pop}(x^{\text{r}}). \tag{2.18b}$$

If we momentarily leave out the reference to the hour h for brevity, Equation (2.18a) 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 MNO's devices with municipality 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 devices in the flow of persons from municipality x^{r}. This homogeneity is not obvious, because, for example, the age distribution within a municipality x^{r} might influence the fraction of people that travel with a mobile device compared to municipality x^{p}_2. Further research is needed to quantify the bias resulting from this assumption.

The second equation (2.18b) results from the assumption "The number of residents of x^{r} who are present in any of the municipalities x considered together equals the number of residents of x^{r}". This assumption introduces a small error since the date (in our case the date of the most recent population registry) for which the figure $\text{Pop}(x^{\text{r}})$ was determined is somewhat different from the observation period of the network event data. A larger error is introduced if the set of municipalities $\{x\}$ 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 is not planned for at this stage of the project. The assumption used hence results in a systematic overestimation of person flows.

The two equations (2.18a) and (2.18b) are easily seen to be equivalent to the single equation

$$\text{Flow}^{\text{pop}}(x^{\text{r}}, x^{\text{p}}, h) = \text{Flow}^{\text{dev}}(x^{\text{r}}, x^{\text{p}}, h) \cdot \frac{\text{Pop}(x^{\text{r}})}{\text{Flow}^{\text{dev}}_{\text{out}}(x^{\text{r}}, h)}. \tag{2.19}$$

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 $\text{Flow}^{\text{dev}}(x^{\text{r}}, x^{\text{p}}, h)$ of a number of devices to the estimate $\text{Flow}^{\text{pop}}(x^{\text{r}}, x^{\text{p}}, h)$ of a number of persons is hence defined to be the fraction

$$\frac{\text{Pop}(x^{\text{r}})}{\text{Flow}^{\text{dev}}_{\text{out}}(x^{\text{r}}, h)} \tag{2.20}$$

and it is independent of the municipality 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 Equation (2.18a) which are derived from this cube.

Before publishing, each value of $\text{Flow}^{\text{pop}}(x^{\text{r}}, x^{\text{p}}, h)$ below 50 is removed from the data and all other values are rounded to the nearest multiple of 50.

2.5.2. Example of calibration

This calibration method is best illustrated via an example. Suppose the Netherlands is partitioned into three municipalities A, B and C, having residential population figures according to the PRD $\text{Pop}(\text{A}) = 5000$, $\text{Pop}(\text{B}) = 750$ and $\text{Pop}(\text{C}) = 1000$, 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 2.3. 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 municipalities A, B and C.

Table 2.3: The slice $\text{Flow}^{\text{dev}}(\,\cdot, \cdot, h)$ at hour h of the flow cube of devices

		Municipality of presence			
		A	**B**	**C**	**Total**
Municipality of residence	**A**	900	20	80	1000
	B	80	120	50	250
	C	70	40	140	250
	Total	1050	180	270	

The residential population figures $\text{Pop}(\cdot)$ for the municipalities 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 remove any counts lower than 50 and round to multiples of 50 to obtain the slice $\text{Flow}^{\text{pop}}(\,\cdot, \cdot, h)$ at hour h of the flow cube of persons as in Table 2.4. The row totals now equal the residential population figures $\text{Pop}(\cdot)$ and the column totals are the incoming flows of persons $\text{Flow}^{\text{pop}}_{\text{in}}(\,\cdot, h)$.

Table 2.4: The slice $\text{Flow}^{\text{pop}}(\,\cdot, \cdot, h)$ at hour h of the flow cube of persons

		Municipality of presence			
		A	**B**	**C**	**Total**
Municipality of residence	**A**	4500	100	400	5000
	B	250	350	150	750
	C	350	100	550	1000
	Total	5100	550	1100	

2.6. Results

With the methodology explained in Sections 2.3 to 2.5 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.

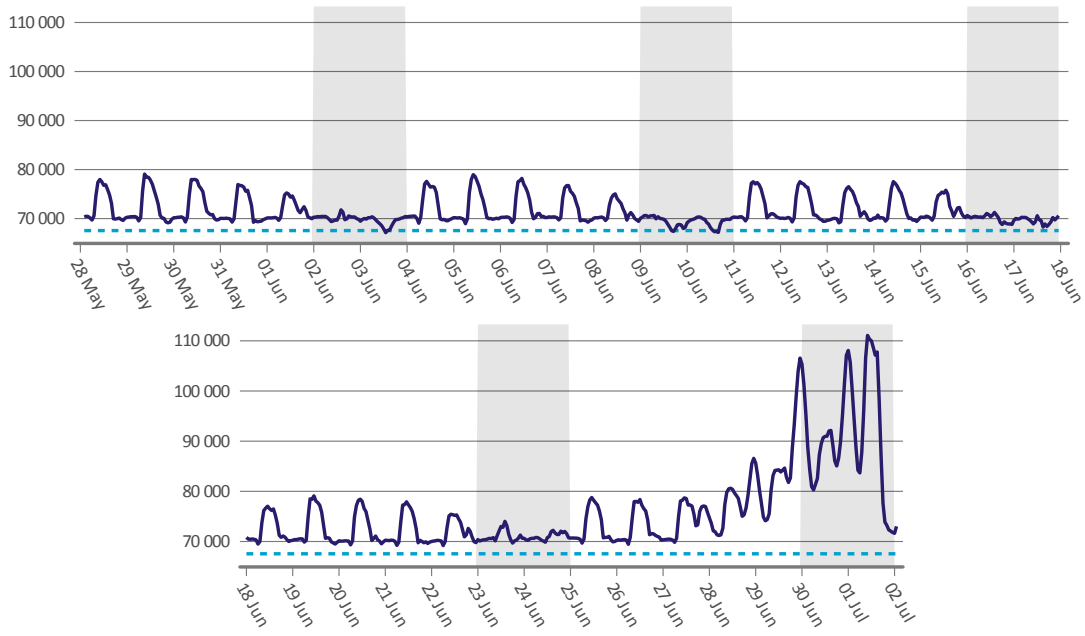
This section shows a flow cube produced during the pilot research with an observation period of five weeks. The cube was generated using the network prior at the level of municipalities, for the period of 28 May up to and including 1 July 2018. The cube will be evaluated for plausibility by checking 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, but these results show some of the quality issues that may occur. A major hindrance for a more detailed investigation of errors is the lack of availability of benchmark data against which population flow estimates can be compared. Visualisations were chosen over large data tables as the tool to evaluate the plausibility of the massive amounts of information in flow cubes.

Given a fixed municipality x, plotting the inflow $\text{Flow}^{\text{pop}}_{\text{in}}(x, h)$ for varying h results in the graph of the *population present in* x. This graph shows how the number of residents of the Netherlands present in x varies over time, in contrast to the residential population of x which can be approximated to be constant over the short observation periods under consideration.

June traditionally is a month in which many summer festivals take place throughout the Netherlands. In the hopes of capturing these well the flow cube for the June period was produced with the network prior.

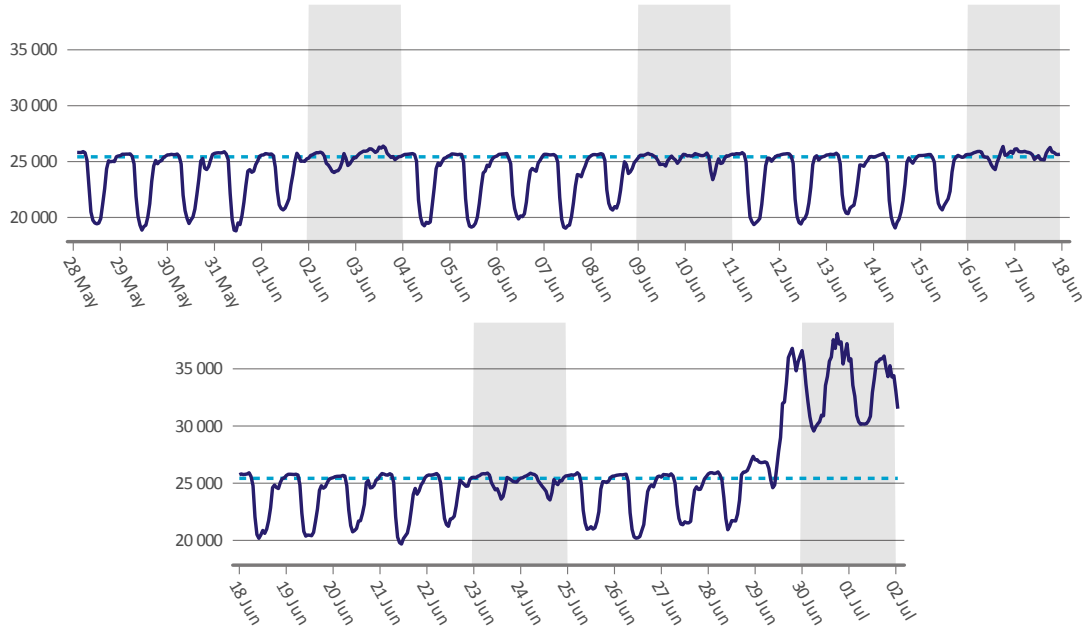
The eight round of the 2018 Grand Prix motorcycle racing season took place on Sunday 1 July 2018 at the TT Circuit in the municipality of Assen. Surrounding events such as qualifying races started on Thursday 27 June, and the preceding TT Festival lasted from the evening of 27 June until the night of Saturday 30 June. These can be clearly read off from the population presence graph of the municipality of Assen shown in Figure 2.5. 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 with the current techniques. Improving the levels of population presence graphs is the subject of ongoing research.

Figure 2.5: A population presence 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 population presence graph of the municipality of Beuningen in Figure 2.6. The words of caution we gave about the events in Assen apply here as well.

Figure 2.6: A population presence graph of the municipality of Beuningen



A dashboard has been published by SN at Statistics Netherlands (2019) in which the flow cube for the June period at the municipal level can be explored interactively. It features population presence 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.

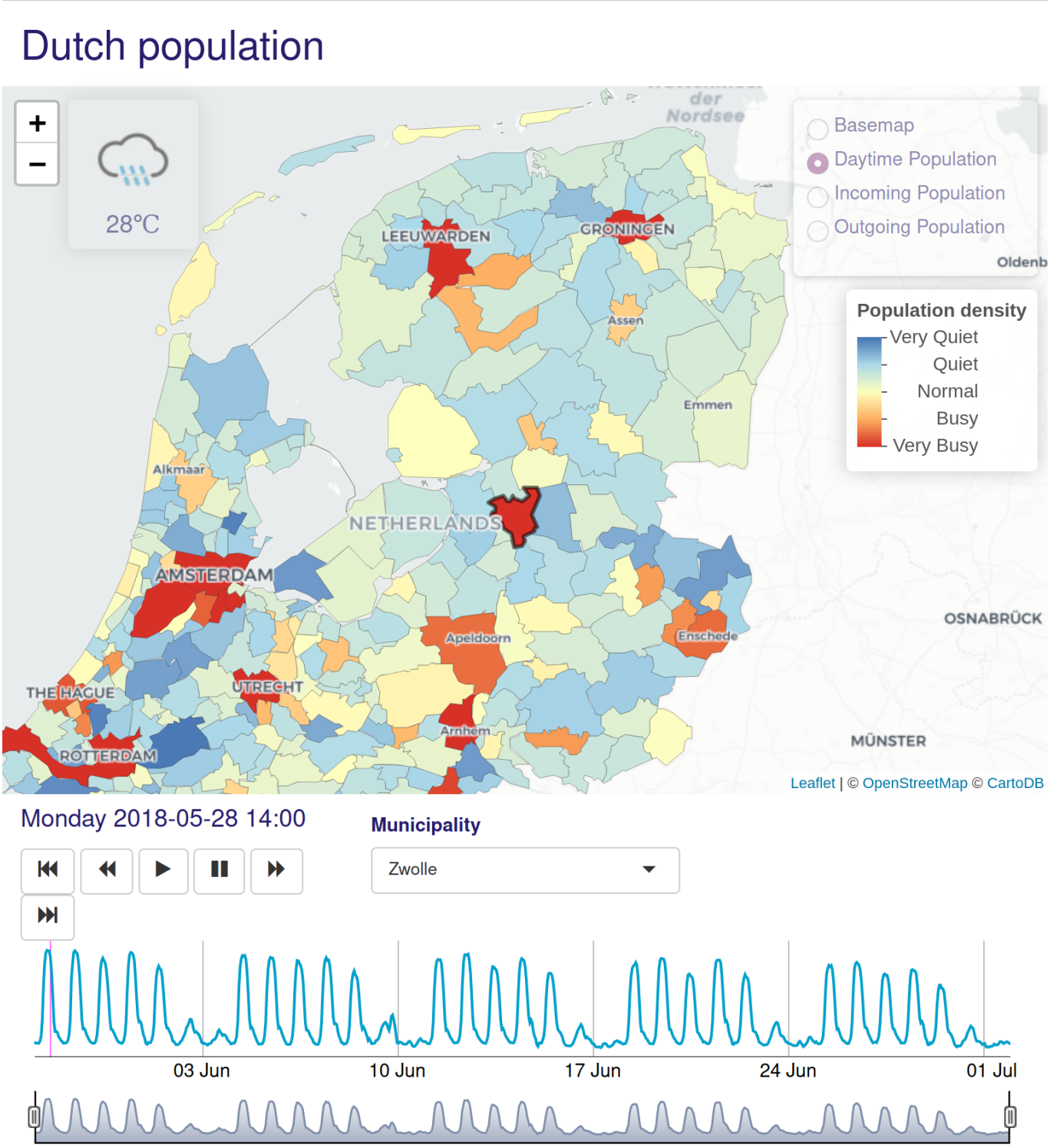
The screenshot in Figure 2.7 shows the daytime population on Monday 28th of May 2018, 14:00. The color indicate the relative population density. In red municipalities, which include almost all Dutch cities, it is more crowded than usual, probably because of work. Zwolle, a small city of about 120,000 inhabitants, is selected in the map. A time series for Zwolle is shown below the map, which clearly indicates the weekday patterns of five peaks from Monday to Friday.

Furthermore, the incoming and outgoing flows can be shown from the selected municipality to the other Dutch municipalities. The screenshot in Figure 2.8 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 is quite small if one expects that the visitors originate from many different municipalities. This is a consequence of the threshold of 15 devices that is enforced on the flow cube of devices before it is exported to SN.

2.7. Concluding remarks

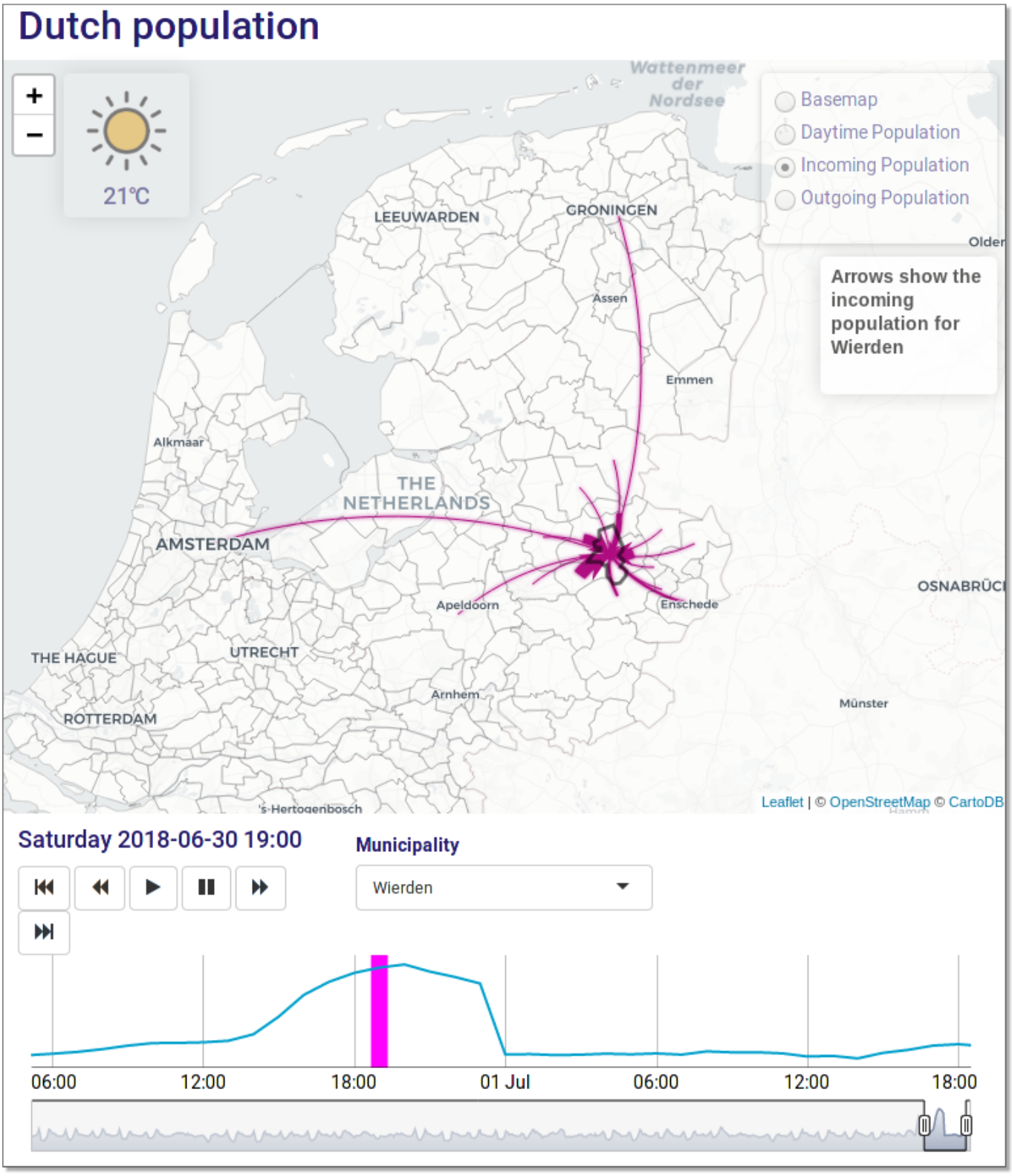
We have presented a methodology to estimate hourly population flows between Dutch municipalities from aggregated anonymous mobile network operator data. These estimates can be used by policy makers from governmental institutes to analyze mobility among the Dutch population. The results from the pilot study from 2018 have shown that population dynamics can be detected at a high temporal detail: workweek-weekend patterns can be seen, as well as major events.

Figure 2.7: A screenshot of the population flow dashboard showing daytime population



For the present situation regarding the COVID-19 pandemic, these population flows are useful in two specific applications. First, they can be used to measure the effect of national social restrictions on population flows. For instance, the number of people travelling to other municipalities is expected to be significantly lower in a (semi) lockdown than without social restrictions. These population flow estimates can be used to verify that. The second application is that these population flows can be used to trace back possible infections

Figure 2.8: A screenshot of the population flow dashboard showing the municipality of Wierden



after hot spots have been found. If, for instance, a lot of Corona infections suddenly have been found in municipality A, we can estimate how many people from other municipalities have visited municipality A, and therefore have a higher risk to be infected. The policies regarding testing can hence be adjusted if necessary.

As mentioned before, used methodology in many ways simpler than the methodology presented in this work package WPI of the ESSNet Big Data II project (deliverable I3). However, we have successfully illustrated how real MNO data can be used for an end-to-end statistical production process. The aim in future projects with MNO data is to apply these new methods as much as possible.

One of the main issues we found was that the estimates of our present population at night time are not always equal to the registered residential population numbers. Additional auxiliary data could partly solve this issue. 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 Tennekes et al. (2020) a *land use prior* is suggested which makes use of administrative data sources on land use to improve the estimated municipality of presence. Another aspect that may help to solve this issue is assigning not one but multiple home cells to a device. A device might switch between different nearby cells even when it remains stationary inside the owner's home.

The applied methodology does not take into account Dutch people who are abroad. Auxiliary data, such as the *Continuous Holiday Survey* Statistics Netherlands (2017a), might be used for calibration of the population flow estimates.

The used methodology assumes that every Dutch residents owns and uses exactly one mobile phone. In reality, there are many people who do not own a mobile phone, in particular young children and elderly people. Furthermore, there are many people who own multiple phones, e.g. one for private and one for business use.

A part of the methodology that is still underdeveloped is the calibration of the estimated numbers of devices to the estimated number of people. This is not straightforward, especially when data from one MNO is used that serve a selective part of the population. Also, people often have more than one device, or no device at all. Calibration methods have been developed in this work package (deliverable I3). It is worthwhile to study how this method can be used within our production system.

Finally, it is assumed that every Dutch resident actually lives at address of residence as registered in the PRD. For instance, students in higher education sometimes live in or near the municipality where they study, but are registered at their parents' address. Auxiliary data is needed to compensate for this issue.

3

Use of mobile network data for official statistics in Germany

3.1. Introduction and Partnerships

As part of the generally increasing use of digital technology, official statistics agencies have the opportunity to explore and employ new data sources and therefore have to organise their processes and procedures accordingly. For that reason, the Federal Statisitical Office (Destatis) is carrying out various feasibility studies to determine the usefulness of new digital data, such as mobile network data, for official statistics. The use of such data is seen as potential for a quicker and more precise production of official statistics.

In Germany, there are three mobile network operators (MNOs), Deutsche Telekom AG (hereafter called Telekom), Vodafone GmbH, and Telefónica Deutschland (hereafter called Telefónica), with a respective market share of one-third each (see Federal Network Agency[1]). As there is no legal access to privately held data in Germany, Destatis has to rely on cooperation agreements with the MNOs in order to examine the value of MNO data and the specific purposes for official statistics. Of the three MNOs in Germany, data from two MNOs in Germany (Telekom and Telefónica) were available for feasibility studies, but only with monetary compensation. Vodafone is not willing to cooperate with official statistics or any other partner in Germany as far as we know.

In order to research the use of mobile network data for official statistics, Destatis entered into cooperation with T-Systems International GmbH and Motionlogic GmbH (both wholly-owned subsidiaries of Deutsche Telekom AG) in September 2017. Destatis has purchased several data sets for different feasibility studies from Motionlogic GmbH. Since May 2020, the data of Telekom is not available any longer, as the responsible subsidiary Motionlogic was closed. Currently, T-Systems is rebuilding the business segment. Since October 2019, Destatis is in cooperation with the company Teralytics AG that processes, analyses and distributes data of the network of Telefónica Deutschland.

3.2. Mobile network data

Due to data protection rules, Destatis only receives anonymised and aggregated mobile network data from each MNO, which correspond to signalling data. The data records currently available to Destatis contain mobile activities in Germany of Telekom and Telefónica customers. This includes contract and prepaid customers. A mobile activity is defined as an event in the mobile network caused by a length of stay at a specific location or in a specific geographical entity without movement (also known as dwell time). Furthermore, signalling data are produced automatically and regularly. Only the location of the cell tower to which a mobile device is connected at a specific time is registered. In addition, mobile network data contain information on socio-demographic characteristics of mobile device users, such as age group, sex and nationality of the SIM card owner. However, the characteristics are only available for contract customers.

[1] https://www.bundesnetzagentur.de/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Marktbeobachtung/Deutschland/Mobilfunkteilnehmer/Mobilfunkteilnehmer_node.html.

As indicated above, Destatis has no permanent and legal access to mobile network data. This means that Destatis must always define and commission new project-specific data for all planned analyses and use cases. Furthermore, the mobile activities are subject to some assumptions or prerequisites in order to obtain the data record. One of the greatest challenges is therefore to adapt these assumptions and requirements in order to prepare the mobile network data as best as possible for the planned analysis.

First of all, a distinction must be made between static and dynamic mobile network data. Static mobile network data is understood as frequency analysis. In this case, only the counted mobile activities per time period and study area are evaluated. Dynamic mobile network data, on the other hand, reflect the movements of mobile activities between and in different study areas, so-called origin-destination matrices. They reflect the counted activities based on the starting point and the destination. Only the counted mobile activities from the mobile network are available and only these can be influenced for static mobile activities as well for dynamic ones.

Furthermore, the mobile activities can be prepared for all possible geographical study areas. Depending on the research question, administrative units such as the federal state level, district or municipal level can be used. However, the advantage of mobile network data is the small-scale availability of the data source. This means that the data can be used to consider smaller study areas than the municipal level, such as INSPIRE-compliant grid cells. Thus, the data can be georeferenced.

Since the number of mobile activities depends on the dwell time of mobile devices (Hadam, 2018), for each research question it has to be considered how to set it. If, for example, the resident population needs to be derived from the mobile network data, long activities of mobile devices corresponding to the length of the dwell time are counted and included in the data record, while short mobile activities are not taken into account. The dwell time of the mobile devices in the corresponding study area is then, for example, two hours in order to filter out short activities of mobile devices that result, for example, from fast movements between grid cells.

However, since no individual data and only aggregated values with a minimum number of 5 mobile activities are transmitted to Destatis, the challenge in data specification is to obtain a fine temporal and spatial resolution without falling below the minimum number of mobile activities. This is particularly problematic in rural areas and with finer temporal resolution. The minimum number of mobile activities may vary between the German MNOs, as this is not mandated by any regulation and depends on the respective anonymization and aggregation process and the MNOs bilateral agreement with the Federal Commissioner for Data Protection and Freedom of Information (BfDI).

The anonymization of the mobile network data takes place via unique pseudo-IDs for each mobile activity and for Customer-Relationship-Management (CRM) data, which are individually matched in a protected area and then delivered as aggregates. CRM data is personal data of the mobile network customers, more precisely contract customers, which is extracted from the billing system of the respective MNO. Due to the German Telecommunications Act, the pseudo-IDs are changed every 24 hours, so that no conclusions can be drawn about the course of individual persons.

3.3. Use cases

Due to the fact that Destatis only receives aggregated and anonymous mobile network data, Destatis mainly investigates possible use cases of these data in official statistics. Since only signals from active SIM cards of the respective MNO are counted, information about the location, time and movements is not available for mobile phones that are turned off or in flight mode. The combination of this information and the socio-demographic characteristics of the contract customers can be used to obtain valuable information for a variety of statistics.

There are basically three possible fields of application with the available aggregated mobile network data. Based on the information on time, place, movement and other socio-demographic characteristics of the mobile activities, primarily use cases can be found in

- Population statistics.
- Commuter statistics.
- Tourism statistics.

The objective at Destatis is to use mobile network data to provide a valid picture of the daytime and resident population and of the mobility of the population in the whole of Germany. The population figures of the 2011 census as well as the 'Commuter Atlas' published by the State Statistical Office in North Rhine-Westphalia (IT.NRW) are used as benchmarks to check the representativeness of the mobile network data. In the following, all feasibility studies and use cases carried out by Destatis are presented. The conceptual designs of the conducted feasibility studies were coordinated with the BfDI.

The use cases described below are divided into already completed and still ongoing projects. An obvious subdivision into static and dynamic mobile network data and corresponding use cases becomes visible. At the beginning of the research work Destatis only had access to static mobile network data. Studies based on this kind of data are described in section 3.3.1. In the course of the research work, more extensive and dynamic data were acquired and were extensively investigated, especially in the context of the Covid-19 pandemic. Therefore, section 3.3.2 mainly focuses on mobility analyses and the use of dynamic data in official statistics.

3.3.1. Completed Projects based on mobile network data

3.3.1.1. Correlation of the mobile network data of two MNOs with census data

In order to make statements on the representativeness and structure of the mobile network data, Destatis used data from the Telefónica network in addition to the data of the Telekom network and concentrates on the federal state North Rhine-Westphalia (NRW). Since there are three MNOs in Germany, each with a market share of about one third on the German mobile communications market, the data records of these two providers can theoretically map about 66% of mobile phone users in NRW. The primary goal that Destatis is pursuing in using both data sets is to increase the representativeness of the mobile network data and thus to carry out a structural comparison of both mobile network data sets.

The data records that were transmitted to Destatis contain mobile activities of Telekom and Telefónica customers in NRW[2] for a statistical week, which contain selected days and months from 2018/2019 without school or public holidays in a 22-hour period. The mobile network activities comprise the average activities on the weekdays selected and are not extrapolated to the whole target population or extrapolated based on the regional market share in contrast to previous or following data sets[3]. The weekdays are categorised according to five types of day, with the days from Tuesday to Thursday being grouped together. In addition, mobile network data contain information, for example, on socio-demographic characteristics of mobile phone users, such as age group and sex. In compliance with data protection rules, the mobile activities were anonymised and aggregated. Only values based on a minimum of 5 activities per grid cell were transmitted to Destatis. The grid cells comply with the INSPIRE directive and correspond to the census grid cells of the 2011 Census Atlas[4]. To verify and increase representativeness, the two non-extrapolated mobile network data sets with mobile activities from the networks of Telekom and Telefónica are merged. For this purpose,

[2]The project implementation at national level and first feasibility studies were carried out by Destatis in cooperation with IT.NRW and was therefore limited to the federal state of North Rhine-Westphalia.

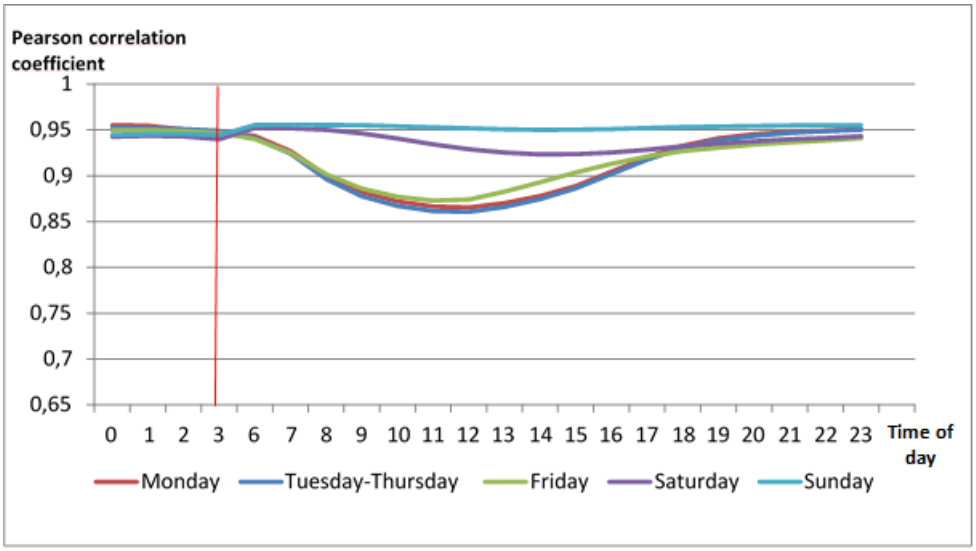
[3]See (Statistisches Bundesamt, 2019) for the first feasibility study. The aim was to analyse whether mobile network data are sufficient to provide a valid picture of the current resident, daytime and working population.

[4]For more information on the Census Atlas please refer to: `https://atlas.zensus2011.de/`.

the mobile activities are filtered according to weekdays and hourly values and linked to each other using the underlying grid cell.

In order to get a first impression of whether and to what extent a combination of both mobile network data sets will lead to an increase in representativeness, the following section examines the relationship between the combined mobile activities (Telefónica & Telekom) in 2018/2019 and the population figures of the 2011 census. The population figures of the 2011 census will be used as a benchmark to check the representativeness of the combined mobile network data. The resulting Pearson correlation coefficient determines the linear relationship or the strength of the relationship between mobile activities and the number of population per hour and study area, as shown in figure 3.1.

Figure 3.1: Correlation between Telefónica & Telekom mobile network activities from 2018/2019 and population figures of the 2011 census in Germany.

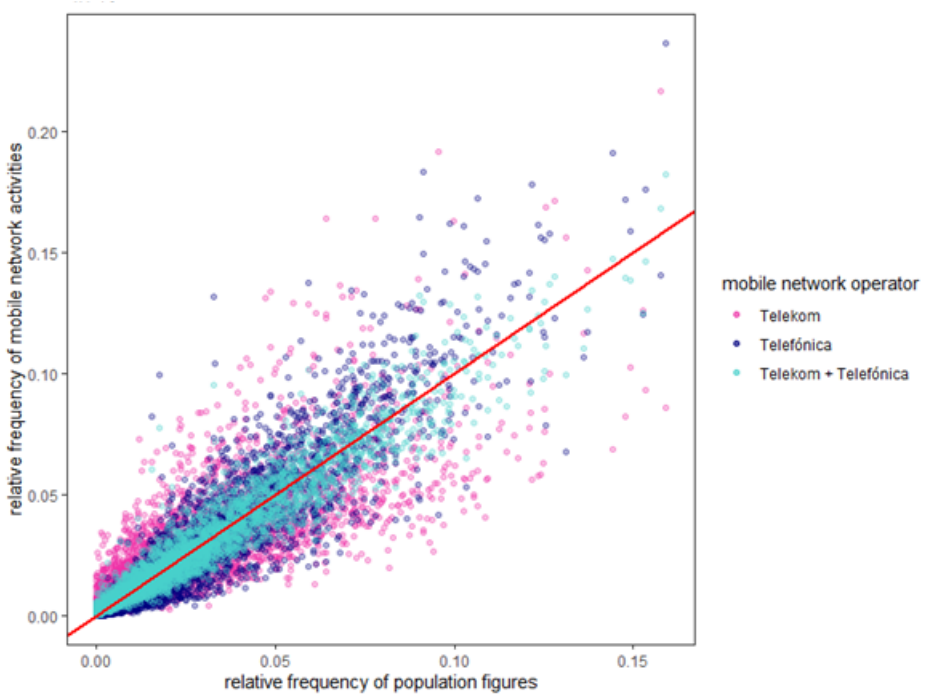


Source: 2011 Census, T-Systems/Motionlogic, Telefónica/Teralytics

The values of the correlation coefficients in figure 3.1 show an overall very high positive correlation of up to 0.95 between the combined mobile activities and population figures in the evening hours, as well as throughout the day on Saturday and Sunday. In comparison to correlation analysis with data from only one MNO (Statistisches Bundesamt, 2019; Hadam, 2018), a combination of data on mobile activities from different or separate mobile network providers in Germany leads to a significantly improved approximation to the distribution of population figures. Furthermore, especially on Sunday evening, this approach leads to an almost perfect linear correlation between mobile activities and the population figures of the 2011 census, as further shown in figure 3.2. Due to the high correlation with the population figures of the 2011 census, mobile activities on a Sunday evening are suitable for deriving the resident population on the basis of mobile network data.

Figure 3.2 shows the distribution of the relative frequencies of mobile activities on a Sunday evening in relation to the relative frequency of the population from the 2011 census. A perfect match of both distributions is indicated by the red straight line. The relative frequencies of mobile activity are shown both separately by MNO and combined (Telekom + Telefónica). Starting with the relative frequency of mobile activities from Telekom's network (magenta-colored dots), it is obvious when comparing the relative frequency of population figures that the dots are strongly scattered around the red straight line. At the beginning of the distribution of the relative frequency of mobile activity, the potential resident population is overestimated on the basis of mobile network data from the Telekom network in areas with a tendency towards low population numbers. Consequently, mobile activities underestimate the potential resident

Figure 3.2: Distribution of the relative frequencies of mobile network activities 2018/2019 to the relative frequency of the population from the 2011 census.



Source: 2011 Census, T-Systems/Motionlogic, Telefónica/Teralytics

population in areas with a tendency to high population numbers. Furthermore, if one considers the relative frequency of mobile activities from the network of Telefónica (dark blue dots), the potential resident population with a tendency to low population numbers is underestimated and in areas with a tendency to higher population numbers is overestimated[5]. If both distributions are now combined - by merging the mobile activities of both MNOs - the relative frequencies (turquoise dots) are visibly approximated to the red straight line and thus to the relative frequencies of the population from the 2011 census. The merged distributions scatter less around the red straight line at the beginning as well as at the end of their distribution and thus largely cancel out the distortions in the individual distributions of both providers. Like the correlation analysis in figure 3.1, the comparison of the relative frequencies of mobile activities and population figures from the 2011 census in figure 3.2 shows that combining mobile activities of different MNOs leads to a significant increase in representativeness using the example of the representation of the resident population based on mobile network data.

3.3.1.2. ESSnet 'City data from LFS and Big Data'

Reliable knowledge on labour force indicators of a country's population is essential for sound evidence-based policymaking. For instance, the geographic distribution of the unemployment rate is used to make decisions regarding the allocation of resources. The Labour Force Survey (LFS) is generally designed to provide reliable estimates for larger domains such as the national or regional level. However, to be able to make policy proposals in urban areas official statistics have to provide information for these areas. The ESSnet project 'City data from LFS and big data'[6] examined whether and to what extent indicators of the LFS can be estimated on spatially disaggregated levels, like Functional Urban Areas (FUA), by using mobile network data as auxiliary information. The production of precise small area estimates relies on the availabil-

[5] Among other things, it must be taken into account that the network quality most likely varies regionally. The population structure also differs between urban and rural areas. This is also an important point, as different population groups prefer different network operators.

[6] See for more details EC (2019).

ity of predictive auxiliary variables. Therefore, in addition to the usage of LFS information, an alternative source of passively collected mobile network data of the Telekom is used to estimate unemployment rates for FUAs.

The motivation for using mobile network data was that they are collected without interruptions and include valuable information on timing, location and intensities of aggregated mobile activities. The major advantages of these data are their finer spatial and temporal resolution. Since Destatis used aggregated mobile network data for the whole of Germany, it was possible to predict the unemployment rate for all FUAs in Germany.

The mobile activities of the Telekom network in this data record refers to the municipality level of NRW and thus can be aggregated up to the FUA sublevel since municipalities are nested within these subareas. The data contain extrapolated mobile activities to the whole target population for a statistical Sunday evening. Due to the high correlation between the population figures of the 2011 census and mobile network activities especially on Sunday evening (see figure 3.1 or Statistisches Bundesamt (2019)), average mobile activities from eight selected Sundays in the months of April, June and July 2018 in a time period from 8 p.m. to 11 p.m. are used. To avoid distortions in the representation of the resident population only Sundays without school or public holidays were selected. This time period is also used as the LFS surveys questioned the resident population.

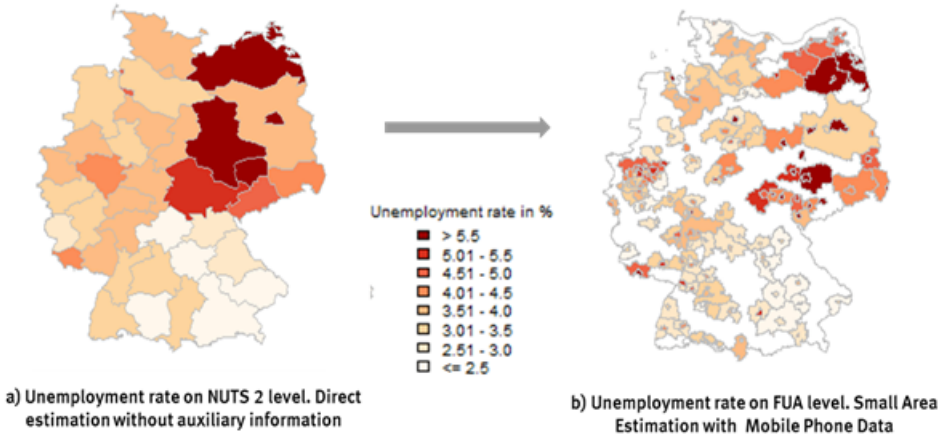
Since the LFS is not designed to produce indicators on smaller areas than NUTS 2-level (government districts), a small area estimation method is used. According to the model described by Schmid et al. (2017) and as part of a collaboration between Destatis and Freie Universität Berlin, mobile network data were geographically associated to the unemployment rate and were estimated for smaller areas by using a Fay-Herriot (FH) approach - an area-level model - with an integrated transformation of the dependent variable.

As can be seen in figure 3.3, unemployment rates of the LFS can be estimated from the NUTS 2-level (government districts) to the smaller FUA level if additional auxiliary information such as mobile network data is used. Since the unemployment rate can only be estimated validly for the NUTS 2-level (without auxiliary information) (see figure 3.3a), a direct estimation on FUA level is no longer possible due to an insufficient sample. The small area method used enables to estimate unemployment rate by using aggregated and anonymised mobile network data at spatially disaggregated level and obtain reliable results for FUAs (see figure 3.3b). Furthermore, reliable estimates for areas without observations can be estimated by this method.

This method revealed also a gain in accuracy compared to the direct estimators, since the uncertainty in the estimates will be reduced due to smaller coefficient of variation. The coefficient of variation is a measure of relative variability and a useful statistic for comparing the degree of variation from one data series to another. Figure 3.4 shows that by using the example of the unemployment rate of women, a reduction of uncertainty by using a FH model instead of direct estimation can be obtained. Thus, the FH estimator is more accurate than the direct estimator for the mean of unemployment rates of women for FUAs, where the red line represents the acceptable coefficient of variation of 20%.

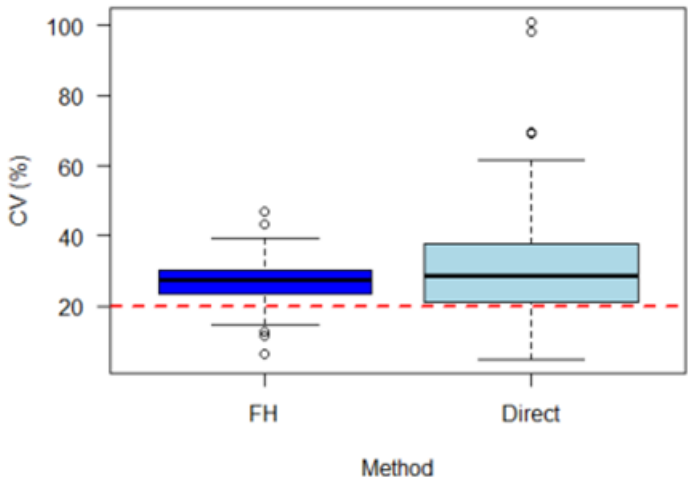
Even if the method and the auxiliary information used here leads to improved estimators, it should be pointed out that there may be better auxiliary information for estimating the unemployment rate as long as it is available at least at the municipal level, such as the data of the Federal Employment Agency of Germany. It can be assumed that these data will lead to more valid estimates. Nevertheless, the focus in this project was to find out to what extent mobile network data can be used for other purposes, such as to provide background auxiliary information.

Figure 3.3: Link between mobile network data and LFS indicators. Unemployment rate of women on NUTS 2 and FUA level.



Sources: Microcensus 2016, T-Systems/Motionlogic

Figure 3.4: Reduction of uncertainty by using a FH model (FH) instead of direct estimation (Direct). Coefficient of variation (CV) for estimating unemployment rates of women on FUA level in Germany.



Sources: Microcensus 2016, T-Systems/Motionlogic

3.3.2. Ongoing Projects based on mobile network data

3.3.2.1. Georeferenced intercensal population update

The previous analyses of population representation with mobile network data show that the distribution of the population can be represented well and in a timely manner. Based on this, Destatis wants to clarify the research question of whether (static) mobile network data are suitable for mapping the intercensal population update of the year 2019 on a small-scale level within the framework of a weighting procedure. For this purpose, suitable data sets will be acquired from the data service provider Teralytics. The data will be processed in such a way that they represent the resident population as accurately as possible at the level of INSPIRE-compliant 1×1 km grid cells. Since Destatis want to represent the population update in a georeferenced form, it is essential that only the resident population is available in the mobile network data. For this purpose, two approaches of data specification are pursued and will be examined:

a) **The first and the last signal of the day**, which captures the location of the first and the last signal of

each SIM card, so that with identical places one can assume the place of residence;

b) **The average activity on a Sunday evening**, whereby the location of an active mobile device is determined on Sunday evening from 8 p.m. to 11 p.m. Feasibility studies for the study area NRW have shown, in comparison to other weekdays and periods that mobile activities on a Sunday evening show a particularly high correlation with the population figures of the 2011 census, so that it can be assumed that the aggregated signals on Sunday evening are a good indicator for the resident population.

The distribution and the relative values of the determined "resident population" from the mobile network data will be used further as weights for the smaller 1×1 km grid cells, in order to calculate the population update of the year 2019 from the municipal level nationwide to 1×1 km grid cells in a weighting procedure. Furthermore, it will be investigated to what extent the described weighting procedure can be applied to socio-demographic data, such as age and gender.

3.3.2.2. Commuter analysis

The project 'Pendler Mobil', which is carried out in cooperation of Destatis and the State Statistical Office IT.NRW, has the objective of identifying domains where dynamic mobile network data may contribute to complementing the present commuter statistics. By the use of origin-destination matrices it shall be investigated, whether data from the mobile network can be used to map commuter flows during the day. In addition, it will be examined whether and to what extent the current commuter statistics for NRW can be supplemented with this data. For example, it will be examined whether mobile network data can contribute to a smaller-scale and temporally faster mapping of commuter flows. The commuter statistics can only be provided at the municipality level and can only be updated annually due to the primary surveys required for this purpose. The Commuter Atlas published by IT.NRW is used as a benchmark to check the representativeness of the data.

Therefore, Destatis has received origin-destination matrices of the Telekom network for the year 2019 in order to carry out mobility analyses of the population with a special focus on commuter flows in NRW. The data contain daily and extrapolated origin-destination matrices for the months August, September and October 2019, with a dwell time of 2 hours. The movement of a signal is only counted if the mobile device was active and has left the place of residence before 9 a.m.[7]

Since one of the research objectives is to determine the working population and commuters in the data, it is important to prepare the data in such a way that these target groups are obtained. The determination of a suitable length of stay is essential. Hence, the data record therefore contains dwell times in a half-hour cycle (half an hour, 1 hour, 1.5 hours, etc.) for each destination. This is intended to further investigate whether it reveals a connection between the length of stay at a destination and the volume of employment (i.e. full-time or part-time employment) or when differentiating between economic sectors.

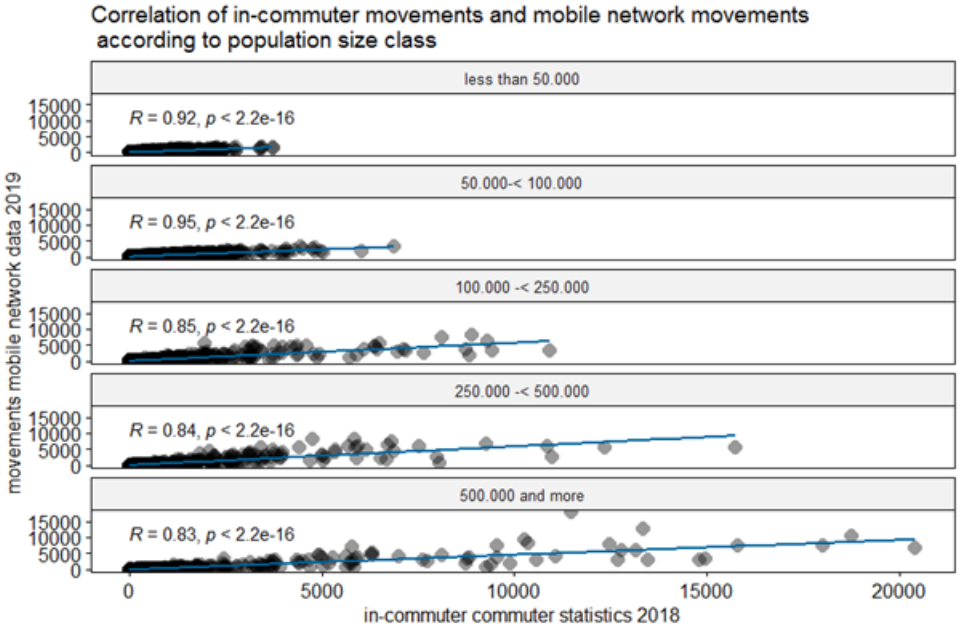
First preliminary results so far show that the dwell time of mobile activities in the study area plays a decisive role in the representation of commuters. The longer a mobile device remains in the study area, the higher the probability that mainly full-time employees are covered. In this case, however, mobile workers as well as part-time workers are compulsorily lost. If the dwell time is too short, the less accurate the filter can be and too many uninteresting non-commuter movements are included in the representation.

In particular, it is already examined whether and to what extent the data from the commuter statistics in NRW correlate with the specified mobile network and whether it correlates more strongly in certain geographical and data specific areas than in others. Therefore, the Pearson correlation coefficient was

[7] This corresponds to a definition of the data provider Motionlogic GmbH.

determined between mobile network data and commuter statistics. The overall correlation coefficient for the in- and out commuter relations contained in the commuter statistics and mobile network data is around 0.85. This indicates a strong linear correlation of both data sources, as shown, in figure 3.5. The figure shows an example of the correlation between in-commuter movements from the 2019 mobile network data and the 2018 commuter statistics for the five common population size classes in Germany. It can be seen that in principle a high correlation of at least 0.83 prevails in all population size classes, but this correlation decreases the larger the population class per study area becomes. This means that the movements in the mobile network data become less precise and less consistent with the commuter statistics, the more inhabitants live in the considered study areas.

Figure 3.5: Correlation Coefficients of in-commuter movements of the commuter statistic and of mobile network data according the five population sizes classes and in terms of the respective absolute values.



Sources: Commuter statistic IT.NRW 2018, T-Systems/Motionlogic

On the one hand, figure 3.5 shows that the mobile network data defined here are strongly related to commuter movements of the commuter statistics. On the other hand, it is already visible that the absolute values of the movements in the mobile network data (vertical axis in figure 3.5) clearly underestimate the in-commuter movements compared to the commuter statistics (horizontal axis).

In further steps, it is planned to compare the regional movements of both data sources and to carry out smaller-scale analyses in order to analyse inner-city movements for further research questions. Since the evaluation of the data from the commuter statistics is currently possible at the municipal level at most, it will be further investigated to what extent mobile network data could contribute to receiving information on a smaller scale. Therefore, a so-called mixed mapping is used for the present analysis, depending on the minimum number of activities in a grid cell (according to the BfDI it has to be at least 5). This means that the municipal level in Germany is used with the so-called official municipality key in rural and less urban areas and grid cells between 250×250 m and up to 4000×4000 m are used for cities with more than 100,000 inhabitants.

3.3.2.3. Transport and traffic analyses

The project 'VerBindungen' is a research co-operation project with the Federal Ministry of Transport and Digital Infrastructure (BMVI). For the first time, the VerBindungen project will combine transport and commuter relations according to departure and destination locations and their accessibility, as well as the resident population on the basis of official statistics and new digital data in such dimensions that this information can be determined on a small scale. The aim of the project is to create a representative determination of nationwide transport demand and accessibility indicators. Therefore, the research project will develop suitable methods to derive reliable statistics on origin-destination traffic and the residence of the population on a small scale. For this purpose, data from the Statistical Offices of the German states and the federal government, statistics of the Federal Employment Agency on the places of work and residence of employees subject to social insurance contributions, as well as mobile network data and floating car data (FCD) will be used. The project started in autumn 2020 with a project duration of three years.

3.3.2.4. Mobility indicators

The measures to contain the Covid-19 pandemic have brought public life largely to a standstill in most countries of the world. In Germany, as in many other countries, the temporary closure of stores, schools and universities, workplaces, cultural institutions, vacation facilities and much more was aimed at reducing social contacts and thus the mobility of the population. Based on mobile network data, Destatis compiles regional indicators that show the development of mobility at administrative district and municipality levels. The mobility indicators developed are used to map the changes in mobility behaviour as a result of the Covid-19 pandemic in Germany and to analyse the effects of restriction measures (Statistisches Bundesamt, 2020).

The aggregated and anonymised data of the Telefónica network, contain the number of movements within a certain period of time (day or month), which are identical with regard to the regions of origin and destination (district or municipality). A movement appears, when a mobile device switches from one radio cell into another. To detect movement into a target area, the mobile device has to remain in a cell for at least 30 minutes. Movements can therefore also be detected within a region, if the mobile device changes from one radio cell to another within this region.

In order to depict the change in general mobility behaviour, all entries to and within the administrative districts are being evaluated for each district. Following this, the daily mobility determined is compared with an average value for the corresponding weekday from the same month of the previous year. The cartographic representation in figure 3.6 shows the daily change in mobility for each district over the past 31 days to the current data status. A value of -20, for example, shows that mobility in 2020 was 20% lower than in the respective month of 2019. This can be used to highlight regional differences (see for more details Statistisches Bundesamt (2020))[8].

In addition to the cartographic representation, figure 3.7 compares the change in aggregated mobility for each federal state in Germany. In contrast to figure 3.6, the calculated results in figure 3.7 are dated back to January 2019. Thus, the changes in daily mobility from the beginning of the COVID-19 pandemic can be displayed. The results in figure 3.7 show a significant decrease in mobility as a result of the restrictions from calendar week 12. In the further course of March 2020, a decline in mobility of around 40% compared to the previous year is shown for Germany. On Sundays, the decline is even bigger. This can be interpreted as an indication that people reduced particularly the number of dispensable movements. This observed, significant deviations on public holidays are partly due to the preparation of the reference values and show not only the change compared to the previous year, but also the difference between public holidays and

[8]Note that the cartographic representation in Statistisches Bundesamt (2020) is available as a dynamic and interactive figure. For reasons of implementation, the map in figure 6 in this article is stored as a static figure, primarily to give the reader a first graphic impression of the topic.

working days (Statistisches Bundesamt, 2020).

Figure 3.6: The daily change in mobility at administrative district level compared to the previous year of the past 31 days.

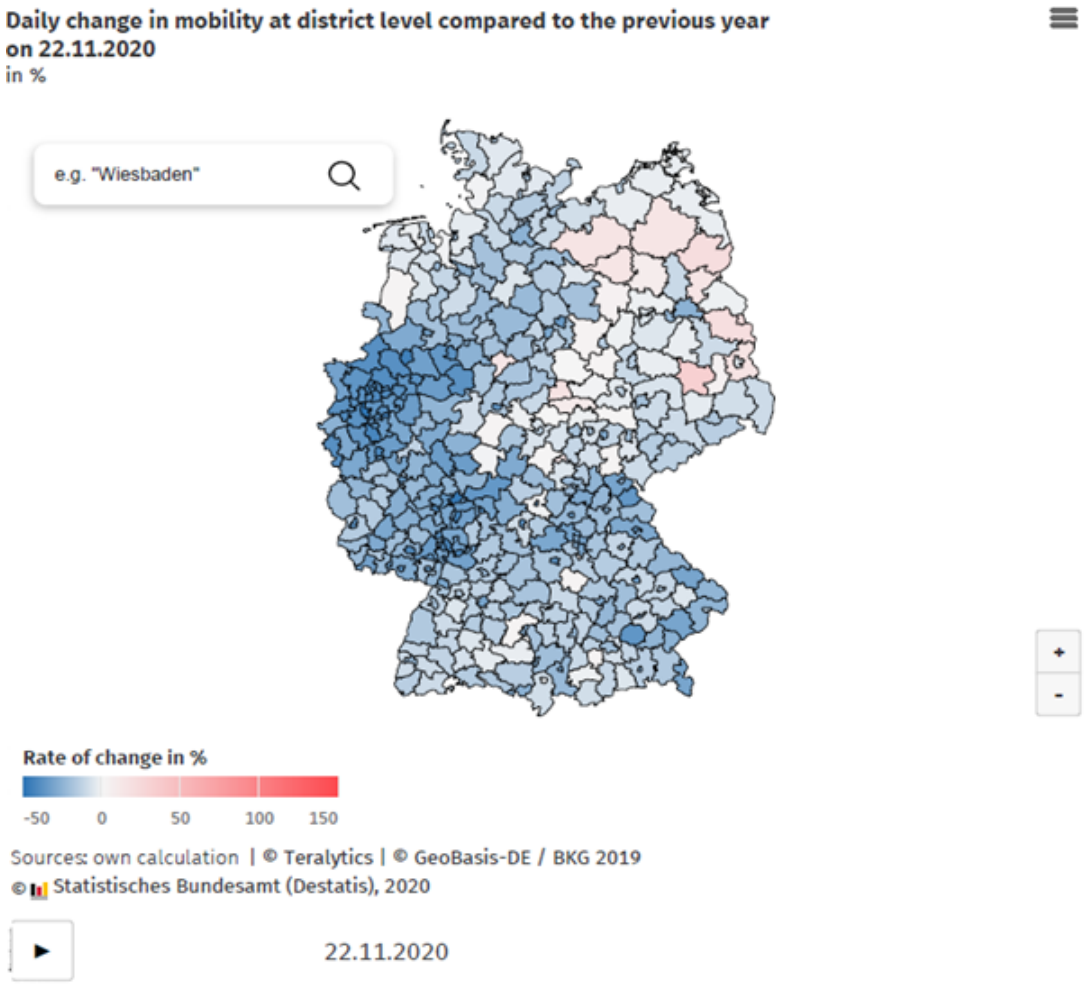


Figure 3.7: The daily change in mobility aggregated for the individual federal states in Germany since January 1, 2020.

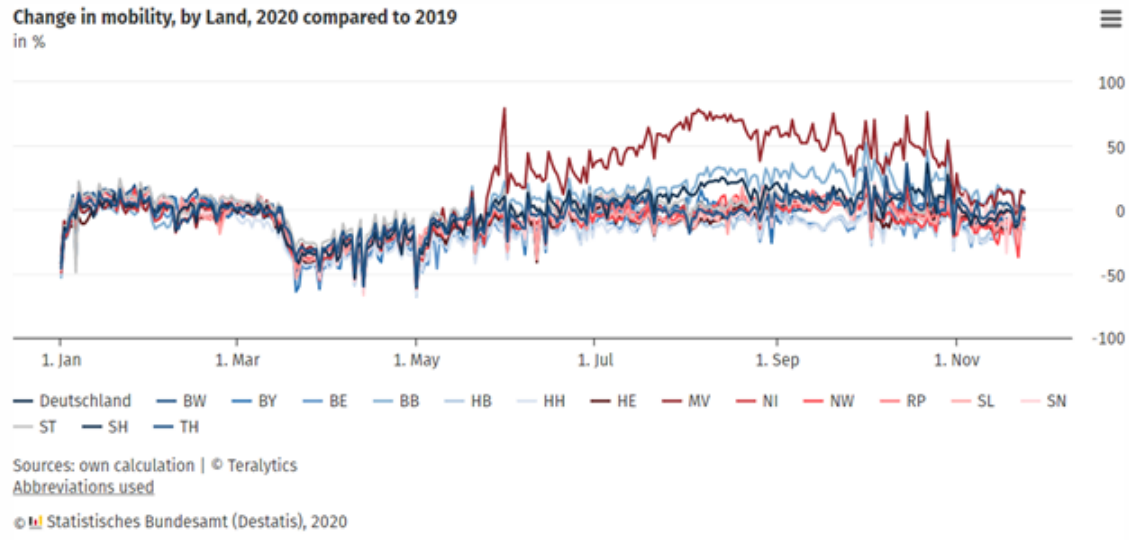


Figure 3.6 and figure 3.7 represent only two exemplary analyses of the mobility change of the population. Statistisches Bundesamt (2020) provides further evaluations, such as mobility changes at administrative district level in Covid-19-hotspots and non-hotspots. The Institute for Research and Development in Federal Statistics is constantly working on the further development of these indicators.

The above-mentioned results rely on experimental analyses. Calendar adjustment of the data is still pending, which makes comparability between two annual values difficult and can lead to biases in some cases. Currently, the comparative values for 2019 are summed up for each weekday over all days of the respective month and divided by the number of weekdays. Public holidays are not taken into account here. Furthermore, reasons for outliers, especially in results on the detailed regional level, need to be investigated.

Another methodological challenge is the granularity of regions. Thus, the regional classification of the underlying data sets provided by Teralytics is not homogeneous. A part of the mobile network data, especially for urban regions, is available at the municipality level or spatially lower, whereas for rural regions a part of the mobile network data is only available aggregated over several municipalities or districts. This can lead to distortions in the calculation of commuter mobility, for example. These and other challenges must be analysed and solved in the further course of the project.

3.4. Conclusion

The first results of the conducted feasibility studies on population statistics, mobility analyses or studies using small area estimation techniques show that mobile network data can – to a certain extent - provide a good picture of the population, of mobility behaviour and serve as promising supporting information for methodological procedures. Mobile network data stands as a potentially rich source of information for this kind of analyses, especially due to its finer temporal and spatial resolution. Each of the individual use cases makes it clear that mobile network data is a valuable data source for future statistics production.

The restriction to have only aggregated mobile network data for the research available, limits the number and the possibilities of the feasibility studies noticeably. Hence, one of the biggest challenges in all the feasibility studies is and will remain the definition of a suitable mobile network data set for each use case, as Destatis has no unrestricted access to individual data. Especially in view of the limited adjustment possibilities and the almost non-existent influence on data processing, the determination of a suitable data set is extremely difficult. Without the urgently needed knowledge of the detailed methodological procedures at the MNOs, it is not possible to exploit the full potential of the data for each question. Restrictions in the use of mobile network data for official statistics in Germany can be divided into 1. legal/strategic and 2. data-specific aspects.

Firstly, a permanent legal access and a secure cooperation environment is needed to ensure that the data can be used with rigour and quality and that there are no sudden data losses due to business dissolution. A solution for sustainable access has not yet been implemented at Destatis. In addition, Destatis currently has a cooperation with only one MNO. Should this operator leave the market, a data gap will arise that cannot be closed easily.

Secondly, despite strong efforts to get insight in the process of data generation (including aggregation and its detailed metadata), it is still non-transparent at least regarding methodology and algorithms, which is mainly due to business secrets. This fact negatively impinges on the value of this source. To be able to create official statistics with this data source in the future, more transparency and insight is required. Hence, a permanent project at Destatis concerns the investigation of the methodological challenges of mobile network data with regard to the aspects mentioned above. A loss of information must also be expected, e.g. through a minimum number of counts per study area or 24 hours evaluation period, especially if a temporally and spatially finer resolution is required. In addition, the different minimum number of counts at the MNOs makes it difficult to compare the data of the individual MNOs. Distortions/uncertainties due

to unclear data influences (e.g. socio-demographic characteristics of mobile phone customers) cannot be corrected or quantified due to the non-transparent data production process. Taking the socio-demographic characteristics as an example, some characteristic attributes are captured more frequently than others in the mobile network data due to the different customer structures of the individual MNOs. This inevitably leads to distortions in the characteristic values, which can only be compensated to a limited extent by the MNO. Reasons for the distortions may be different market shares in a study area of each operator, i.e. an over- or under-representation of a provider in the federal state and therefore different sized customer bases in the investigated area. Moreover, since the socio-demographic characteristics are only recorded by contract customers, it is still unclear how the characteristics are distributed from contract to prepaid customers. Family contracts, duplicate SIM cards and missing information on prepaid customers, among others, make it even more difficult to represent the socio-demographic characteristics on the basis of mobile network data in a representative and differentiated way.

Last but not least, the presented and the still ongoing feasibility study present promising results and highlight mobile network data as a potential data source. Nevertheless, all results must be treated with caution. Due to the aforementioned challenges regarding undisclosed methodology, the quality of the mobile network data and thus of the results cannot be sufficiently verified and quantified so far.

4

Italian experiences with CDRs

In 2015, the Italian National Statistical Institute (in short, Istat) launched a project in collaboration with the University of Pisa and WIND with the aim of studying the use of Call Details Records (CDRs) data to build an "Origin/Destination Matrix" to be benchmarked against the "commuting matrix" produced by Istat starting from the population census data.

In this first implementation of the project, the CDR data were provided by a specific Italian MNO, namely WIND, to the University of Pisa that processed it according to the conceptual and methodological framework provided by Istat. Furthermore, Istat was also involved in the results evaluation phase. This project, conducted for research purposes, was very successful, especially for the encouraging achieved results. Hence, it was lately resumed for possible follow-up activities. A short description of the project and its main results is provided in section 4.1.

In order to manage CDR data, the institute has entered into an agreement with the Provider free of charge and has defined a secure protocol for the acquisition of anonymised data from the provider for a province for 5 consecutive weeks. Istat also stated the aims of the research project in the National Statistical Plan (PSN) [1] that establishes the statistical surveys of public interest entrusted to the National Statistical System and the related information objectives. The projects included in the PSN must receive a pass from the Guarantor, in order to be authorized to enter an implementation phase.

The dialogue with the Guarantor in addition to technical clarifications required the drafting of the Data Protection Impact Assessment (DPIA) document. The assessment of the impact on the protection od the data and of the risk for privacy of citizens is reported in section 4.2

4.1. Mobile phone data for population estimates and for mobility and commuting pattern analyses

4.1.1. Objective and data description

Statistical outputs on the use of mobile phone data are particularly important in Official Statistics and strictly related to the census output. One of the aims of Istat is to produce high quality estimates on population density at a very small scale on the basis of population census data and administrative data.

Currently, Istat is implementing a census transformation program, as well as other National Statistical Institutes (NSIs) in developed countries. The new framework provides for leaving the traditional door-to-door decennial census in favor of the combined use of statistical registers based on administrative data

[1] The National Statistical Program (PSN) of Italy is the regulatory act which, based on art. 13 of Legislative Decree no. 322 of 1989 and subsequent additions, establishes the statistical surveys of public interest entrusted to the National Statistical System and the related information objectives. `https://www.sistan.it/`.

and social surveys. Specific aspects, like coverage of sub-population and other information that cannot be derived by administrative data and ongoing social surveys, will be investigated by yearly ad-hoc sample surveys, so to guarantee yearly high quality population estimates at small scale.

The usability and potentialities of Mobile Phone Data (MPD) are analysed with respect to the new census framework, underlining the steps in which MPD may increase the information already available via administrative data and social surveys.

To this aim, MPD can be used in different ways, both as complementary data source and primary data source, as well as to validate population estimates. In this report, the abovementioned aspects are investigated, even if, firstly, the reliability of MPD is assessed through the comparison with the official estimates.

To analyse and understand the spatio-temporal behaviour of people it is important to assess the localization of MPD, i.e. the information concerning the geographical referencing of mobile phones during their activity. In the case of CDR, we have a passive localisation that corresponds to the antenna/sector code to which the calling device and the call end antenna has been linked. The data of the CDR, in addition to those of localization, are made up of a code that identifies the device, also called SIM (Subscriber Identity Module), that makes the calls, other than the information related to the type and time of the event.

The cellular signal is picked up by an antenna and enters the network. The antennas are the mobile telephone systems that receive and re-transmit the signals of mobile phones which are distributed throughout the territory in a capillary manner, according to population density. Each antenna is designed to serve a portion of limited territory, called "cell".

The cells are divided into several sectors. Each sector is a service characterized by a technology, a direction and a coverage area of antenna. The coverage area is named Service Area, in figure 4.1 we can see an example of three service area identified by three different colors.

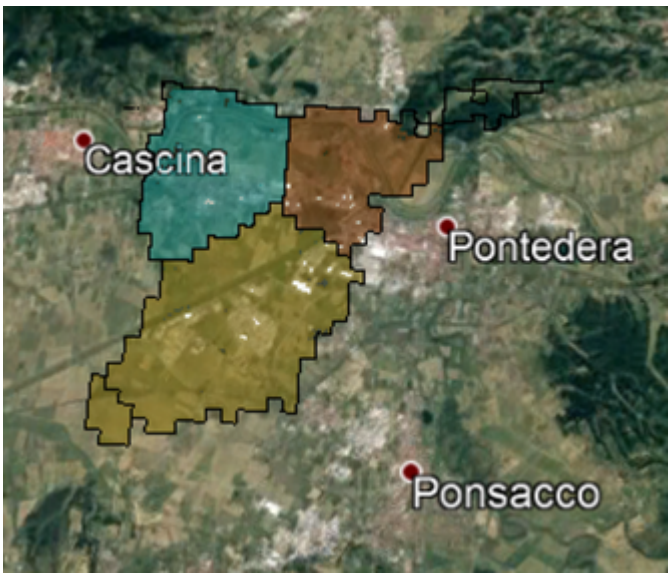


Figure 4.1: Best Service Area of an antenna. Example of three Best Service Areas in 3G technology. The antenna tower is located at the intersection of the 3 areas.

ISTAT received from WIND the data of the CDR of its subscribers concerning calls made in the province of Pisa from 1st January 2017 to 12th February 2017. This supply was managed as part of a Persons & Places research project.

The CDR data available to ISTAT are composed as follows:

- caller ID, that is a numeric code associated to each SIM by an algorithm that guarantees the de-identification of direct personal identifiers;
- the municipality where the call originated;
- the date and time at which the call is originated;
- the duration of the call;
- the municipality where the call ended.

For text messages, data report the date and time of the text message and the municipality from which the text message was sent.

The information concerning the call location can be obtained by using two different techniques. In the former, the localisation is based on the position of the antenna, i.e. all the calls and text messages are assigned to the municipality where the antenna tower is located. This location concentrates all calls in the municipality where the antenna is placed, even if the coverage of the antennas is very large and often covers areas that belong to more than one municipality. For example, in Figure 4.1 the antenna tower is located in the municipality of Cascina, but the coverage is in the municipalities of Cascina, Pontedera and Ponsacco. The latter methodology, applied in cooperation with the MNO (Mobile Network Operator), splits the proportion of the territory according to the Best Service Area (BSA), that is the coverage area which is best served by each sector, according to complex proprietary algorithm based on many features of the antennas and of the land, and assigns each call and text message on percentage to all the municipalities served by the specific BSA. In particular, knowing the percentage of the coverage area of each service for each municipality, and considering the uniform distribution of calls for each BSA, the percentages of calls of each BSA for each municipality were calculated according to the percentage of coverage. The following analyses were carried out by using this type of pre-processed aggregated data. The CDRs are processed so that anonymity is ensured.

The supply consists of a number of CDRs, just fewer than eighteen millions, divided into: fewer than eleven millions of Calls and seven millions of SMS. The total number of Calling SIM is just over four hundred thousand.

4.1.2. Results for population estimates

To properly use MPD for population estimates, we firstly investigated the correlation between MPD and official population estimates.

CDRs provide information on the activity of MP (mobile phone) users at a given date (with detailed time) and a very small spatial scale. However, calls-in and text messages can be used to produce population estimates given some basic assumptions, such as:

1. High level of MP penetration rate
2. High level of MP coverage over the field
3. the knowledge of the MP operator market share

One of the highest MP penetration rates in developed countries can be observed in Italy; indeed, the percentage of MP connection per 100 citizens is about 154% in 2016 , as well as a high coverage of MP networks over the territory. Moreover, working in strong cooperation with the MP operator ensures to be able to assess the market share at small spatial scale.

To investigate the correlation of MPD with official population figures derived by statistical registers, we firstly concentrated the analyses on the nighttime population. The approximation of residential population with nighttime mobile phone users has been stated in several works ((Kang et al., 2012), (Deville et al., 2014), (Douglass et al., 2015)). In this work, we identify mobile phone users with SIMs.

In cooperation with the University of Bologna, a first study was carried out by taking into consideration the localisation of the SIMs using the antenna tower position. This work highlighted the limits of this kind of localization and suggested the necessity to implement a finer geolocalization methodology. To this purpose, in cooperation with the MPO, we adopted the BSA-based approach. So, the residential municipality is assigned to each SIM according to the following procedure: a percentage is assigned to each municipality on the basis of the coverage of the BSA that most frequently registers calls-in and text messages during the nighttime. In the relative literature, the nighttime is from 8pm to 7am.

Figure 4.2 shows a scatter plot of the count of nighttime active SIMs versus the January 2017 residential population estimates for the province of Pisa at municipality level. It shows that there is a reasonable good relationship, approximately linear as depicted by the LOESS regression interpolation, in blue in the graph. In the linear regression model, the correlation coefficient is 0.94, proving the adequacy of the model in predicting residential population via the nighttime mobile phone users. Similar results in terms of high correlation are also obtained when considering logarithmic transformation and whether the extreme value represented by the city of Pisa is excluded from the analysis.

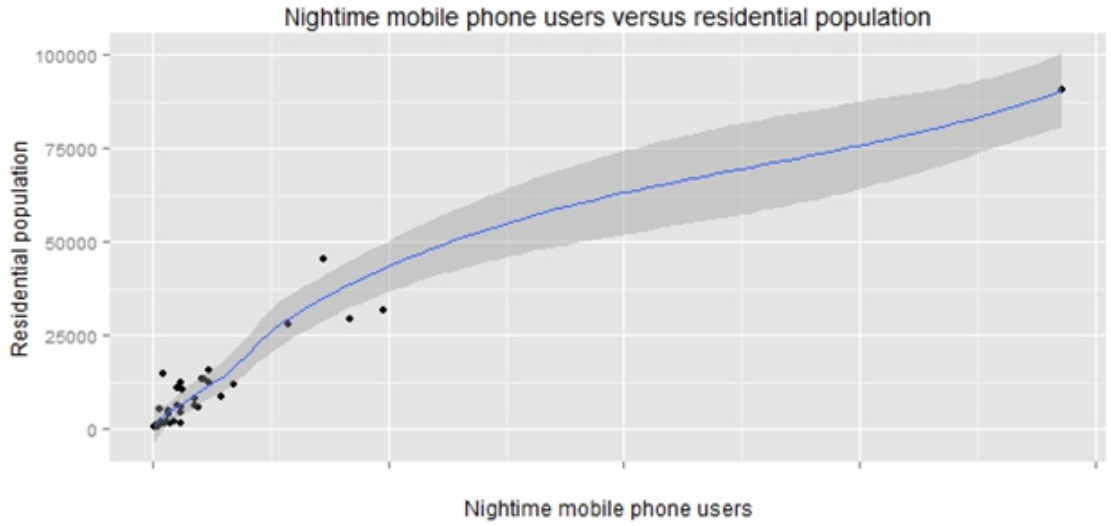


Figure 4.2: Scatter plot of nighttime mobile phone users versus residential population (municipalities in Pisa province)

The high correlation between phone users and residential population is also proved when the analysis is focused on the SIMs active during the day. In particular, we analysed SIMs calling between 5pm and 6pm: the most frequent peak hour in the observed period, from Monday to Friday. We used the phone users' population identified by these SIMs as predictor of the residential population, in this case the SIMs are assigned to the municipality resulting from the nighttime position. Figure 4.3 shows the scatter plot of the count of 5-6pm active SIMs against January 2017 residential population estimates for the province of Pisa at municipality level. We again observe a reasonably good approximation in the linear relationship, as depicted by the LOESS regression interpolation (in blue in the graph). In the linear regression model, the correlation coefficient improves up to 0.95.

These results are considered as satisfactory and encourage us to use CDR to estimate something that currently cannot be observed with sample survey and administrative data, but which could be a new output

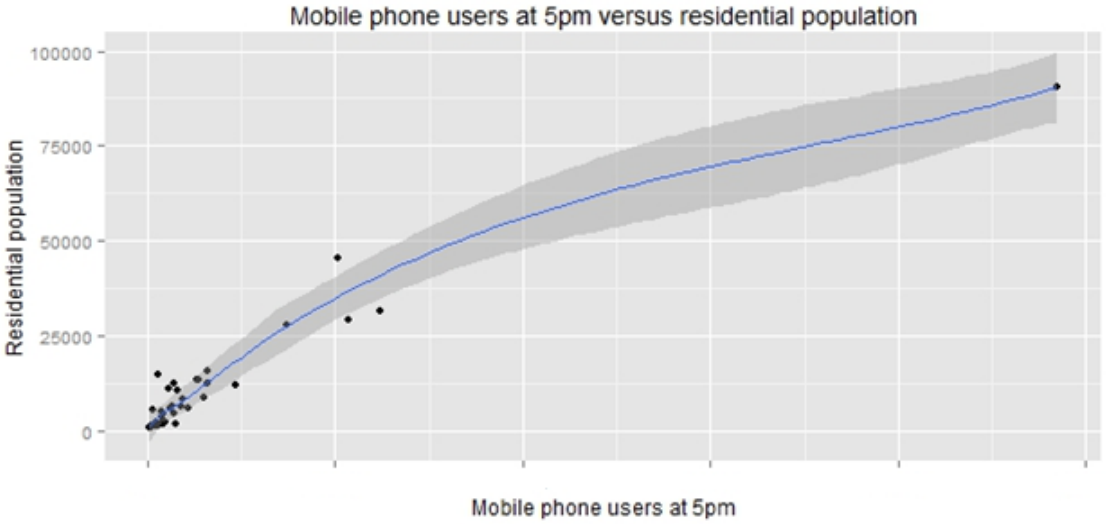


Figure 4.3: Scatter plot of mobile phone users at 5 pm versus residential population (municipalities in Pisa's province)

within official statistics.

An example of these new opportunities provided by MPD is shown in a video. The video reveals how mobile phone users' population move across the province of Pisa during daytime. The baseline time is between 4am and 6am and the video updates every two hours based on the activities of MP users across the municipalities of the province. In yellow we represented the municipalities where the MP users population density remains quite stable, i.e. changes are less than 10% compared to the baseline. In light red we represented the municipalities where the population density increases if compared to the baseline up to 20% and in dark red the municipalities with the highest increase. On the other side, in light blue we represented the municipalities where the population density decreases if compared to the baseline up to -20% and in dark blue the municipalities with the highest decrease. Another interesting dynamic in changes of the population density across the province can be shown with respect to the days of the week. These outputs exemplify some of the potentialities of the MPD if compared to the official statistics currently produced by Istat on the basis of the traditional data sources.

A further opportunity offered by the MPD is related to their use in combination with the coverage surveys included in the new population census framework. In the Census transformation program, the MPD allow us to identify areas that might be problematic for census counts, for instance, areas at risk of over or under coverage can be identified by comparing population estimates from MPD with the counts of people enrolled in registers. The risk of over/under-coverage can be defined at a very small scale and this information can be used both at the sample stage, when designing the coverage sample survey, and at the estimation stage, when small area population estimates have to be provided.

Figure 4.4 shows areas at risk of under/over coverage for the province of Pisa on the basis of comparison between MPD population estimates and January 2017 residential population. The MPD population estimates are based on the nighttime MP users by applying a multiplier estimator that takes into account the market share of the MP operators in the Pisa province. The municipalities where the MP population estimates are similar to the official estimates, with differences lower that 10% are in yellow. The municipalities with lower MP population estimates are in the light red area (up to 50% lower than the official estimates) thus highlighting a moderate risk of over-coverage. At the same time the higher ones are in the light blue area and show a moderate risk of under-coverage. The municipalities with the highest risk of over-coverage are in dark red, since the MP population estimates are lower up to 1.5 times than the counts enrolled in the

registers; instead, the data referring to under coverage, as the estimates are higher, are reported in blue.

Figure 4.4: Municipalities at risk of over/under-coverage, Pisa province

4.1.3. Mobility pattern analysis: the Origin-Destination Matrix

Another opportunity provided by the analysis of MPD is related to understanding how and where people move, the so-called mobility pattern analysis. There are at least two ways to study the mobility by means of the MPD: the first is based on the relative densities of CDRs, both across different areas and across time, as shown in the previous section; the second is based on the anonymised individual-level data. Some outputs of the first kind of analysis are illustrated in the previous section, while in this section are shown the results of a mobility analysis by using as input the second typology. In this case, a meaningful positioning for MPD, such as "home" and "work/study" are determined as follows:

1. The "home" is the municipality where a MP user is more frequently located during the nighttime, as in the previous section for the residential population estimates;
2. The "work/study" is the municipality where the MP user is repeatedly observed during the daytime.

By aggregating individual-level data for which home and work/study have previously been derived, it is possible to produce home and work/study origin-destination flows. In figure 4.5 we propose an origin-destination matrix for the Pisa province at municipal level, where only movements within the province are taken into account. The main diagonal should represent people who live ("home") and "work/study" in the same municipality. To make the graph analysis clearer, this kind of people are not taken into account, even if they represent 70% of the analysed data. The intensity of the movement is represented by the intensity of the colors on the matrix. In this case too, the results are in line with the information coming from other sources, the most commonly used routes are those involving Pisa and its nearby municipalities, as well as the municipalities where most of the largest establishments of the province are located (e.g. Pontedera).

One drawback of this analysis, in comparison with the results deriving from administrative data, is that it is not possible to detect the reason of the mobility; on the other hand, MPD allow us to assess the frequencies of the mobility, which cannot be derived from statistical registers.

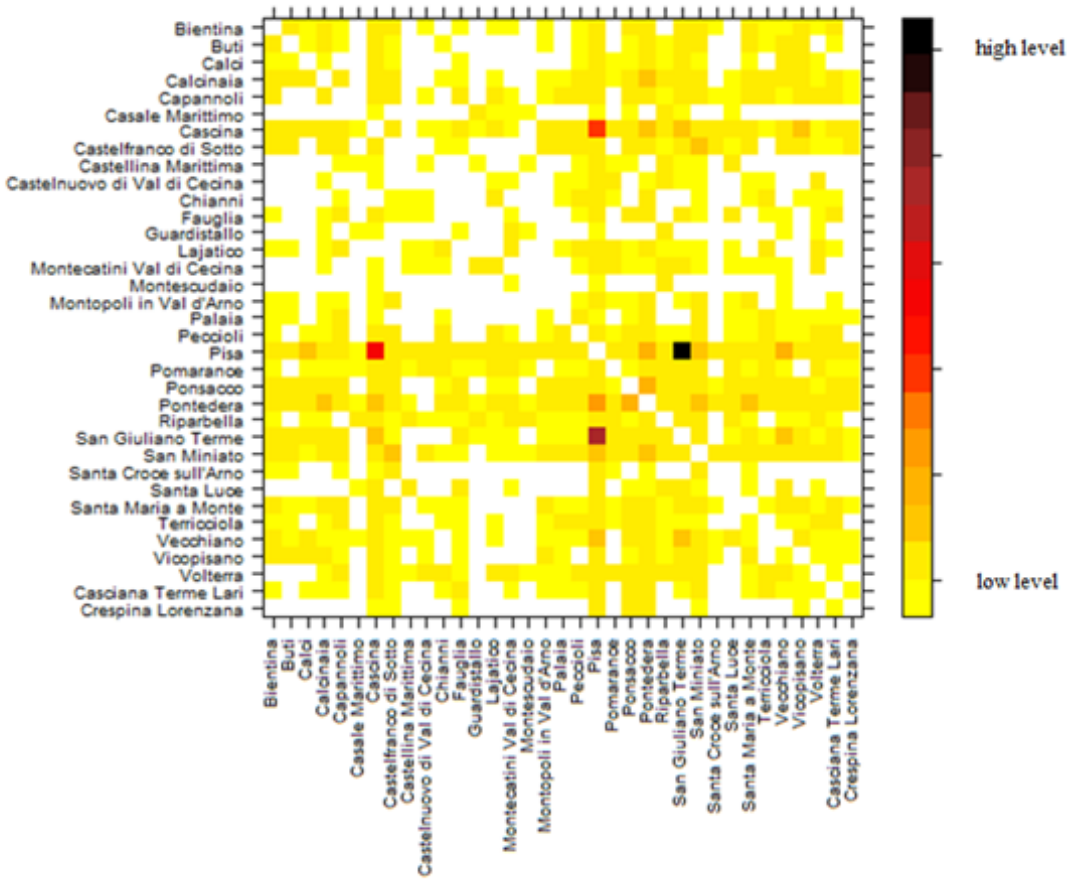


Figure 4.5: This Origin-Destination (OD) Matrix describes people movement in the Pisa Province

4.1.4. Lesson learned and next steps

The results of the analyses of the CDRs described in this report are definitively encouraging as to the potentialities of MPD both for population estimates and mobility pattern study.

A key point for a successful exploitation of CDRs is the small scale localization of the MP users' activities (calls and text messages). At present, the localization based on the position of the antenna (i.e. all the calls and text messages are assigned to the municipality where the antenna is located) shows weaknesses that compromise the reliability of population estimates, as well as all the other statistics connected with this concept. We overcame these disadvantages thanks to the cooperation with the MN (Mobile Network) operator, who provided us the proportion of territory served by the BSA. On this basis, we were able to develop a procedure that assigns each MP activity on percentage to all the municipalities served by the specific BSA, and, in this way, we highly improved the CDRs localization and obtained reliable population estimates at municipal level.

In the future, in agreement with the MPO, we will manage to produce statistics at a smaller scale than the single municipality, i.e. at census sections, so to fully exploit the huge amount of information that MPD supply us with "urban rhythms" for designing and optimizing mobility in dense urban centres.

The analyses proposed in this report are still a local observation, restricted to only one province. The availability of new data will allow us to examine these results at a wider scale, e.g. regions and finally at country level.

4.2. Data Protection Impact Assessment of CDR in Italy

This document describes in detail an assessment of the impact on the protection of the data collected by the provider, the definition of the statistical purposes and the treatments performed, the organizational responsibilities, the assessment of the risks for the privacy of citizens and the risk containment measures adopted.

The document writing follows the guidelines [2] on the conduct of DPIA established by working group 29 or WP29 [3] and it is focused on identifying and assessing all potential privacy risks for citizens involved in the data and presenting strategies for managing, mitigating or minimizing risks.

The DPIA follows a specific process divided into phases:

1. **Threshold Assessment**: the DPIA is recommended when data processing is "likely to involve a high risk to the rights and freedoms of individuals"Ḟurthermore, this is also recommended on certain types of data and if the use involves linking to other data.

   From the literature, e.g. (De Montjoye et al., 2013), CDRs are often represented as personal data. Therefore, the Italian Guarantor explicitly requested the DPIA, pronouncing a suspension of the Italian research projects related to usage of the MPDs in official statistics.

2. **Identifying Privacy Risks**: the purposes of the processing, the information flows and the organizational and IT security provisions of the project are described. With particular attention to the description of privacy management practices. The information also includes the methods and software and hardware environments used to record, archive, transmit and disseminate, and process the knowledge acquired by analyzing the data.

   A relevant part of the DPA of Istat is the definition of the risk of re-identification on the set of data provided. In this phase, the risks of each phase of the CDR processing process and the software and hardware data management architecture were defined.

[2] https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.
[3] https://edpb.europa.eu/our-work-tools/article-29-working-party_en.

3. **Evaluating Privacy Risks**: for each process that processes the data, the measures of the demands inherent in the data processing and the probability (measure) that such risks will occur and the consequences (impact) if they occur must be specified. In this phase, the benefits obtained from the acquisition of the processed information are weighted against the potential costs that follow a possible violation of privacy.

   The analyzes for the measurement of the re-identification probability confirmed the results already known from the literature (De Montjoye et al., 2013) regarding the uniqueness of the trajectories in the MPD. Actually, a trajectory defined by 4 points in space and time correspond to a unique trace for the 95% of the CDR' users. Furthermore, by decreasing the spatial and temporal resolution of the data, the power of identification decays very slowly. Unfortunately, these results are often interpreted as a sufficient condition for the identification of a mobile user, while the uniqueness in mobile data is only a necessary condition for the identification.

   Then a study was conducted by simulating the possibility of attack according to two distinct conditions of knowledge of information by the attacker which in the literature are known as: "nosey neighbour adversary" (Pedreschi, 2017) e "journalist adversary" (Pellungrini et al., 2017). In this last case, the attack simulation adapts to the knowledge of microdata available in Istat and of the risk associated with their "matching" with telephony data.

   This type of re-identification risk analysis is based on the analysis of the linkability of the data. In fact, the uniqueness together with the linkability of the data may lead to the identification of mobile users, therefore this type of analysis contributes to the definition of Data protection by design and by default (Art. 25 GDPR).

4. **Identifying Arrangements and Controls to Mitigate Risks**: the safeguards adopted to eliminate risks where possible or to reduce them must be described. A second residual risk measurement must be carried out to assess that risk reduction does not make it impossible to achieve the project objectives.

   Further interventions were requested to implement methodologies and a general Framework for the Security of processing (Art. 32 GDPR). In particular, the implementation of the pseudonymisation and encryption of personal data. The complexity of this type of operations must always be conducted "Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons".

   Therefore, as a first action, the elimination by physical or logical cancellation of the singularities in the data is adopted if that activity doesn't compromise the output. Then pre-treat the raw data to adopt aggregations that make singularities disappear from the data. This type of action is linked to the specific analysis conducted, for example in the processing of data for the realization of the commuting matrix, the single trip must be deleted, or the processing of aggregation of cells and/or time must be carried out to ensure that the trip is no longer unique.

   In reality, privacy output measures have always been adopted in statistical institutes, but the characteristics of telephony data also require the application of Input privacy measures that are commensurate with the data processing activities.

5. **Documenting the Data Protection Impact Assessment**: The report must have the appropriate approval at the level of the board of directors and the Data Protection Officer will acquire the opinion of the guarantor before starting the activities of the project.

Istat has produced the DPIA document and at the moment, it is awaiting the opinion of the Guarantor. Istat plans to launch a research project on Input privacy to be applied to MPDs.

5

Experience with real data in France

In 2020, INSEE explored with Orange Labs the construction of present population statistics based on three months of raw signalling data combined with official population data. This contribution presents the results of this collaboration. Due to privacy protection, raw signalling data were available only for a few months at the premises of the operator before being erased. The collaboration required strong synchronization between INSEE and Orange labs teams as processing of individual data was performed by each data owner. For INSEE, it represents an opportunity to develop a preliminary present population statistics paying particular attention to

- consistency of the final statistics with official measures of population.
- simplicity over complexity and easiness of implementation, in order to allow for gradual upgrades in the future depending on quality feedback on this preliminary version.

In this document, we aim at sharing this experience with other official statistical institutes and we hope to fuel future methodological improvements targeted at some identified "data challenges". We focus on estimating hourly present population at fine spatial scale over several weeks over the whole France metropolitan territory, focusing on French residents only.[1] We start with documenting the main identified challenges in mobile network signaling data for present population estimation while sketching proposals to address them. We then explain the present population methodology we used. Finally, we point out the relevant literature or explain what were the obstacles which prevented us from applying it off the shelf.

Partnership between Orange Labs and INSEE SSP Lab. Since 2016, INSEE, Eurostat and Orange Labs have been collaborating in exploring the usefulness of Mobile Phone Data (MPD) for official statistics. In this context, we have been able to benefit from the work of a French national collaborative research project (ANR Cancan[2]) which collected in 2019, 3 months of mobile signaling data. The raw data had to be deleted after 12 months. The approach described in this document only required exchanges of anonymous aggregates between INSEE and ORANGE. Processing of individual data was performed by each data owner. However, methods and algorithms starting from the raw data were developed jointly and transparently allowing both parties to evaluate and validate the outputs.

Mobile Network Data. Three months of raw signaling data from Orange clients were collected from the 16 of March 2019 to the 15 of June 2019. It included all Orange client device interactions (active and passive) with Orange metropolitan France network, for 2G, 3G and 4G.

Let us introduce the device set $\{d \in D\}$. D records all the devices in the scope, detected on the network over the period of interest. $t \in T$ is the hourly time grid of interest, each hour in three consecutive months. P is the population of interest, which is assumed constant over the period. D_t denotes devices observed during t.

[1] Henceforth, we refer as French residents for residents of metropolitan France

[2] `https://cancan.roc.cnam.fr/`

Target statistics. We aim at an estimate $\hat{u}_{i,t}$ giving the distribution of France metropolitan residents over hour t in T over several months and over tiles $i \in I$, a regular grid covering metropolitan France. As an intermediate output, we aim at estimating $\hat{u}_{i,t,r}$ giving the distribution of France metropolitan residents who are resident in location r and present in tile i at hour t. Thus, $\hat{u}_{i,t} = \sum_r \hat{u}_{i,t,r}$.

Minimal consistency constraints. For consistency, we require that the present population estimate matches an external official source: $\forall t \in T,\ \sum_i \hat{u}_{i,t} = P$. We will further require that we have as many residents of r contributing to $\hat{u}$ as there are residents of r in the official source to balance our estimates across residencies: $\forall t \in T,\ \sum_i \hat{u}_{i,t,r} = P_r$. This requires to estimate a residency for each device. It allows to break down population presence by place of residence. We note that we exclude from the onset the existence of inbound and outbound trips, absent reliable sources on daily population flows in and out the country.

Table 5.1: **Notations (Section 1)**

$d \in D$	Devices in the scope, detected on the network over the period of interest
P	The target population set
$t \in T$	Hourly time grid of interest (several weeks)
D_t	Set of devices in the scope, detected on the network during t
D_l	Set of devices in the scope, detected on the network at date l
$i \in I$	Tiles that covers the territory of interest.
$j \in J$	Cells of the MNO network
$u_{i,t}$	Population count in location i at time t (target statistics)
$\hat{u}_{i,t}$	Estimated present population in location i at time t

5.1. Challenges deserving particular attention in producing statistics from Mobile Phone Data

We begin by describing the challenges that remain in deriving a reliable present population statistics from mobile phone data.

An ideal setting for statistical purposes would be a constant equation between observed devices and units in the target population, that is $D_t \Leftrightarrow P$, in which case simple counts of devices in location i at hour t would provide our target statistics: $u_{i,t}$.[3]

In what follows, we define the device scope D as any device (as identified by its International Mobile Subscriber Identity) appearing at least 30 distinct days over Orange 2, 3 and 4G networks within the three-month time window while being identified as an Orange client (based on the Mobile Network Code of its IMSI) and as a mobile phone (based on its Type Allocation Code and an external register of known mobile phones TAC).[4] Notations are summarized in Table 5.1.

5.1.1. Counting only active mobile devices leads to unreasonable variations in aggregates

The first challenge for producing an official present population statistics from mobile phone data is that variation in counts of active devices confound many mechanisms unrelated to population variation. The large volatility of active devices total counts is in turn a major stylized fact of mobile phone data. A given device presence may be very sporadic in the collected data, *both within and across days.* Although relying on

[3] Although in this simple example, we have set aside issues regarding the location of each device from radio cell to actual tiles.

[4] Note that about 20% of daily unique identifiers are filtered as they can not be identified as mobile phones based on their TAC. We expect a large part of these excluded devices to be carried out by machines rather than persons (M2M and IoT devices, such as cameras, vehicles, alarms, sensors...).

both passive and active data improves time sampling rate, it remains a major issue to document population variations.

Within a given day, the main reasons are linked to mobile phone users' behaviour and to network coverage: users may choose to shut down their phone, may run out of battery or out of signal.[5] Across days, we may expect a large role for client churning, the telecom market in France being competitive and changing operators being increasingly easy for customers. For instance, over the two first semesters of 2018, 4 millions of mobile phone numbers were kept while changing MNO. Telco market dynamics is also at play: the sim cards number quarterly growth was of +200 000 on Q2-2019 over our period of interest (All MNOs, excluding M2M).[6] We have also to take into account failure in data collection of the richer passive data - as probes may punctually malfunction.[7] Aside these mechanisms entailing undesirable user disappearance, users may disappear as well due to outbound trips - but as for now, they are indistinguishable from the former although of interest for present population estimation.[8]

Figure 5.1 illustrates the issue on the dataset. Left Panel of Figure 5.1 represents the ratio of observed devices $\frac{|D_l|}{|D|}$ for each date l. Over a long period, aggregate variations in percentage of users observed on a given date may vary considerably: several percentage points on a regular basis, occasionally by more than 10p.p. Underlying causes of the aggregate variations remain ultimately speculative (reported data collection failure, outbound trips e.g. on bank holidays...). While we can detect relatively easily unreasonable variations at the aggregate level, we may suspect that the very same issues go undetected at local levels. Within-day, the hourly detection rate varies between 70 and almost 90% - which is arguably rather high and considerably higher than for CDR only (Right Panel of Figure 5.1). Variations seem highly driven by devices disconnecting at nighttime.

Challenge 1. Device observations have missing hours within days and may regularly be unobserved from one day to the next (for the whole day, for several days..). In practice, even when taking into account the three technologies and with passive data, the observation process at the device level is sparse for most "macro" applications intended to be comparable across time periods: $\{(d,t) \text{ such that } d \in D_t\}$ is notably bellow size $|D| \times |T|$, particularly when $|T|$ involves several days.

A good property we want to maintain for our longitudinal present population statistics is to be the least sensitive as possible to mobile device activity and data collection issues. A simple method to avoid this issue is to adopt a device-centric view and to interpolate device trajectories when unobserved - building a panel of devices trajectories. The issue seems in turn not avoidable when longitudinal data is not available - making attempts at building reliable population counts very speculative. We can not reasonably disseminate population statistics whose total may vary by 10 p.p. for network-related reasons. External calibration sources are therefore key to make the most of this rich though secondary dataset.

5.1.2. Mapping presence over the network in space

As events are located at the cell level, the location task involves mapping the information onto the grid of interest. Voronoï tessellation has been extensively used as it requires limited information (cell tower coordinates) and is simple to implement. Yet, it has been shown limited (Sakarovitch et al., 2018; Ricciato et al., 2020). In turn, probabilistic approaches accounting for overlapping cells have been advocated as

[5] Network coverage is uneven across space - and can not be ignored in a country such as France where low density areas still host a large fraction of the population.

[6] There were about 75 millions (resp. 77 millions) of active sim cards by the end of Q2-2018 (resp. Q2-2020), excluding M2M. Arcep - Services Mobiles - Q2-2018, Q2-2019 & Q2-2020 - Observatoire des marchés des communications électroniques.

[7] Probes may punctually malfunction without a too high cost for the MNO as it does not affect the network communications but is rather a monitoring tool. Thus, investing in their continuous reliability is not a priority.

[8] Characterizing places where outbound trips originate (airports, train station, borders) could be considered to discriminate absence from the territory from other phenomena.

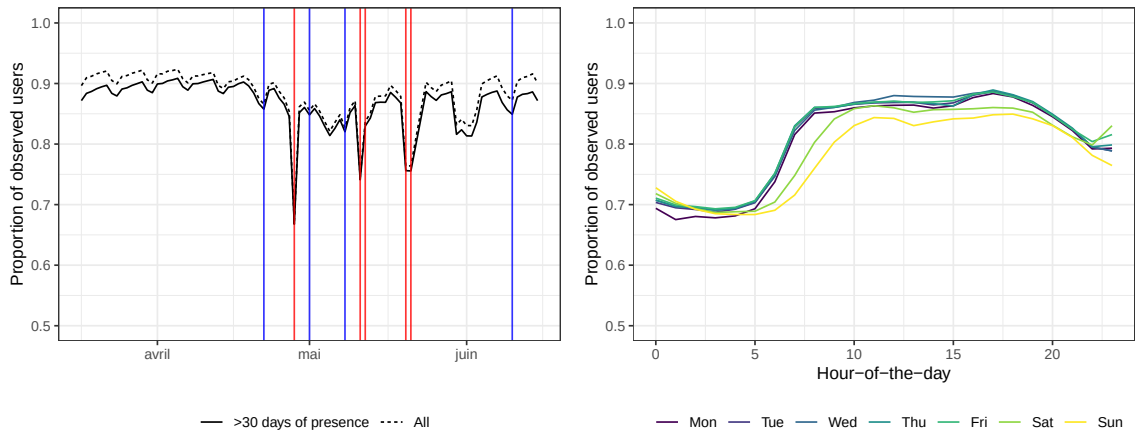


Figure 5.1: **Proportion of observed devices by date and hour-of-the day.** *Left:* % among all orange devices. We distinguish all users (dotted line) and users present at least 30 days. We represent dates with reported data collection issues (in red) and within-week bank-holiday (in blue). *Right:* % among devices appearing during the 16-31 mars 2019 period. Scope: Orange metropolitan France network, Orange-client devices identified as mobile phones.

more realistic thus preferable in a context where the mapping choice entails large discrepancies in outputs (Tennekes et al., 2020; Ricciato et al., 2020; Salgado et al., 2020).

Let us define coverage probability matrix A, such that A_{ji} represents the probability of being detected at cell j while being in tile i:

$$A_{ji} = \mathbb{P}\{\text{device detected in cell } j | \text{device in tile } i\}$$

In the Orange datasets, cell location information takes two forms. First, weekly extractions of a cell registry covering the data collection period were performed. The cell registry exists for network maintenance purpose and sees regular entries and exits following network life. The extracted data were limited to cell tower coordinates and to cell technology type. Second, Orange Fluxvision provided a cell-specific coverage map instantiating matrix A, modelises the network and devices population as of February 2019. The latter is static and is obtained from a radio-propagation model taking into account network specificities, local topography and land use, and device diversities through simulations. Applications of these data include helping in providing emergency call location.

The dimensionality of this matrix is important when taken for the whole French territory ($\approx$ 55 millions of tiles of $100m^2$ for several hundred thousands cells). If we take this matrix as a good approximation of "reality", radio cells are massively overlapping with each other over the territory and the signal is quite diluted, in the sense that the mapping cell-tiles is highly non exclusive. On average over France in the coverage map, there are 33 cells with positive coverage per 100 meter tile. However, most of these links are really low: it we restrict to links with significant coverage (say $A_{ji} > 0.1$) there are only 2.4 cells per tile and 16% of tiles without any linked cell. Figure 5.2 illustrates the distribution of areas covered (resp. significantly covered) by cells (First panel). The median cell cover significantly 25 tiles (1706 tiles with non zero coverage). The second panel of Figure 5.2 illustrates the number of cells per tile. The median tile is covered by 30 distinct cells, but significantly by 2 cells.

In practice, the location precision we can expect when mapping in space an event located at a given cell is reliant on the extent and precision of the cell covered area. If taken as the ground-truth, the estimation A encodes a relatively low network precision which is highly heterogeneous over space.

Challenge 2. Defining a mapping from cells to tiles based on A (and possibly external sources of information) as precise as possible while managing the dimensionality.

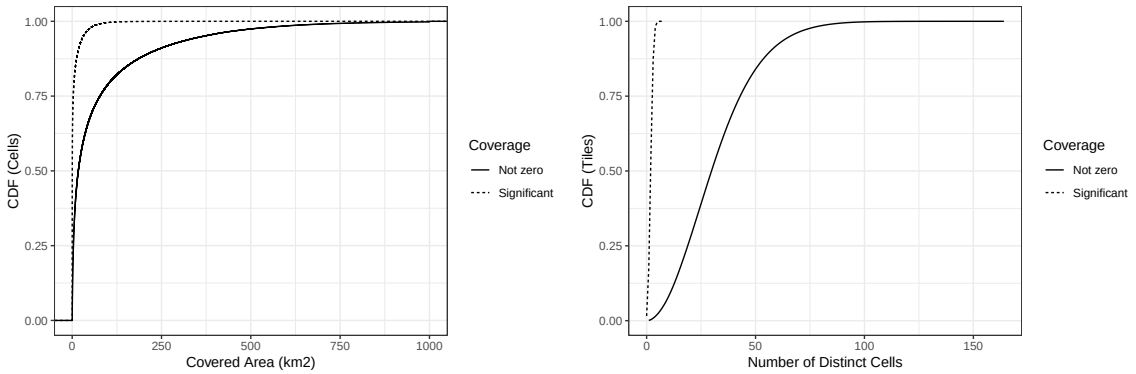


Figure 5.2: **Distributions of cells coverage areas and of cells per tiles.** Here, a tile i is said covered (significantly covered) by a cell i if $A_{j,i} > 0$ ($A_{j,i} > 0.1$).

5.1.3. Mapping devices to the population by characterizing residency

By definition, we do not observe the target population (the French residents) but a selected subset through their device(s).[9] However, we have external information on the precise location of the resident population. If we were randomly sampling devices pertaining to the French residents, we would expect no systematic association between devices characteristics and population characteristics, provided we can map both to each other. In turn, if we found (all) the determinants of inclusion in the sample, we could use these determinants to correct for sampling bias.

Characteristics of devices which can be of help are their trajectory "anchor points". From time use surveys, we can draw a picture of the typical day (including weekends and holiday) of (over-15) persons in France. On average in 2010, 8:30 hours are spent sleeping, 1:02 hour spent for washing/health care, 2:13 hours spent eating, 4:04 hours in leisure (including 2:06 hours spent watching tv), 3:10 hours in domestic work, 2:51 hours spent working or studying, 0:24 minutes of work-home commute... This average includes unemployed, retiree and housewife and may be rather heterogeneous across population type.[10] But for the "average" person, the vast majority of the time is spent at home. Therefore, if we are able to derive the place where the device spent most of its time - we would have a fair approximation for "home".

The literature on home detection based on mobile phone data has focused on Call Details Records (CDR), that is active events which are stored for billing purposes (see e.g. Vanhoof et al. (2018)). For instance, the Orange CDR data from 2007 record only a few events per device per day, and a percent of observed devices which goes from less than 5 at nighttime to about 50 percent in late afternoon (Galiana et al., 2020a). In turn, the signaling data set ensures a very large detection, even at rate at night - but the literature is still at its infancy for lack of access. In addition, computational challenges are really high here, with about 20 billions recorded events per date. In the CDR 2007, there were about 100 millions events per date.

Challenge 3: Defining a scalable method for home detection and design a validation test.

5.1.4. Active devices might not be representative of the population

Our final statistics are consistent under the assumption that the mobility and presence patterns we observe for a subset of the target population can be extrapolated to the target population. A milder assumption can be formulated when some characteristics of the devices are observed, for instance their home environment. In the application, we assume that mobility and presence patterns of unobserved

[9] It is a subset to the extent that we have a reliable method to exclude devices which are not used by human being (M2M, IoT), and that remaining Orange devices identified as mobile phones are indeed belonging to French residents which carry them along. We maintain this assumption in this work.

[10] See Ricroch and Roumier (2011) for the full picture.

residents can be extrapolated from mobility and presence patterns of persons carrying an Orange mobile device *when they share the same home environment*.

Challenge 4 Assess representativeness - which is still an open-ended question.

A related question is whether Orange clients places of residence are representative of places of residence in France. Locally, we may derive a ratio between residents and MNO-detected residents.[11] We may expect that the lower is this ratio and the better is the local representativity, as we have more devices per resident. This ratio informs on differential local representativity but also divulgates Orange market shares and is thus not reported on a map. In practice this ratio is highly heterogeneous over space. The local representativity tends to deteriorate in poor neighbourhoods in some urban areas. As an illustration, Figure 5.3 represents how this ratio distribution evolves by municipality median disposable income. At the municipality level, there is no clear association between disposable income and local representativity as captured by the ratio between residents and MNO-detected residents, except at the lower hand of disposable income. However, it hides a large heterogeneity at lower spatial scales. If the median ratio is about 0.33 detected residents per residents, the D1 ratio is 0.16 while the D9 is 0.58.

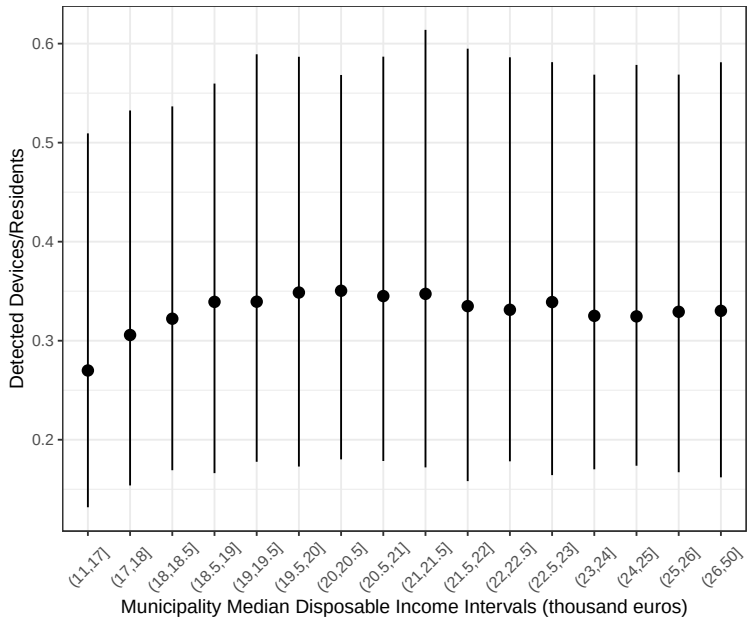


Figure 5.3: **Detected residents per actual residents and Municipality-level Disposable Income.** *Distribution (D1, Median, D9) of municipality-level ratio by municipality median disposable income range.*

On top of users' socio-economic characteristics which may differ from one MNO to another, mobile phone data is better able to capture the behaviour of active users who benefit from a good network coverage. Typically, we observe that the total number of devices connected to the 3G network is low compared to devices' presence over the 4G network. However, if each device is rescaled to represent the population of its home location - residents "potential" presence is comparable over the 3G and the 4G network. That is, devices which tend to appear less over the network may represent a large share of the population and raw mobile phone activity data under-represents population from less dense areas.

5.2. Measuring present population: a first approach

In the second part of this report, we present the methodology we use for building the hourly present population statistics over metropolitan France using 3 months of signaling data from Orange network and

[11] We describe below how we estimate this ratio.

the geography of French residents from INSEE. As we do not restrict the analysis to a sub-period or part of the territory, we favored simple method to deal with the dimensionality of the data.

Table 5.2: **Notations (Section 2)**

$j_{d,t} \in J$	Estimated *presence cell* for device d during time interval t
P_r	Official residents in place r
D_r	Devices with estimated residency in r
$\mathbf{m}_{r,t} \in \mathbb{N}^{\lvert J \rvert}$	Devices count in *presence cells* at time t with residency r
$\mathbf{u}^0 \in \mathbb{N}^{\lvert I \rvert}$	Population count from official source over tiles
$\hat{\mathbf{u}}_t \in \mathbb{R}^{\lvert I \rvert}$	Estimated population count at time t over tiles
$\hat{\mathbf{u}}_{r,t} \in \mathbb{R}^{\lvert I \rvert}$	Estimated population count at time t who are resident in r over tiles

5.2.1. Overview of the method

Our population estimation relies on several modules with building device presence panel and residency-based weighting being the critical ones.

Device presence panel. This first module role is to bypass the temporal sporadic presence of users over the network, by interpolating the trajectory of each device before any aggregation. In practice, we define and estimate for each hour and each device a *presence cell* $j_{d,t}$ - whether or not the device was observed during t. The presence cell $j_{d,t}$ is meant to represent the cell where the device d would mostly connect during the time interval t if it was active.

Residency characterization. This second module role is to estimate the residency r of each device d at an adapted geographical level to be defined. We denote D_r the set of devices with residency in place r, which can be compared to the set of residents in the official source, P_r.

We can then define presence over cells of devices residing in r, denoted $\mathbf{m}_{r,t} \in \mathbb{R}^{|J|}$ with

$$m_{j,t,r} = \sum_{d \in D_r} 1\{j_{d,t} = j\}$$

$m_{j,t,r}$ is the count of devices who are resident in r and are considered present in cell $j \in J$.

Residency-based weighting. This third module role is to extrapolate the number of devices to an estimate of the present population. If we assume that the sample D_r has been randomly chosen among P_r with sampling rate $\frac{|D_r|}{|P_r|} = \frac{1}{w_r}$, a valid estimation of the expected presence of residents of r over the network cells is $w_r \times \mathbf{m}_{r,t}$. We extrapolate the presence patterns of D_r to P_r. It amounts to apply a rescaling factor $|P_r|$ to the density of resident devices. The pseudo-weight $w_r = \frac{|P_r|}{|D_r|}$ is the ratio of residents from the official sources to the resident devices in place r. Instead of counting for 1 person, each device in the scope will participate in the aggregate with weight $w_d = w_{r(d)}$. Of course, this approach is valid if D_r is indeed close to a random sample draw from P_r. If we could add additional inferred characteristics on the devices, we could improve this stage by stratifying weights beyond residency.

Spatial Mapping. This fourth module role is to transform an active device at the cell level to an active device at the tile level. We do it by defining a linear spatial mapping $Q : \mathbb{R}^{|J|} \to \mathbb{R}^{|I|}$ which distributes a vector of presence over the network cells in the tiles with $\sum_i Q_{ij} = 1$, by specifying

$$Q_{ij} = \mathbb{P}\{\text{device mapped to tile } i \mid \text{device connected to cell } j\}$$

Then, $Q\mathbf{m}_{r,t} \in \mathbb{R}^{|I|}$ represents the presence over tiles of devices who are resident in r, at time interval t.

Presence Estimation. We estimate the presence over tiles of the residents of place r denoted $\hat{\mathbf{u}}_{r,t}$ by attributing to residents of r the presence distribution of devices who are resident in r. That is,

$$\hat{\mathbf{u}}_{t,r} = w_r Q \mathbf{m}_{r,t} \tag{5.1}$$

We finally estimate the total population presence over tiles with

$$\hat{\mathbf{u}}_t = \sum_r \hat{\mathbf{u}}_{t,r}$$

These definitions enforce our minimal consistency constraints. Note that $\hat{\mathbf{u}}_t$ can be written as reweighted projection of device-level trajectories:

$$\hat{u}_{it} = \sum_{j \in J} Q_{ij} \sum_{d \in D} w_d 1\{j_{d,t} = j\} \tag{5.2}$$

5.2.2. Implementation

The raw signaling data contain about 20 billions of events per date, totalizing 130 Tb of parquet files. It was handled using the big data framework Spark on an HDFS infrastructure at the MNO office. The Spark cluster was configured to stop any job lasting more than 24 hours. Given that the cluster was shared with other projects, the estimated resources available for the project were at maximum of 300 CPU for 1.2 Tb of RAM. The data was queried through PySpark, the python API for Spark - given available skills in the project. No algorithms beyond what could be built from PySpark API functions were used on raw signaling data - which limited device-level algorithms. Even with simple algorithms, whole-network longitudinal analysis entailing device-level sorting over a long period were challenging given the available resources.

5.2.2.1. Device-level simplifications to manage dimensionality

Given these constraints, the following simplifications were adopted in the implementation:

1. **The location information was kept at the cell-level for all device-level calculations.**
   The spatial mapping was performed on aggregates only. Aggregates can be kept for the project and exported to INSEE premises when anonymous - as opposed to device-level data. It avoids any device $\times$ tiles bottleneck operations in calculations (such as using matrix A). On one hand, this leaves the spatial mapping from cells to tiles easily adjustable downstream and leaves room for comparison among several choices of spatial mapping which have been shown to matter a lot (Ricciato et al., 2020). On the other hand, as the spatial links between cells are never considered at the device-level, some loss of spatial information is likely.

2. **For device presence, the location information is restricted to one cell per hour.**
   It seems a reasonable simplification for INSEE low-frequency purposes (at most, presence per hour). It stabilizes by design the oscillation phenomenon by which a motionless device may switch cells for network-related reasons (Katsikouli et al., 2019). However, if it is probably a good approximation for motionless devices, it is not for non-stationary devices visiting a high number of distant cells during an hour. Therefore, we oversimplify the presence of non-stationary devices by attributing them one cell on their path.

Figure 5.4 summarizes the different steps given these constraints. On top of raw signaling data and cell geographic information, this project uses the 2016 geolocalized fiscal residents counts estimated from tax sources - from which we can derive population counts in any tiles grid.[12] We denote $\mathbf{u}^0$ this official source counts.

[12]Using the same internal files, INSEE publishes social and fiscal data in a slightly more aggregated grid at https://www.insee.fr/fr/statistiques/4176305

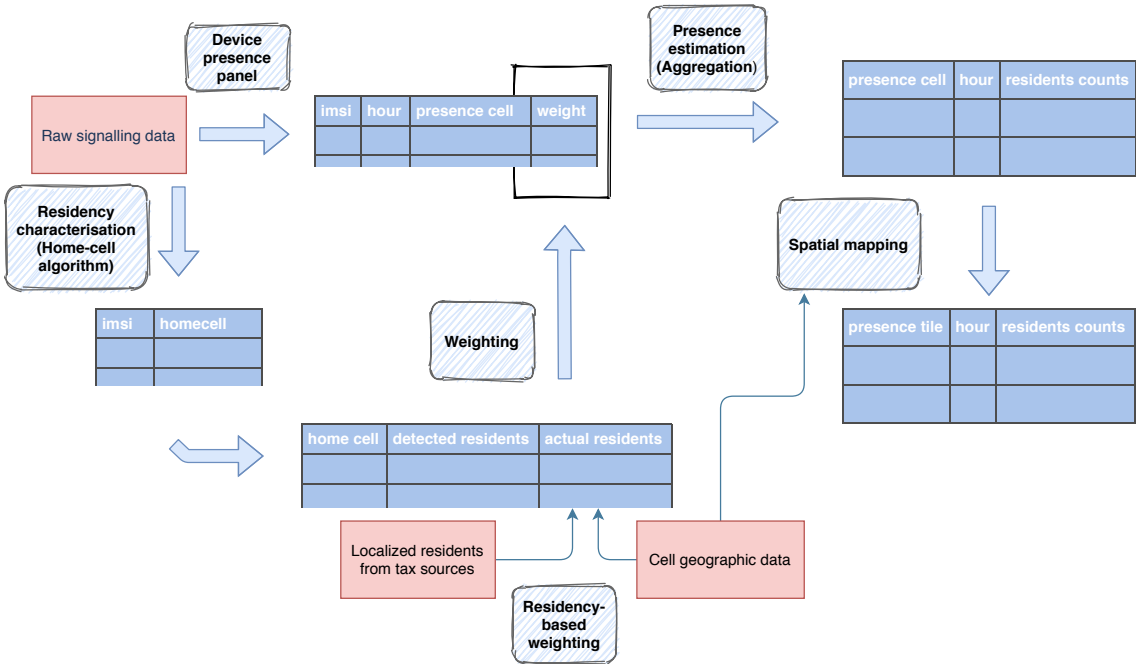


Figure 5.4: Overview of the method implementation

5.2.2.2. Device presence panel

We first filter devices which are identified as mobile phones (to filter M2M) and retain devices which are present at least 30 days out of the three months so as to ensure a relative stability of the scope (e.g. to filter movements due to client churning - irrelevant to inform on total counts).

We build the panel of presence cells $j_{d,t}$ by 24-hour rolling windows, for t in 5 a.m. to 5 a.m. the next day. When the device appears on the 24-hour window during a time interval t, the presence cell is taken as the cell recording the most events during t for this device. When it does not, the presence cell is searched within the closest time interval t' in 0 a.m. to 5 a.m. the next day when device d is observed.[13]

Figure 5.5 presents the mean interpolation error, defined as the absolute difference between the reference hour t and the hour of actual observation, used to estimate the *presence cell*. It is on average lower than one hour, but varies following the daily user behaviours and by technology. It is particularly high when the presence cell is a 2G or 3G cell.

5.2.2.3. Residency characterization

For the characterization of residency, we used the cell with the most time spent without favoring an heuristic with time spent at night although we run it as an alternative. It may suffer from an increased observation bias (Figure 5.2). In addition, we do not want to assume or constraint where the population is at night but rather deduce it from the data. For instance, we note that 1.8 millions employees work at nighttime (8p.m. to 5a.m.) more than half of their working hours a given month (Létroublon and Daniel, 2018).

According to a survey of time use, the major place of presence is the homeplace. As we are interested in recurring points of presence, we define the presence of a device at cell j in hour t as soon an event is recorded at this cell during the time interval t. For this task, a device is therefore counted as present in all

[13] We extended the search window to 29 hours for helping interpolation in the early morning with nighttime observations.

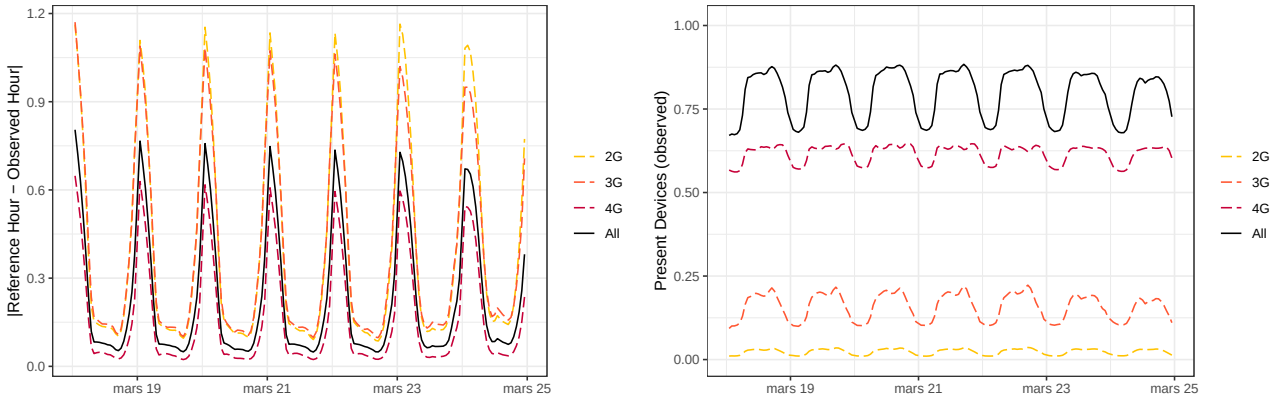


Figure 5.5: **Active devices and Interpolation in Time.** The right panel presents the proportion of active devices among all devices in dotted lines, in total and by cell technology. The left panel presents the mean interpolation error defined as the absolute difference between the reference hour t and the hour of actual observation.

cells which have detected it at some point - even for a single event.[14] Then, the max-presence cell is the cell where the device has been recorded present the most.[15] It turns out that we find evidence of at least one strong "anchor point" for most of the devices in the scope. Figure 5.6 represents the distribution of the number of distinct hours of presence in the max-presence cell when the latter is defined over two weeks. 75% of the devices in the scope are observed at least 27% of the hours over the period in the same max-presence cell. Overall, signaling data prove very promising for pinpointing anchor points.[16] If we define residency as the place with the most time spent - for sure longitudinal signaling data offer large perspectives.

This step requires us to use all events[17] and to sort the longitudinal data by device pseudo-identifier to be able to rank cells. To derive the max-presence cell over three months, we run a max-presence cell algorithm by two-weeks windows and kept per device only 10 max-presence cell candidates. Filtering the least likely candidates cells allowed to keep the computational burden manageable. Finally, we define the home-cell as the max-presence cell over the pooled max-cell candidates. At this stage, this step is highly stylized from a methodological point of view but benefits from the richness of the data over a long period.

5.2.2.4. Residency-based weighting

We then map French residents over home cells using realistic information on the coverage of each tile of 100 meters by Orange cells as provided by Orange (matrix A). Indeed, if all French residents were Orange clients, active and at home, we would expect to observe over the Orange network cells the following counts:

$$A\mathbf{u}^0$$

We define places of residency r as home cells or groups of contiguous home cells gathering at least 20 detected resident devices. This allows us to adapt our definition of residency to the MNO varying local market share, having enough devices per residency place while keeping the place of residence relatively precise. We start from all home cells, search for the closest home cells for home cells with less than 20 detected resident devices, group both and iterate until all groups of contiguous home cells reach the condition. This ensures a minimal size for D_r while keeping a high level of disaggregation in r. In turn, $P_r = \sum_{j \in r} [A\mathbf{u}^0]_j$ is the expected number of residents in the home-cell group r. Figure 5.7 presents at level r (group of contiguous home cells) and at municipality level detected residents per actual residents.

[14] This is therefore distinct from the presence cell as defined to track longitudinal presence of devices. Here, all events are used and unobserved periods are not inferred.

[15] For simplicity in computation and scalability, we kept the analysis at the cell-level but note that this approximation could probably be improved by considering several cells and their geography.

[16] Note that for this step, we did not interpolate device trajectories over an unobserved time period.

[17] In fact, all events with the cell information.

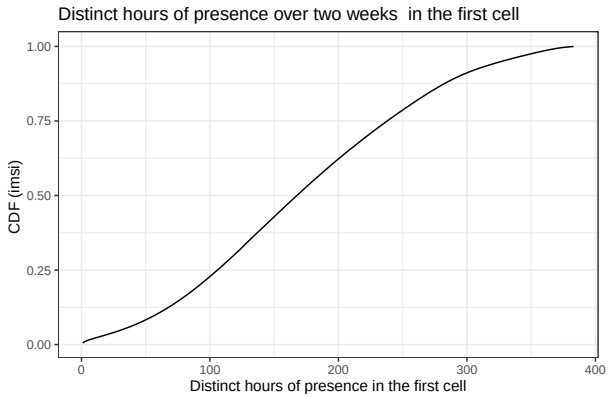


Figure 5.6: **Distinct hours of presence in the max-presence cell.** The max-presence cell is defined as the cell recording the highest number of distinct hours of presence over a two-week time period. One event within the hour is enough to consider presence in this cell at hour t. Scope: Orange metropolitan France network, Orange-client devices identified as mobile phones, 16-31 mars 2019 (384 hours).

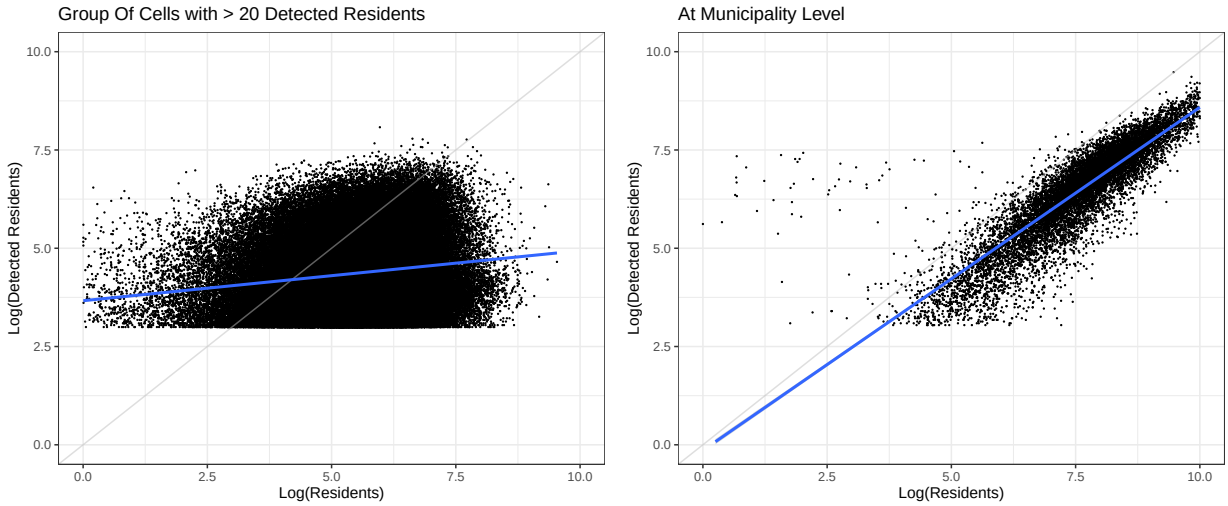


Figure 5.7: Detected Residents and Actual Residents, by aggregation level.

We define weights with the ratio of actual residents P_r divided by the network-detected residents D_r at the level of contiguous groups of home cells r which contain at least twenty detected resident devices. We end by trimming weights at their 2nd and 98th percentiles - weights fall in $[0.07, 53.5]$.

5.2.2.5. Spatial mapping

We use the coverage probability matrix A as provided by Orange Fluxvision which modelises the network as of February 2019. A_{ji} represents the probability of being detected at cell j while being in tile i:

$$A_{ji} = \mathbb{P}\{\text{device detected in cell } j | \text{device in tile } i\}$$

In particular $\sum_j A_{ji} = 1$. From a vector of presence over tile $\mathbf{u_t}$, we expect to observe on network cells $\mathbb{E}[\mathbf{m_t}] = A\mathbf{u_t}$ translating the presence of devices.[18] The estimate $\hat{\mathbf{u}}_t$ can be written in general as $\hat{\mathbf{u}}_t = g(A, \mathbf{m}_t)$ where g is a chosen *spatial mapping* (Ricciato et al., 2020). In this work, we focus on a linear estimator $\hat{\mathbf{u}}_t = Q\mathbf{m}_t$. Although any spatial mapping could be used, for our empirical results we follow Tennekes et al. (2020) who suggest to deduce Q from A using Bayes' rule by introducing a prior that reflects where the

[18] We here assume $D \Leftrightarrow P$ for clarity of exposition.

population is most likely located (e.g. based on land-use). In the results presented here, we use a uniform prior.

In addition, we propose a general framework to evaluate the location estimation precision of cellular network events. This evaluation combined with a quadtree algorithm enables us to build an adaptive spatial grid featuring small tiles for high accuracy areas and large tiles for low accuracy areas. The spatial precision is embedded within the dissemination grid.

Estimating accuracy locally. The accuracy of the linear estimator Q can be approached locally by defining the probability to localise in i a device who is in i_0 and connects to the network probabilistically through A.

$$N_{i,i_0} = \mathbb{P}\{\text{device mapped to tile } i | \text{device in tile } i_0\}$$

Formally, $N = QA$. A good estimator Q should lead to a high N_{i_0,i_0} probability (correct mapping), or at least a high probability of tiles i in the neighborhood of i_0. With previous notations, if we take the example of localizing a single device d who is in i_0, that is $\mathbf{u}_t = 1_{i_0}$, $N1_{i_0}$ can be interpreted as $\mathbb{E}[\hat{\mathbf{u}}_t | \text{device in tile } i_0]$. N encodes the spatial error by integrating the uncertainty from A and Q. $N1_{i_0}$ provides a local evaluation of spatial accuracy.

Embedding precision within dissemination. We build a quadtree which directly embeds the calculated spatial precision by gathering tiles until the probability of correct location in the macro tile I_0 (group of tiles) is higher than a threshold: $N_{I_0}(I_0) > s$. We derive present population estimates within this reduced spatial grid, which visually provide a clear idea of the achievable precision (Figure 5.8). In what follows, the tile grid $i \in I$ should be understood as this quadtree-derived grid for $s = 5\%$.

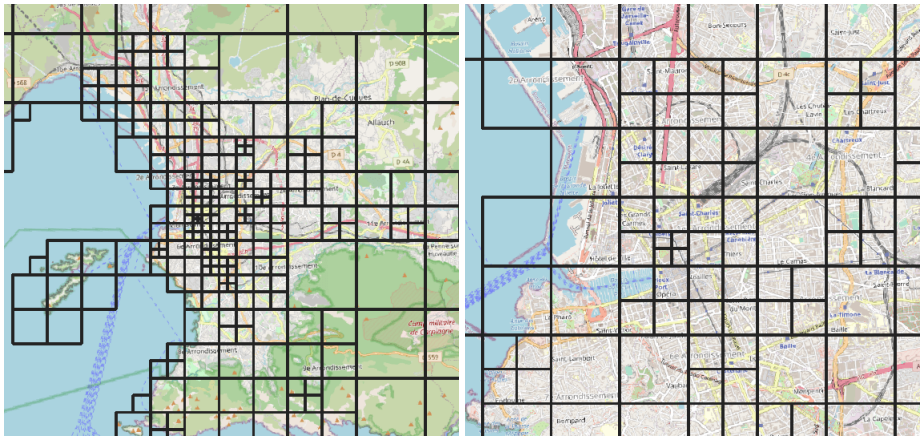


Figure 5.8: Reduced grid with a threshold $s = 1\%$. The larger the tiles, the less accurate the precision. *Note:* The grid was based on Orange matrix A and uniform prior.

5.2.2.6. Presence Estimation (Aggregation)

In practice, the set of devices present a given day varies (Figure 5.1). We denote this set $D_l \subset D$ for each date l. We only interpolate device-level trajectories within-day. To respect our consistency constraint, we finally define $w_{d,l} = w_{r(d)} \times \frac{|D_r|}{|D_r \cap D_l|}$. If all detected residents of r are here on date l, their weights remain to w_r. If some are missing, their mass is transferred to the remaining residents' devices.

Precisely, our final estimator writes, for the chosen spatial mapping Q:

$$\hat{u}_{it} = \sum_j Q_{ij} \sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j\}$$

To illustrate the difference with counts of active devices, we reproduce Figure 5.5 by re-weighting each device with $w_{d,l}$ in Figure 5.9. The mean interpolation error increases - showing that relatively less present devices have been reweighted by more than relatively more active devices. By construction, the total number of present residents is constant. Estimated presence over the 3G network is now comparable to estimated presence over the 4G network: $\sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j \ \& \ j \in \text{3G}\} \approx \sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j \ \& \ j \in \text{4G}\}$.

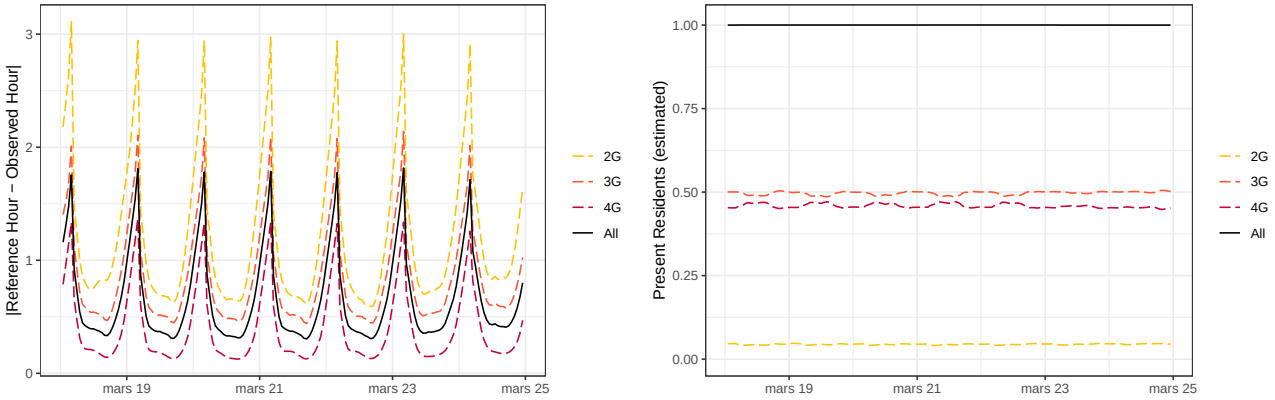


Figure 5.9: **Estimated Present Population and Interpolation in Time.** The right panel presents the proportion of estimated present residents among all residents, in total (by assumption, all residents are represented) and by cell technology. The left panel presents the weighted mean interpolation error defined as the absolute difference between the reference hour t and the hour of actual observation. Weights are $w_{d,l}$.

5.3. Results and Comparison with External Sources

The clear advantage of present population estimates derived from mobile phone data are their timeliness, their granularity in time and (relatively) in space. We first provide a rapid overview of the hourly and weekly fine-grained patterns which can be uncovered. This dynamic nature and spatial extent is rarely achievable with other sources. We then compare some present population estimate snapshots to other external, more static, sources of population density.

5.3.1. Daily and weekly cycle, local and national variations.

Figure 5.10 illustrates the daily and weekly cycles recovered from present population estimates. The 24-hour cycle features the daily pendulum movement of suburban commuters: while the present population tends to be higher in the periphery of urban areas at night, at 9a.m. these peripheries have seen their population decrease for the benefit of urban centers and the reverse in the evening. At a finer scale in the Paris surroundings, the present population estimate variations discriminate places mainly characterised by their economic, leisure and touristic activities from places mostly residential, and shows the attractiveness of a multi-polarized center. The weekly cycle discriminates the nights from Friday to Saturday and from Saturday to Sunday, where some locations in coastal areas and in the mountains fill up. In Paris, the nights from Friday to Saturday and from Saturday to Sunday have overall less present population than during the week. We however observe some nighttime excess in the present population in some places in these nights, probably reflecting nightlife activity or touristic overnight stays.

5.3.2. Comparison with external sources

Measuring the quality of present population statistics is difficult as there is no source of truth. But, we can assess how comparable are our estimates to other high-quality population measures. We here choose two points of comparisons: residents geolocalized at their tax address and day and nighttime population density estimation from Batista e Silva et al. (2020).

5.3.2.1. Resident Population (2016)

Our first comparison is with the resident population according to fiscal data from 2016. In France, this data is publicly available on a very granular grid (up to 200m granularity). The native data is geolocalized at the tax address and is here aggregated in our dissemination grid. By definition, this data only measures residents' population at their tax address. It tends to be of lower quality to measure residency for young adults when they are fiscally attached to their parents home address. Yet, it is the most granular source of resident population localisation available for France. We note that we use this datasource to build our weights.

5.3.2.2. The enhancing activity and population mapping ENACT (2011)

The second comparison source is the ENACT database that use data fusion to map population at daytime and nighttime at the European level on a 1km grid.[19] Batista e Silva et al. (2020) use a top-down approach disaggregating NUTS3 population counts based on groups' assumed place of activities using land use information. In contrast to our estimation, foreign tourists' presence is estimated and contributes to the population density.

5.3.2.3. Comparisons

We compare population densities in our dissemination grid. Figure 5.11 presents maps over France and a focus on the Paris area. At the country scale, the present population estimate respects the distribution of the population found in the other sources. However, the present population estimate tends to dilute the

[19]This data is made freely available at https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/ENACT/

(a) Daily Cycle *(b) Weekly Cycle*

(c) Daily Cycle *(d) Weekly Cycle*

Figure 5.10: **Variation of Present Population, within day and within week, at national level and in Paris.**
Note: The 24-hour within-day variations are relative to the average present population on Wednesday, March the 20th. The first image corresponds to the time interval 0 to 1 a.m. The 7 days within-week variations are relative to the week average present population at 3 to 4 a.m., from the Monday 18th to the Sunday 24th of March 2019. The first image thus corresponds to nighttime from Sunday to Monday.

population mass in space. Around dense urban areas, we observe a halo of presence absent from other sources. The first reason is the lack of precision of the cell-level localisation. Another reason is that by definition, the external sources considered here locate population in buildings. A straightforward way to close the gap between our estimate and external sources would be to use a land-use prior in the spatial mapping task. However, it may create bias in particular during daytime and during weekends. We here choose a static spatial mapping, that is, independent of t.

At the level of the Paris area, the structure of the present population distribution differs strongly during the day from during the night. Present population at 3.a.m. on a weekday (f) tends to offer a smoothed but quite accurate version of the high-resolution image of the resident population (h). We report a contextual

map of the Paris region in appendix.[20] The population variation from nighttime to daytime is similar if we consider either present population as estimated from mobile phone data from (f) to (e) or from a disaggregation of NUTS3 official sources counts as obtained in ENACT data, from (j) to (i). For instance, the core center near the *Seine* river and the *Défense* neighborhood attract population during the day, as predicted from activity-related presence with the ENACT methodology.[21]

In addition, Table 5.3 reports comparisons along three metrics: correlation, rank correlation and allocation accuracy. Allocation accuracy can be interpreted as the percentage of population density allocated in the same tiles in both sources and is defined as:

$$\mathrm{AA}(\rho_i^1, \rho_i^0) = 1 - \sum_i \frac{\frac{1}{2} \times |\rho_i^0 - \hat{\rho}_{i,t}^1|}{\sum_i \rho_i^0}$$

The present population at nighttime is as close to the resident population as is ENACT estimation during the night (slightly closer according to correlation and allocation accuracy - and farther according to rank correlation). Both present population measures get more distant from the resident population during the day, although the night/day difference is more pronounced in the MPD-derived estimation. Finally, MPD-derived presence estimation and ENACT presence estimation are closer to each other during both day and night than they are to the resident population.

Overall, at nighttime, MPD and ENACT densities fall in the same metrics range when compared with the resident population and are aligned with each other. If we had used a prior based on land uses for spatial mapping Q, we would probably be even closer to the ENACT estimation - which by definition follows land use.

5.4. Discussion

This experimental present population estimate was built with knowledge and inspiration from a number of existing works notably Salgado et al. (2020), CBS (2020) and Ricciato et al. (2020), plus discussions with ESS net colleagues which should be warmly thanked. However, we found that off-the-shelf solutions were never fully applicable to our case. In addition, a number of methodological improvements could be considered and studied in the future to make the most of this promising data source.

5.4.1. WPI proposed production framework

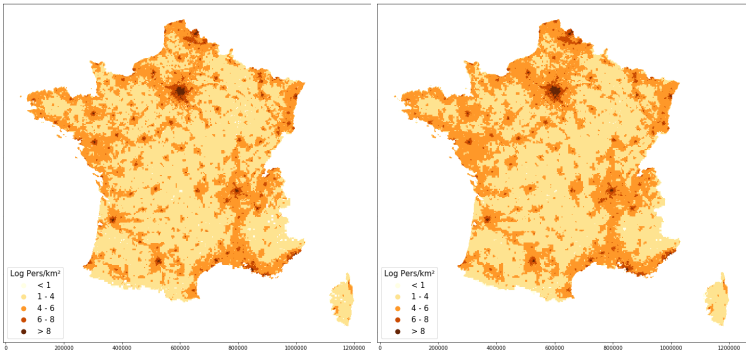
Although applying the WPI proposed production framework was appealing for a number of reasons, our approach does not fully fall in the generic framework proposed in Deliverable 3 "Proposed production framework with mobile network data", Salgado et al. (2020). We explain below differences to fuel future discussions.

Modularity. We kept in mind one of the main messages of WPI proposed production framework: modularity. The implemented method was designed to be modular and therefore easily upgraded, as Figure 5.4 shows.

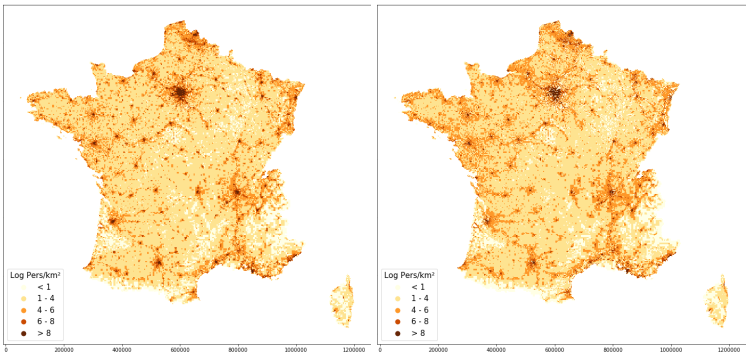
The main differences with WPI proposed production framework are

[20] For instance, the large parks in the east and the west (Boulogne and Vincennes) display a non null density according to our present population estimates, most likely due to imprecision in the spatial mapping.

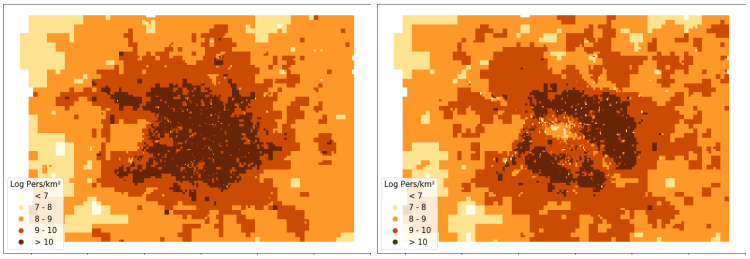
[21] Figure 5.13 in appendix presents the differences between the present population statistics and both external sources over France. It makes clearer the tendency of mobile phone data estimates to create halos around dense areas. It shows that at night in dense areas such as Paris, except for particular places such as parks, the error resembles a white noise (no tendencies over space to either overestimate or under-estimate the population compared to the resident population). During the day, it tends to offer even more contrast between places density that the ENACT estimates do.



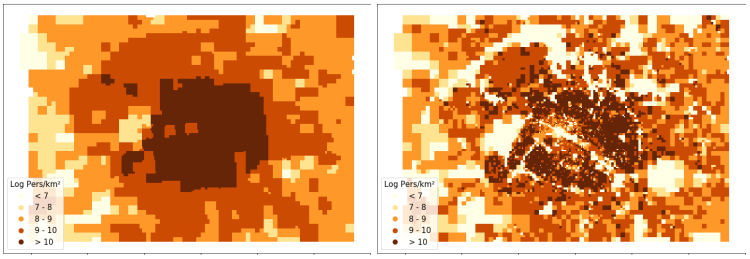
(a) Present Population at 3p.m. (week day) *(b) Present Population at 3a.m. (week day)*



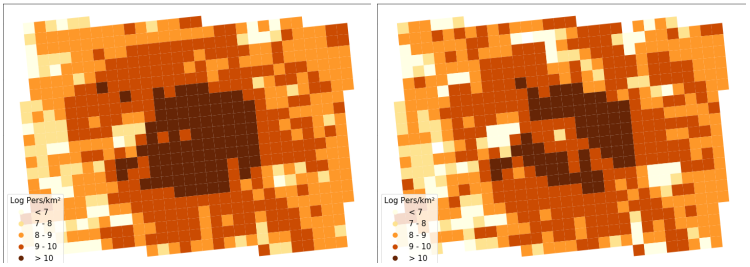
(c) ENACT Daytime Population *(d) Resident Population (Tax sources)*



(e) Present Population at 3p.m. (week day) *(f) Present Population at 3a.m. (week day)*



(g) ENACT Daytime Population *(h) Resident Population (Tax sources)*



(i) ENACT Daytime Population *(j) ENACT Nighttime Population*

Figure 5.11: **Population Densities in the Dissemination Grid (a-h).** Present Population: March 2019. ENACT: March 2011, daytime (original grid 1km² in (i-j)). Resident Population: 2016.

	Residents	ENACT	
		Day	Night
Allocation Accuracy			
Present at daytime (3 p.m.)	0.58	0.74	.
Present at nighttime (3 a.m.)	0.73	.	0.79
ENACT - Day	0.66	1	.
- Night	0.71	.	1
Correlation			
Present at daytime (3 p.m.)	0.55	0.75	.
Present at nighttime (3 a.m.)	0.81	.	0.87
ENACT - Day	0.77	1	.
- Night	0.78	.	1
Rank Correlation			
Present at daytime (3 p.m.)	0.73	0.89	.
Present at nighttime (3 a.m.)	0.75	.	0.88
ENACT - Day	0.78	1	.
- Night	0.81	.	1

Table 5.3: Population Density Comparisons. *Note:* All densities are computed in our dissemination grid, using proportional area estimation for ENACT data and direct calculations for resident density from tax files. ENACT: March 2011, Present Population: March 2019, Resident Population: 2016

1. We did not resort to a Hidden Markov Model (HMM) modelisation for device-level trajectories but a simpler interpolation method at cell-level.
2. We combine mobile phone data estimates and official sources of resident population at the device level through device-level characterisation and reweighting.
3. We perform a static spatial mapping on cell-level aggregates, as opposed to the dynamical spatial mapping at the device level delivered by the HMM model.
4. We do not propose any inference framework.
5. We did not perform any deduplication at the device-level to account for persons carrying several devices (from the same MNO).

Limitations. Of course, point 4. is a downside, as we do not compute any confidence intervals which would be in theory possible with WPI proposed production framework. Point 5. is a limit which will need to be evaluated.[22] No attempts were made at deduplication. Users carrying several Orange mobile phones will be considered as distinct residents - hence participating relatively more to aggregates compared to single carriers. Point 3. was considered to deal with the computational burden, although one non negligible advantage is to be able to vary the spatial mapping after the intensive computations step.

Scalability. As for point 1. the computational complexity of resorting to simple interpolation has nothing to do with setting up a HMM estimation in a high-dimensional states space (up to 55 millions tiles), emissions probabilities (connecting these millions tiles to hundred of thousands cells) and devices (about twenty millions three-months trajectories). Although the problem is embarrassingly parallelizable at device level, the single device problem can be quickly high dimensional in space and time in the simple setup proposed to the in the R package destim. We did not work in the direction of encoding a priori constraints

[22]For instance, is it a limit more binding than representativeness?

to match a realistic transition probability matrix over the whole France territory at our targeted scale. Thus, as noted in WPI proposed production framework futures prospects, scalability is a point for future work, to be addressed with the right computer-science skills and which will be decisive in applications.

Minimize sensitivity to active devices variations when unrelated to population variations. One of the strengths of the HMM model is to probabilistically recover trajectories when the device is unobserved, from future and past observations. Given Figure 5.1, it is a guarantee against network and behavioural effects which seems highly desirable. We see this figure as urging for longitudinal views to derive sensible statistics. We therefore resorted to a simple interpolation method.[23] Only a few works mention interpolation as a key feature for deriving reliable present population statistics whereas we tend to consider interpolation as essential for sustainable and comparable-in-time statistics. Ricciato et al. (2020) point it has a promising line for future research. Interpolation techniques were explored for mobility analysis when targeted time granularity is high and especially when working with sparse Call Detail Records data (Hoteit et al., 2016; Chen et al., 2019; Bonnetain et al., 2019). The issues of data time sparsity and sensitivity to user behaviour is generic, and apply as well when working with signaling data on a present population estimation use case. Up to now, existing literature derived snapshots of "dynamic population" (e.g. within a given day) for lack of access to longitudinal data for research. During the covid crisis when people were under lockdown, some MNO methods have shown sensitivity to changes in behaviour (increased usage of the phone, change in the timing of usage). Roughly, the increased presence over the network translated into more detected population but unexpectedly large mechanical increases in present population estimates while borders were closed. Estimates based on reprocessed multi-MNO data were conducted by INSEE (Galiana et al., 2020b). To derive longitudinal present population which is steadily comparable over a long period, a desirable property of the statistics should be to be the least sensitive to network-related and user behavioural effects. In addition, interpolation has a positive effect on representativeness if the extent of network detection is correlated with socio-economic background: relatively less active users (or users having access to a less performant mobile phone or local network) participate to aggregates more equally after interpolation relative to more active users.

Reweighting (at device-level). We argue that residency-detection is not only useful for statistical filtering, but also to balance the estimates across residency to correct for unbalanced representativeness (as illustrated in Figures 5.3 and 5.7). Imposing a constraint of equality at local level between "usual" residents detected from mobile-phone data and actual residents as estimated from official sources appears milder than one alternative that would consist in equalizing resident population P_r to population present at night: $u_{r,t_0} = P_r$. A non negligible part of the population spend nights regularly outside of its main residency or works at night. These atypical location behaviours may be relatively more captured by MPD than by traditional sources. Figure 5.10 shows how the present population at night can vary greatly during the week - and it is also even more the case for holidays and bank holidays. Note that if the pseudo-weights are based only on residency location, it could be based on as many characteristics that we can accurately recover at the device-level and for which we have an external population-level estimate.[24] This framework could be promising for future sources combination aiming at facing selectivity issues to derive representative statistics.

5.4.2. Other notable works

CBS has recently published a report on its methodology CBS (2020) which shares similarity with our implementation. Working on signaling data of one MNO in the Netherlands, they integrate a device presence estimation with a residency detection module. They perform a home-cell detection step, where the computation barrier appears to be important as well - and based on a similar heuristics. Their calibration step

[23] Although inference is clearly a plus of the HMM approach, it is balanced with its computational costs. A pragmatic approach would be to compare how the final aggregates differ would we employ one method or the other - which will probably depend on the use case (e.g. monthly populations vs fine-grain mobilities).

[24] See the discussion on pseudo-weights in Beręsewicz et al. (2018)

is based on rescaling the estimated number of active devices to the number of local residents independently of their places of presence. Implicitly, the minimal consistency constraints are therefore the same than the ones we impose.

In CBS (2020), calibration is not performed at the device-level but by rescaling aggregates of presence broken down by place of residence (the so-called *Flow cube*) register data for residence as benchmark. It is equivalent to our device-level reweighting as long as only residency is used and that no statistical disclosure control step trims aggregates at some threshold level before summing (see equations 5.1 and 5.2). One difference is that to bridge home cells and resident data, we project fine grained residential data to the network cells using matrix A, and not home-cells to tiles using a (bayesian) spatial mapping ("A^{-1}"). The other difference is that CBS (2020) does not interpolate device-level trajectories and it is only the calibration step which ensures the consistency constraints.

Ricciato et al. (2020) stresses the practical importance of the geolocation step. In this work as here, the geolocation step is performed after cell-level aggregation. Ricciato et al. (2020) and Ricciato and Coluccia (2020) propose several classes of estimators based on matrix A and cell-level aggregates which could be considered in place of our bayesian spatial mapping, for instance to deliver confidence intervals. However, only device densities are considered in these works so that we will have to work on how to integrate our minimal set of constraints to match our population totals.

5.5. Conclusion

This first version of experimental present population statistics for France is promising. However, to develop a full-fledge methodology, access is of utmost importance. A legal basis for processing MNO data for official statistics under due privacy protection, as well as cooperation of MNOs are of primary importance. It is all the more challenging today that this work tends to demonstrate that

- Some form of access to longitudinal individual data would be needed for deriving reliable population estimates, even aside interest in mobility analysis (interpolation to avoid being plagued with activity bias, home detection or device-level characterisation to ensure representativeness, deduplication ...)
- Data management from the MNO side seems decisive for the final statistics quality (network topology modelisation and cell register management, filtering of IoT and M2M, reports on data collection failure...)
- Combination of sources from the MNO and the NSI sides (e.g. build weights for representativeness) seems very promising and requires cooperation, privacy-preserving transfer of information and a transparent sharing of computations - which has been at this stage possible only within research projects.

5.6. Appendix

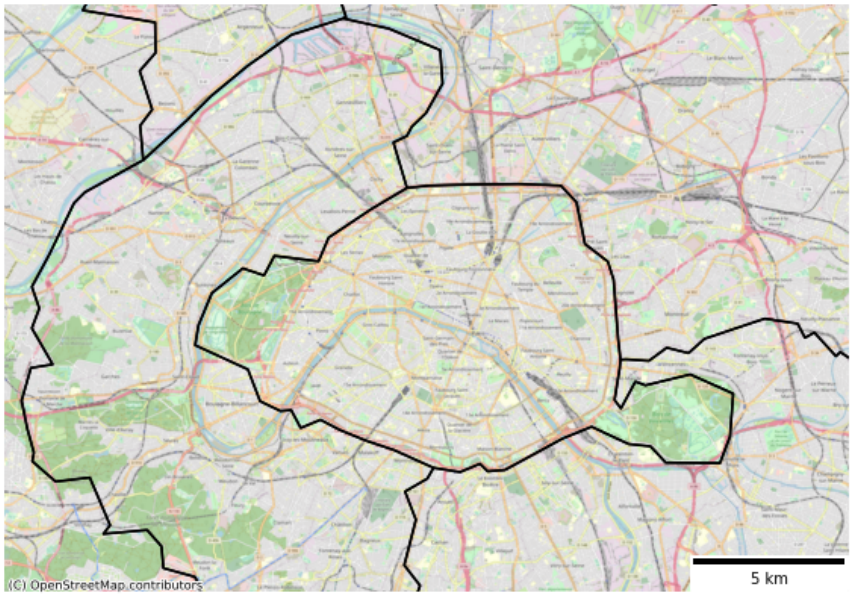
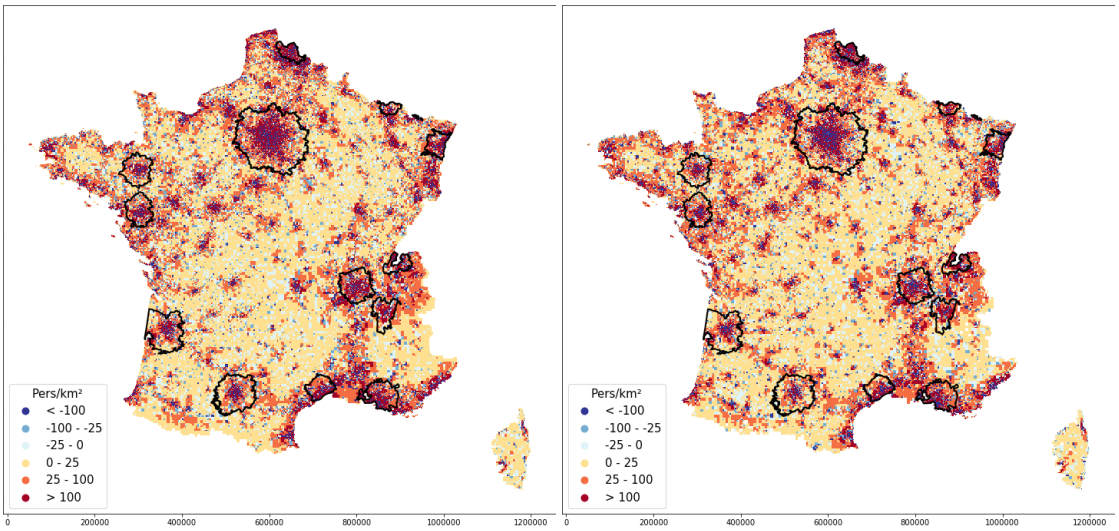
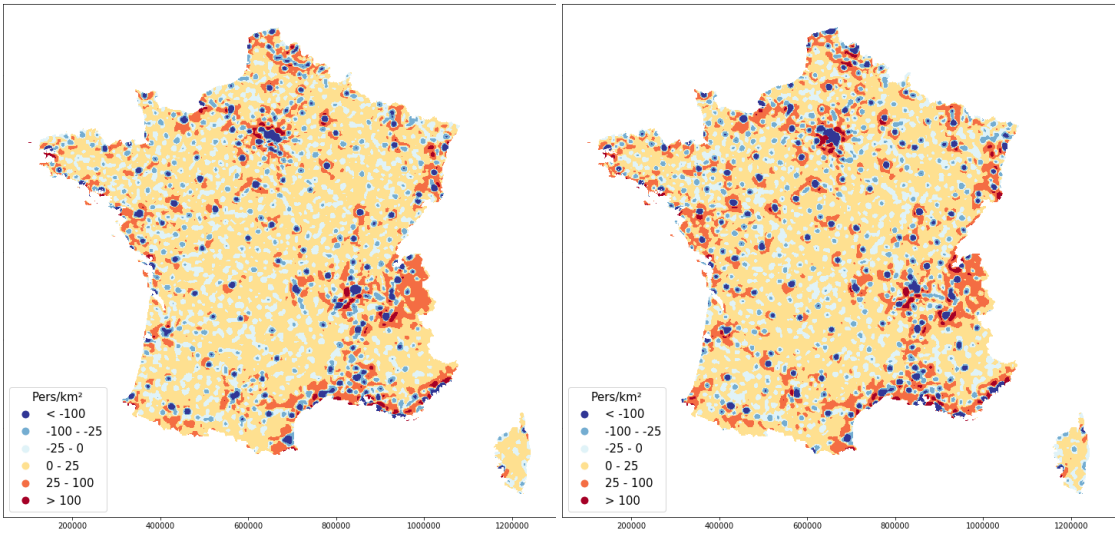


Figure 5.12: Paris area context, with *Département* administrative boundaries

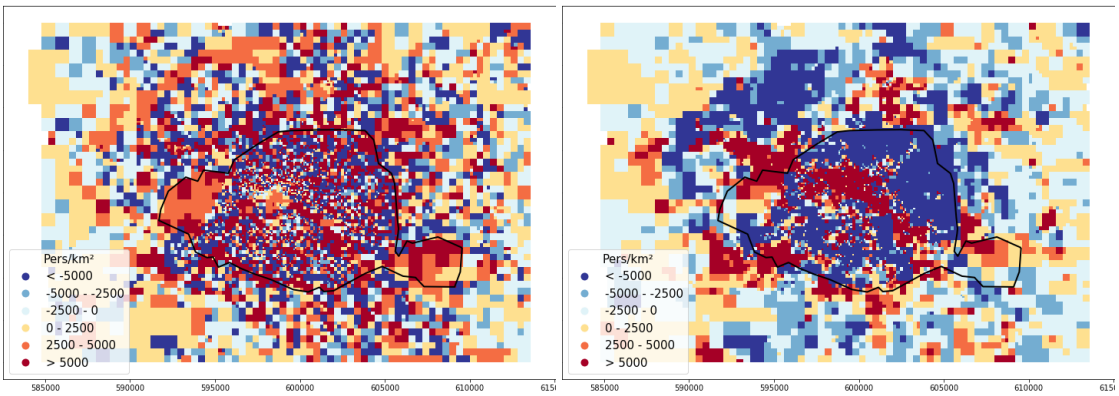
Night: Residents *Day: ENACT*



(a) Differences Present Population - External Source Population



(b) Smoothed Differences Present Population - External Source Population



(c) Differences Present Population - External Source Population

Figure 5.13: Differences between Present Population at 3a.m. and Resident Population (Left) and between Present Population at 3p.m. and ENACT Day Time Population (Right). The first panel shows urban attraction areas with more than 700 000 inhabitants

Bibliography

Batista e Silva, F., Freire, S., Schiavina, M., Rosina, K., Marín-Herrera, M. A., Ziemba, L., Craglia, M., Koomen, E., and Lavalle, C. (2020). Uncovering temporal changes in europe's population density patterns using a data fusion approach. *Nature communications*, 11(1):1–11.

Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L., and Karlberg, M. (2018). An overview of methods for treating selectivity in big data sources. Technical report, Eurostat Statistical Working Paper. Doi: https://doi. org/10.2785/312232.

Bonnetain, L., Furno, A., Krug, J., and Faouzi, N.-E. E. (2019). Can we map-match individual cellular network signaling trajectories in urban environments? data-driven study. *Transportation Research Record*, 2673(7):74–88.

Calabrese, F., G. D. Lorenzo, L. Liu, and C. Ratti (2011, 10). Estimating origin-destination flows using mobile phone location data. IEEE *Pervasive Computing 10*(4), 36–44.

CBS (2020). Estimating hourly population flows in the netherlands. *Statistics Netherlands*.

Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies 68*, 285–299.

Chen, G., Viana, A. C., Fiore, M., and Sarraute, C. (2019). Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1):30.

De Montjoye, Y.A., C.A. Hidalgo, M. Verleysen, and V.D. Blondel (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports 3*, 1376.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D. and Tatem, A.J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111(45)*, 15888–15893.

Douglass, R., Meyer, D.A., Ram, M., Rideout, D., and Song, D. (2015). High resolution population estimates from telecommunications data. *EPJ Data Science 4*, 1–13.

European Commission (2019). City data from LFS and Big Data. Final report. `https://ec.europa.eu/regional_policy/sources/docgener/studies/pdf/city_data_en.pdf`. Accessed on 23 November 2020.

Galiana, L., Sakarovitch, B., Sémécurbe, F., and Smoreda, Z. (2020a). Residential segregation, daytime segregation and spatial frictions: an analysis from mobile phone data. *INSEE's working paper*, G2020-12.

Galiana, L., Suarez Castillo, M., Sémécurbe, F., Coudin, É., and de Bellefon, M.-P. (2020b). Retour partiel des mouvements de population avec le déconfinement. INSEE ANALYSES(54).

Hadam, S. (2018). Use of mobile phone data for official statistics. *METHODS – APPROACHES – DEVELOPMENTS Information of the German Federal Statistical Office 2018(2)*, 6–9. https://www.destatis.de/EN/Methods/Quality/mad2_2018.pdf?__blob=publicationFile. Accessed on 23 November 2020.

Hadam, S., N. Würz, and A.-K. Kreutzmann (2020). Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany. Refubium - Freie Universität Berlin Repository. https://refubium.fu-berlin.de/handle/fub188/27030. doi: 10.17169/refubium-26791.

Hoteit, S., Chen, G., Viana, A., and Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the eleventh ACM workshop on challenged networks*, pages 45–50.

Kang, C., Liu, Y., Ma, X., and Wu, L. (2012). Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology) 19(4)*, 3–21.

Katsikouli, P., Fiore, M., Furno, A., and Stanica, R. (2019). Characterizing and removing oscillations in mobile phone location data. In *2019 IEEE 20th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pages 1–9. IEEE.

Létroublon, C. and Daniel, C. (2018). Le travail en horaires atypiques: quels salariés pour quelle organisation du temps de travail? *DARES ANALYSES*, 2018-30.

PDOK - Publieke Dienstverlening Op de Kaart (2019). Dataset: Actueel hoogtebestand nederland (AHN1). https://www.pdok.nl/introductie/-/article/actueel-hoogtebestand-nederland-ahn1-.

Pedreschi, D. (2017). Data ethics and privacy-preserving analytics. Lecture at the ACM Summer school, July 15, 2017. Athens, Greece.

Pellungrini, R., L. Pappalardo, F. Pratesi, and A. Monreale (2017). Fast estimation of privacy risk in human mobility data. In International Conference on Computer Safety, Reliability, and Security (pp. 415-426). Springer, Cham.

Prins, K. (2017). Population register data, basis for The Netherlands population statistics. Statistics Netherlands. https://www.cbs.nl/-/media/_pdf/2017/38/population-register-data.pdf.

Ricciato, F. and Coluccia, A. (2020). On the estimation of spatial density from mobile network operator data. *arXiv preprint arXiv:2009.05410*.

Ricciato, F., Lanzieri, G., Wirthmann, A., and Seynaeve, G. (2020). Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing*, page 101263.

Ricroch, L. and Roumier, B. (2011). Depuis 11 ans, moins de tâches ménagères, plus d'internet. *INSEE PREMIÈRE*, 1377.

Sakarovitch, B., Bellefon, M.-P. d., Givord, P., and Vanhoof, M. (2018). Estimating the residential population from mobile phone data, an initial exploration. *Economie et Statistique*, 505(1):109–132.

Salgado, D., Sanguiao, L., Bogdan, O., Barragán, S., and Suarez-Castillo, M. (2020). A proposed production framework with mobile network data. *Workpackage I Mobile Network Data Deliverable I.3 (Methodology)*.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing socio demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Social Sciences) 180(4)*, 1163–1190. doi: 10.1111/rssa.12305.

Statistisches Bundesamt (2019). Mobile phone data representing the population. https://www.destatis.de/EN/Service/EXDAT/Datensaetze/mobile-phone-data.html. Accessed on 25 November, 2020.

Statistisches Bundesamt (2020). Mobility indicators based on mobile phone data. https://www.destatis.de/EN/Service/EXDAT/Datensaetze/mobile-phone-data.html. Accessed on 25 November, 2020.

Statistics Netherlands (2017a). Continu vakantie onderzoek (cvo), vanaf 2017. https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/continu-vakantie-onderzoek--cvo---vanaf-2017.

Statistics Netherlands (2017b). Onderzoek verplaatsingen in Nederland (ovin). https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/onderzoek-verplaatsingen-in-nederland--ovin--.

Statistics Netherlands (2017c). Kerncijfers wijken en buurten 2017. https://www.cbs.nl/nl-nl/maatwerk/2017/31/kerncijfers-wijken-en-buurten-2017.

Statistics Netherlands (2017d). Wijk- en buurtkaart 2017. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2017.

Statistics Netherlands (2019). Dutch population. https://dashboards.cbs.nl/v1/dtp/.

Tennekes, M., Y. A. Gootzen, and S. H. Shah (2020). A Bayesian approach to location estimation of mobile devices from mobile network operator data. CBDS Working Paper 2020-06. https://www.cbs.nl/-/media/_pdf/2020/22/cbds_working_paper_location_estimation.pdf.

van der Valk, J., M. Souren, M. Tennekes, S. Shah, M. Offermans, E. de Jonge, J. van der Laan, Y. Gootzen, S. Scholtus, and A. Mitriaieva (2019). Experiences of using anonymized aggregated mobile phone data in The Netherlands. In *City data from* LFS *and Big Data*. European Commission. https://ec.europa.eu/regional_policy/en/information/publications/studies/2019/city-data-from-lfs-and-big-data.

Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 34(4):935–960.