Journées de méthodologie statistique de l'Insee 2022

Estimation de la précision spatiale des données de téléphonie mobile

François Sémécurbe (*), Milena Suarez Castillo(*), Tom Seimandi(*), Haixuan Xavier Tao(*), Cezary Ziemlicki (**)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale
(**) Orange Labs

francois.semecurbe@agriculture.gouv.fr
milena.suarez-castillo@insee.fr

Mots-clés. (6 maximum) : Données de téléphonie mobile ; précision spatiale.

Domaines. 11.2 ; 7.1

Résumé

Cette communication se concentrera sur les aspects spatiaux de la méthode présentée dans cet article visant à construire une statistique de population présente à partir des données passives de téléphonie mobile. L'exploitation des données de téléphonie mobile collectées au niveau des antennes relais pour produire des populations présentes est un enjeu majeur de la statistique publique. L'abondance des antennes en France métropolitaine (plus de 130000 pour les trois opérateurs historiques selon l'ARCEP en septembre 2021) masque des situations locales contrastées. Une antenne peut couvrir de larges pans de l'espace rural et réciproquement un quartier urbain peut être couvert par plusieurs dizaines d'antennes, de sorte que la précision spatiale des populations présentes est très variable entre les territoires. L'estimation de cette précision, rarement documentée, est l'objet de cette communication.

Abstract

Mobile phone data records are promising for measuring temporal changes in present population. This promise has been boosted since high-frequency passively-collected signaling data became available. Its temporal sampling rate is considerably higher than that of Call Detail Records - on which most of previous literature is based. Yet, we show it remains a challenge to produce statistics consistent over time, robust to changes in the "measuring instruments" and conveying spatial uncertainty to the end user. In this paper, we propose a methodology to estimate - consistently over several months - hourly population presence over France based on signaling data spatially merged with fine-grained official population counts. We draw particular attention to consistency at several spatial scales and over time and to spatial mapping reflecting spatial accuracy. We compare the results with external references and discuss challenges which

remain. We argue data fusion approaches between fine-grained official statistics datasets and mobile phone data, spatially merged to preserve privacy, are promising for future methodologies.

Temporally Consistent Present Population from Mobile Phone Signaling Data and Official Statistics

1 Introduction

Mobile Phone Data (MPD) has the potential to significantly enhance population statistics by increasing their levels of spatio-temporal details and their timeliness. This potential has been recognized through several international initiatives aiming at incorporating this new data source into the production of official statistics, such as the United Nation Big Data dedicated Task Team, the European Statistical System working groups or national initiatives [Statistics Netherlands, 2020, Coudin et al., 2021]. Yet, despite the new interest in timely measuring dynamic population during the Covid crisis, it has remained a great challenge for official statistics. Lack of access to real data, privacy protection and models to efficiently cooperate with private data holders have been long discussed challenges. The need to invest in a transparent methodology may be one of the distinctive feature of official statistics relative to private producers who already disseminate statistical products worldwide.

Historically, a huge body of research provided insight on human mobility by relying on Call Detail Records (CDR), data generated by the user when actively communicating through its cell phone (texting, calling...) and collected for billing purposes. The reliance on user activity and the low time sampling rate has proven an obstacle for inferring from CDR present populations fitted for official statistics purposes [Sakarovitch et al., 2018, Vanhoof et al., 2018]. Today in France and more generally in Europe, private actors producing population statistics from their networks use instead re-purposed signaling data as base material when available. This new generation of mobile phone data is characterized by a much greater spatio-temporal sampling rate. To ensure the centralized network knows about the mobile's state and location to reach it efficiently, passive communications are pervasive with the device when switched on, and are increasingly collected by Mobile Network Operators (MNOs) for network optimization and monitoring purposes. For statistical purposes, both types of data are however similar in the unitary information they contain : records convey *non-continuous* information on the proximate *radio cell* ($\approx$ antennas) mobile *devices* are wirelessly communicating with. Thus, three main dimensions of uncertainty may be listed to infer from MNOs records information on population presence : temporal, spatial and population coverage uncertainties [Ricciato et al., 2020].

To be disseminated as official statistics, dynamic population counts should be steadily comparable over time, that is the least sensitive to network-related and user behavioural effects or

MNOs client churning. They should also be consistent with other approaches to estimate population counts when other high quality datasets are available. Indeed, if many case studies exist providing dynamic maps of mobile phone users, focusing only on mobile phone users without any extrapolation to the total population is of limited interest for official population statistics [Panczak et al., 2020]. Mobile phone data not only represent an opportunity for developping countries. In developped countries, dynamic population counts be it within typical days or over the year, are not part of official statistics production today although they convey policy relevant information in many domains.

The literature on addressing spatial uncertainty is insightful. Location at radio-cell level is imprecise and depends on network local density. To model radio-cell coverage, Voronoi tessellation has been extensively used as it requires limited information (the cell tower physical coordinates) and is simply implemented by partioning space. Yet, it has been shown imprecise and is detached from network functioning [Sakarovitch et al., 2018, Ricciato et al., 2017]. In turn, probabilistic approaches accounting for overlapping cells coverage have been advocated as more realistic and 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., 2021]. As for temporal uncertainty, interpolation techniques were explored for mobility analysis when targeted time granularity is high or when working with sparse Call Detail Records data [Hoteit et al., 2016, Chen et al., 2019, Bonnetain et al., 2019]. For instance, Orange CDR data from 2007 recorded only a few events per device per day, and a percentage of observed devices which goes from less than 5 at nighttime to about 50 percent in late afternoon [Galiana et al., 2020a]. In turn, signaling data ensures a very large detection, even at night - but the literature is still at its infancy for lack of access. The availability of signaling data may allow for simpler strategies for the use case of dynamic population measurement and a reassessment on this dimension. Finally, only a specific fraction of the population is observed and thus scaling is needed. To link mobile phone data to other data sources, such as census data, home detection algorithms aim at identifying where the user lives, to rescale mobile-phone counts by residency location, e.g. as in [Fekih et al., 2021]. Absent home detection steps, cruder alternatives include rescaling factors based on mobile-phone counts at night and census counts [Deville et al., 2014]. Home detection may be considered as a data augmentation technique and rescaling could be stratified across more device characteristics to improve representativeness. Data fusion approches - ensuring coherency across several sources with distinct strengths and weaknesses - could enhance mobile phone data through the use of additional sources. Indeed, daytime and nighttime population estimates may be obtained by combining other data sources. For instance, [Batista e Silva et al., 2020] combine official statistics (residents, employees, students, tourist and other population counts and estimated flows between region) with geospatial data to produce and validate a European dataset of population grids taking into account intraday and monthly population variations, called "ENACT" [Schiavina et al., 2020].

In this article, we focus on estimating hourly population counts at a fine spatial scale over several months over France metropolitan territory. Our approach is based on spatially merging passively-collected cellular network signaling data with fine-grained official population counts by place of residence. First, we draw particular attention to consistency at several spatial scales and over time. We require that the dynamic population counts stratified by living places are locally consistent with the official number of residents. This requires to estimate a residency for each device, and allows to break down dynamic population counts by place of residence. We focus on building dynamic flows for French residents only [1] and exclude from the onset the existence of inbound and outbound trips, absent reliable sources on daily population flows in and out the country. Second, we design the spatial mapping of population counts to reflect spatial accuracy

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

by building an adapted quadtree grid - independent of administrative borders. We compare the results with the ENACT dynamic population counts. Finally, we discuss challenges which remain and argue that data fusion approaches between fine-grained official statistics datasets and mobile phone data, spatially merged to preserve privacy, are promising for future methodologies. We review how this work relates to recent initiatives in the European Statistical System to build a reference methodological framework for mobile phone data integration into official statistics production [Salgado et al., 2021, Ricciato et al., 2020, Statistics Netherlands, 2020].

Data Statement. 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 twelve 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.

2 Data

Three months of raw signaling data from Orange clients were collected from the 16 of March 2019 to the 15 of June 2019. These data included all Orange client device interactions (active, such as text or calls and passive, such as hand-overs between antennas) with the Orange metropolitan France 2G, 3G and 4G networks. All personal information was removed and device identifiers were pseudonymised before storing the data, which were erased twelve months after collection. Data are collected through probes positionned on the Orange networks for monitoring performance. Events are information exchanges between the devices and radio cells, base station antennas which communicate wirelessly with mobile devices (BTS, NodeB and eNodeB respectively for 2G, 3G and 4G networks hereafter, cell or radio cell). Users of the 4G networks generate on average about a thousand events per day, while users of the 3G and 2G networks generate respectively about 50 and 20 events per day.

In the Orange datasets, radio 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, which modelises the network coverage as of February 2019 context. 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 for these data include helping to provide emergency call locations.

The official source which serves as reference for localized population counts is fiscal data from 2016 (Filosofi). The native individual data is geolocalized at the tax address and can be aggregated over our regular grid of interest covering metropolitan France. As this data records tax addresses, it tends to be of lower quality to determine 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.

2. https://cancan.roc.cnam.fr/

Let us introduce the notations summarized in Table 1. The device set $\{d \in D\}$ is defined as all devices (as identified by their International Mobile Subscriber Identity) appearing at least 30 distinct days over Orange 2G, 3G and 4G networks within the three-month time window T while being identified as an Orange client (based on the Mobile Network Code of their IMSI) and as a mobile phone (based on their Type Allocation Code and an external register of known mobile phones TAC). [3] The hourly time grid of interest $\{t \in T\}$ contains each hour in the three consecutive months. P is the total population of interest, which is assumed constant over T. D_t denotes devices observed during t and D_l devices observed at date l. The grid $\{i \in I\}$ on which we want to estimate present population is made up of tiles, and $\{j \in J\}$ is the set of cells in the MNO network.

TABLE 1 – **Notations**

$d \in D$	Devices in the scope, detected on the network over the period of interest
$t \in T$	Hourly time grid of interest (several weeks)
P	The target population set
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 cover 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)

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 $u_{i,t}$. [4] This is not the case, since not all devices in the scope D are active at a given hour t. In fact, there are several dimensions of uncertainty in the data which must be dealt with to compute good estimates for present populations. We detail these dimensions now.

2.1 Temporal uncertainty : only counting active mobile devices leads to unreasonable variations in aggregates

The first major uncertainty observed in the data is the high temporal variation in counts of active devices. This variation confounds many mechanisms unrelated to population variation and is actually 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 we improve time sampling rate by relying on both passive and active data, this remains a major issue to document population variation.

Within a given day, the main reasons for the observed variation 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,

3. 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...).

4. Although in this simple example, we have set aside issues regarding the transposition of device locations from radio cells to the grid I, which are detailed later.

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.

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 more than 200 000 in the second quarter of 2019, over our period of interest (for 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 1 illustrates the issue on the dataset. The first panel 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 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% (second panel of Figure 1). Variations seem highly driven by devices disconnecting at nighttime. Although the detection rate is arguably rather high and considerably higher than for CDR data, only counting active mobile devices would lead to unreasonable temporal variations in aggregates.

A simple method to deal with this source of unconsistency is to account for all devices across the entire time grid by adopting a device-centric view and interpolating device trajectories when unobserved - building a panel of devices trajectories. This method is detailed in 3.

2.2 Spatial Uncertainty : Mapping presence over the network in space

Recorded events are located at the level of radio cells, not on the grid of interest I on which we estimate present populations. Spatial information on devices over time must be mapped onto the grid of interest.

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 :

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

Orange Fluxvision provided an instantiated cell-specific coverage map matrix A, which modelises the network and devices population as of February 2019.

The dimensionality of matrix A is high when considering the whole French territory (approximately 55 millions of tiles with sides of 100 meters for several hundred thousands cells). If we take this matrix as a good approximation of "reality", radio cells are massively overlapping

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.

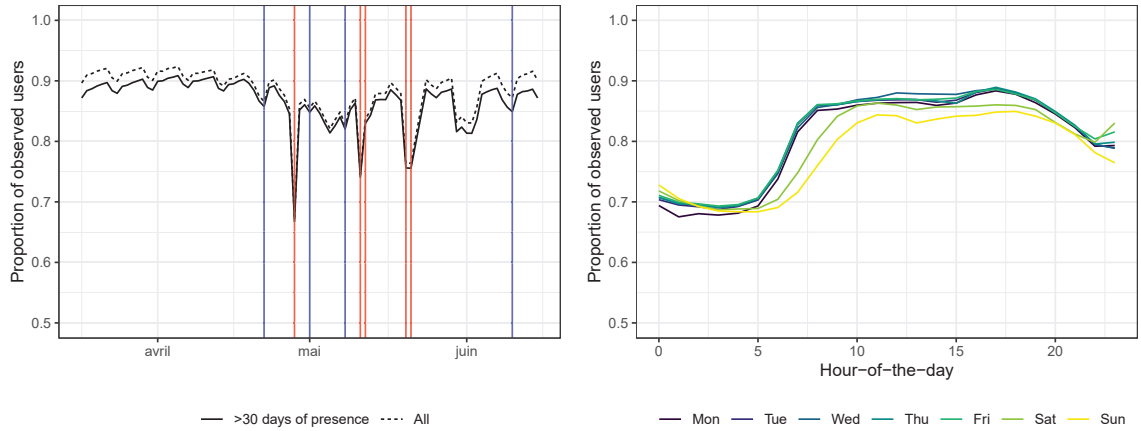


FIGURE 1 – **Proportion of observed devices by date and hour-of-the day.** *Left :* % among all orange devices. We distinguish all users (dotted line) and users present at least 30 days. We represent dates with reported data collection issues (in red) and within-week bank-holiday (in blue). *Right :* % among devices appearing during the 16-31 mars 2019 period. Scope : Orange metropolitan France network, Orange-client devices identified as mobile phones.

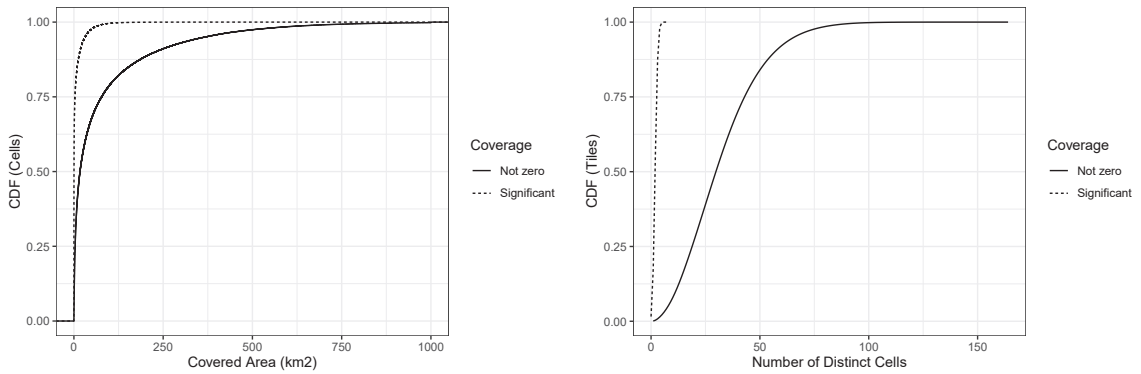


FIGURE 2 – **Distributions of cells coverage areas and of cells per tiles.** Here, a tile i is said covered (significantly covered) by a cell i if $A_{j,i} > 0$ ($A_{j,i} > 0.1$).

with each other over the territory and the signal is quite diluted, in the sense that the cell-tiles mapping 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. The first panel of Figure 2 illustrates the distribution of areas covered (resp. significantly covered) by cells. The median cell covers significantly 25 tiles (1706 tiles with non zero coverage). The second panel of Figure 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. The method to map cells to tiles based on A as precisely as possible while managing dimensionality will be detailed in 3.

2.3 Coverage uncertainty : mapping devices to the population

A third dimension of uncertainty in the data is linked to population coverage. By definition, we do not observe the target population (all French residents) but a selected subset through their device(s). [9] Present population estimates relying on device counts are consistent under the assumption that the mobility and presence patterns we observe for the selected subset of the target population can be extrapolated to the target population. However, active devices might not be representative of the population. A milder assumption can be formulated when some characteristics of the devices are observed, that mobility and presence patterns of unobserved residents can be extrapolated from mobility and presence patterns of persons carrying an Orange mobile device when they share these characteristics. High-frequency signaling data fittingly allows to observe one such characteristic - the home environment.

A related question is then whether the places of residence of Orange clients are representative of places of residence in France. Locally, we may derive a ratio between residents and MNO-detected residents. [10] We expect that the lower this ratio is, the better is the local representativity, as we observe 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 it is highly heterogeneous over space. The local representativity tends to deteriorate in poor neighbourhoods in some urban areas. As an illustration, Figure 3 represents how this ratio distribution evolves by municipality median disposable income. At the municipality level, there is no clear correlation between disposable income and local representativity as captured by the ratio between residents and MNO-detected residents, except at the lower hand of disposable incomes. If the median ratio is about 0.33 detected resident per resident, the D1 ratio is 0.16 while the D9 is 0.58. A large heterogeneity may also exist at lower spatial scales.

On top of users' socio-economic characteristics which may differ from one MNO to another, mobile phone data is able to better capture the behaviour of active users who benefit from a good network coverage. In this case, raw mobile phone activity data under-represents population from less dense areas where the network is less developed.

9. It is a subset to the extent that we have a reliable method to exclude devices which are not used by human beings (M2M, IoT), and that remaining Orange devices identified as mobile phones indeed belong to French residents which carry them along. We maintain this assumption throughout this work.
10. This is detailed in 3

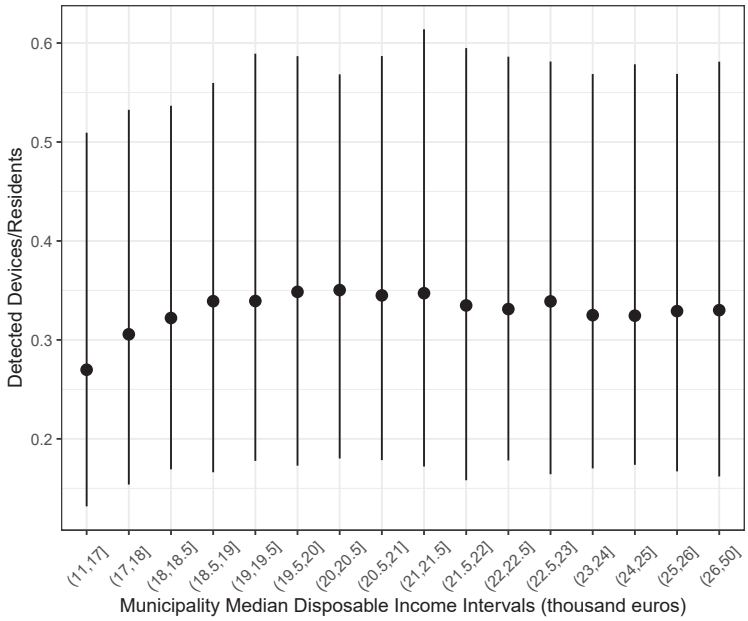


FIGURE 3 – **Detected residents per actual residents and Municipality-level Disposable Income.** *Distribution (D1, Median, D9) of municipality-level ratio by municipality median disposable income range.*

3 Measuring present population : a first approach

We now present the methodology we use to compute hourly present population estimates over metropolitan France using 3 months of signaling data from Orange network and the geography of French residents from fiscal data. Formally, we aim at computing estimates $\hat{u}_{i,t}$ giving the distribution of France metropolitan residents at hours t in T over several months and over tiles $i \in I$. As an intermediate output, we aim at estimating $\hat{u}_{i,t,r}$ giving the distribution of France metropolitan residents in location r 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 total 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, due to the absence of reliable sources on daily population flows in and out the country.

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

TABLE 2 – **Additional notations**

$j_{d,t} \in J$	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

3.1 Overview of the method

Our population estimation relies on several modules, the critical ones being the construction of a device presence panel and residency-based weighting.

Device presence panel. This first module's 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's 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 $\mathbf{m}_{r,t} \in \mathbb{R}^{\lvert J \rvert}$ with

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

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

Residency-based weighting. This third module's 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 applying 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's 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 which reside 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,

$$\hat{\mathbf{u}}_{t,r} = w_r Q \mathbf{m}_{r,t} \tag{1}$$

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 a reweighted projection of device-level trajectories :

$$\hat{u}_{i,t} = \sum_{j \in J} Q_{ij} \sum_{d \in D} w_d 1\{j_{d,t} = j\} \tag{2}$$

3.2 Implementation

The raw signaling data contain about 20 billions of events per date, totalizing 130 Tb of parquet files. They were 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.

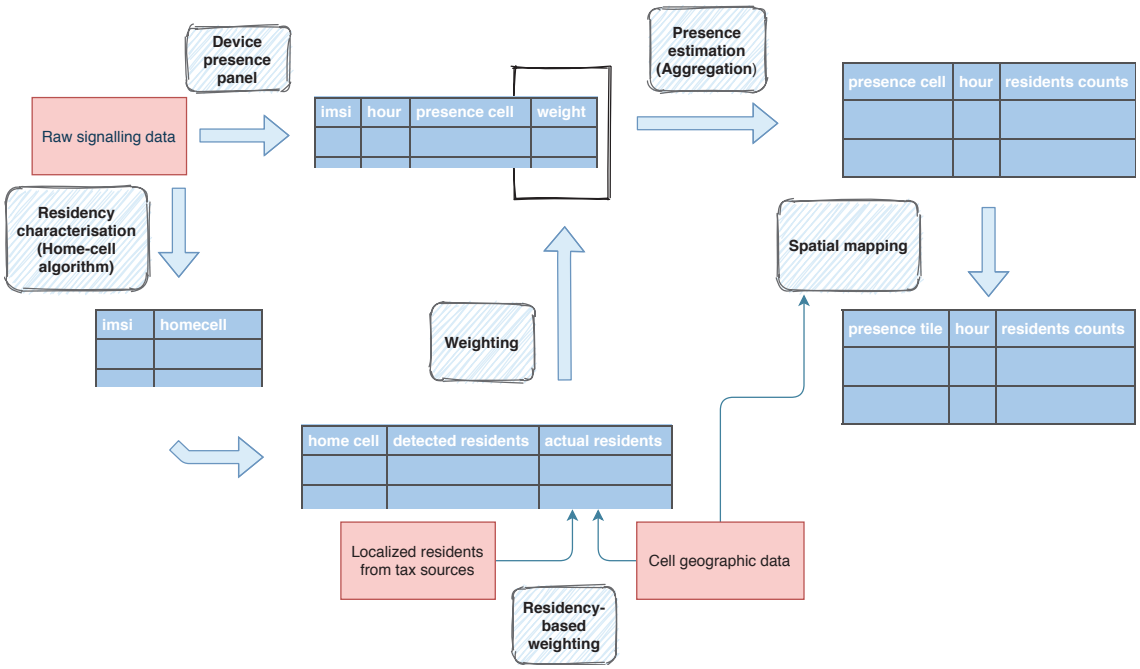


FIGURE 4 – Overview of the method implementation

3.2.1 Device-level simplifications to manage dimensionality

Given these constraints, the following simplifications drove the choice of the method presented in 3.1 and were adopted in the implementation :

1. **For device presence, the location information is restricted to one cell per hour.** It seems a reasonable simplification for national statistical institute 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.
2. **The location information was kept at the cell level for all device-level calculations.**
   We were allowed to export aggregates to the national statistical institute premises, as opposed to device-level data. Hence, the spatial mapping was performed on aggregates only. This avoids bottleneck operations on $|D| \times |I|$-sized matrices during calculations (such as operating involving A on such matrices). On one hand, this leaves the spatial mapping from cells to tiles easily adjustable downstream and leaves room for comparison between several choices of spatial mapping, which has 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.

Figure 4 summarizes the different steps of the method which are now described in more depth.

3.2.2 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

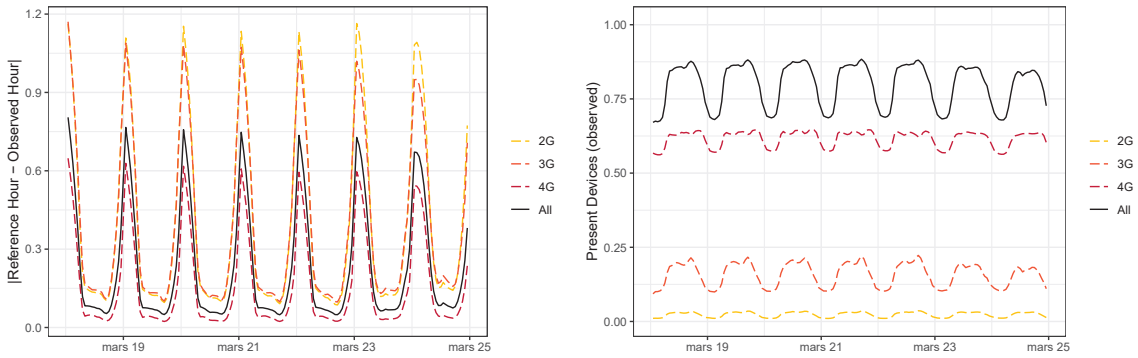


FIGURE 5 – **Active devices and Interpolation in Time.** The right panel presents the proportion of active devices among all devices in dotted lines, in total and by cell technology. The left panel presents the mean interpolation error defined as the absolute difference between the reference hour t and the hour of actual observation.

of the scope (e.g. to filter movements due to client churning as they are 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 during a time interval t, the presence cell at t is taken as the cell recording the most events. When it does not, the presence cell is searched within the closest time interval t' when the device is observed in a window going from 0 a.m. to 5 a.m. the next day. [11]

Figure 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.

3.2.3 Residency characterization

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 and 0 :24 minutes of work-home commute. This average includes unemployed, retirees and housewifes and may be rather heterogeneous across population type. [12] However, for the "average" person, the vast majority of the time is spent at home.

For the characterization of residency, we therefore choose the cell with the most time spent rather than favoring a heuristic using time spent at night although we run it as an alternative. It may suffer from an increased observation bias (Figure 2). In addition, we do not want to assume or constrain 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].

11. We extended the search window to 29 hours for helping interpolation in the early morning with nighttime observations.
12. See [Ricroch and Roumier, 2011] for the full picture.

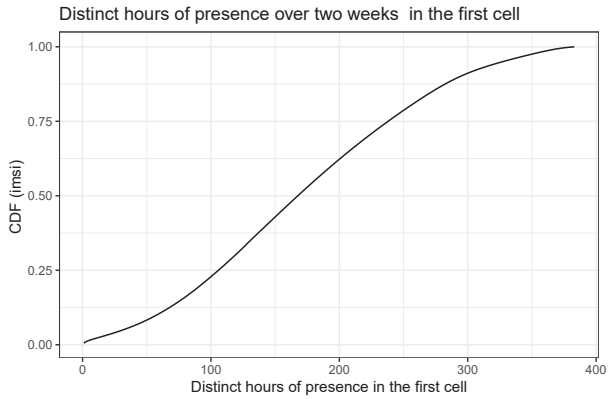


FIGURE 6 – **Distinct hours of presence in the max-presence cell.** The max-presence cell is defined as the cell recording the highest number of distinct hours of presence over a two-week time period. One event within the hour is enough to consider presence in this cell at hour t. Scope : Orange metropolitan France network, Orange-client devices identified as mobile phones, 16-31 mars 2019 (384 hours).

To estimate the residency of each device, we thus want to identify recurring points of presence. A device is present 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 cells which have detected it at some point - even for a single event. [13] Then, the max-presence cell is the cell where the device has been recorded present the most. [14] It turns out that we find evidence of at least one strong "anchor point" for most of the devices in the scope. Figure 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. [15] 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 [16] 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.

3.2.4 Residency-based weighting

In this step, we 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). Let us denote $\mathbf{u}^0$ the resident counts over grid I estimated from 2016 fiscal data. If all French residents were Orange clients, active and at home, we would expect to observe over the Orange network cells the following counts :

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

13. 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.

14. 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.

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

16. In fact, all events with the cell information.

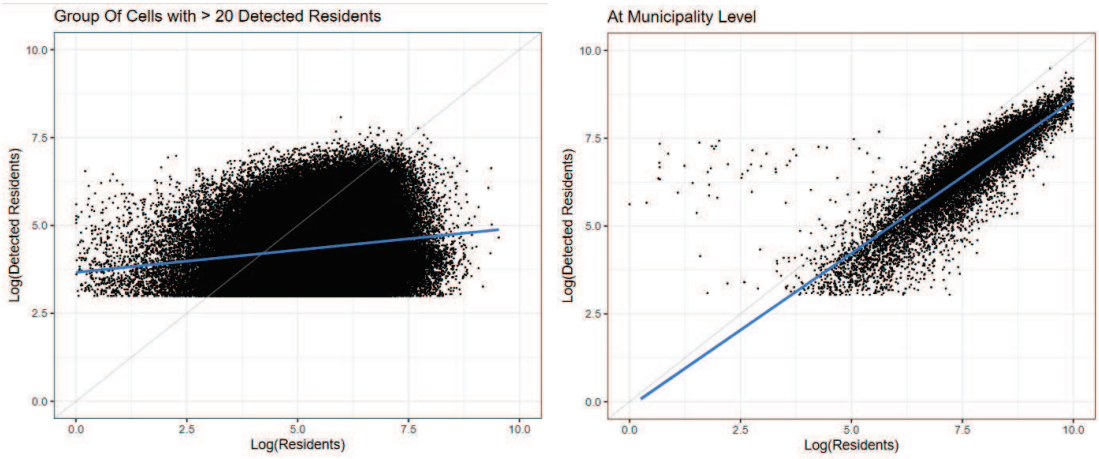


FIGURE 7 – Detected Residents and Actual Residents, by aggregation level.

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 shares, 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 7 presents at level r (group of contiguous home cells) and at municipality level the number of detected residents against actual residents, hinting at weights heterogeneity.

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 $[0.07, 53.5]$.

3.2.5 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 :

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

In particular $\sum_j A_{ji} = 1$. From a vector of presence over all tiles $\mathbf{u_t}$, we expect to observe on network cells $\mathbb{E}[\mathbf{m_t}] = A\mathbf{u_t}$ translating the presence of devices. [17] 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 population is most likely located (e.g. based on land-use). In the results presented here, we use a simpler uniform prior. Specifically,

$$Q_{ij} = \frac{A_{ji}}{\sum_{i'} A_{ji'}}$$

17. We here assume $D \Leftrightarrow P$ for clarity of exposition.

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 which 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 8). In what follows, the tile grid $i \in I$ should be understood as this quadtree-derived grid for $s = 5\%$.

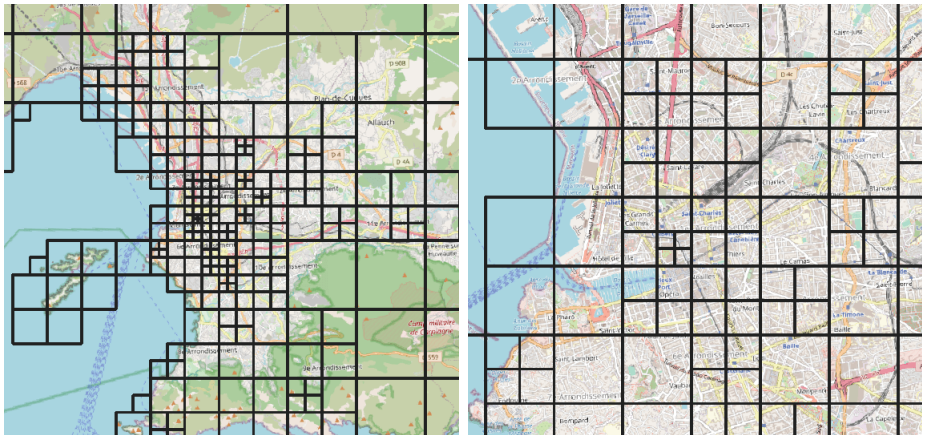


FIGURE 8 – Reduced grid with a threshold $s = 1\%$. The larger the tiles, the less accurate the precision. *Note :* The grid was based on Orange matrix A and uniform prior.

3.2.6 Presence Estimation (Aggregation)

In practice, the set of devices present a given day varies (Figure 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 are set 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 :

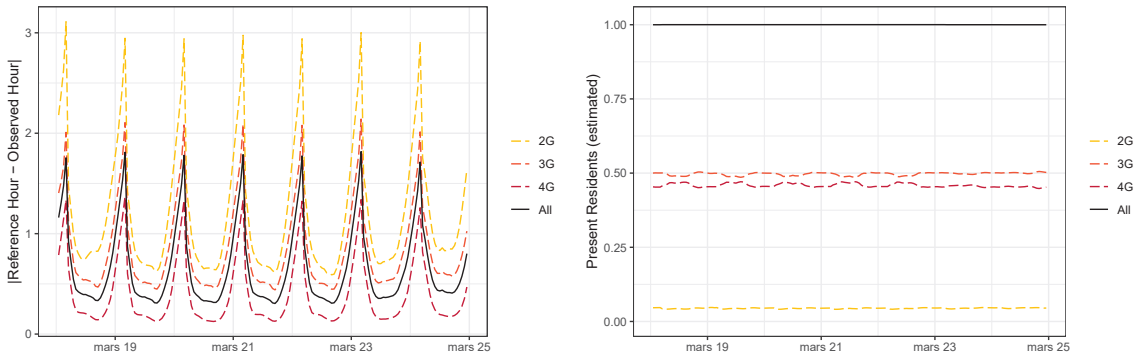


FIGURE 9 – **Estimated Present Population and Interpolation in Time.** The right panel presents the proportion of estimated present residents among all residents, in total (by assumption, all residents are represented) and by cell technology. The left panel presents the weighted mean interpolation error defined as the absolute difference between the reference hour t and the hour of actual observation. Weights are $w_{d,l}$.

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

To illustrate the difference with counts of active devices, we reproduce Figure 5 by reweighting each device with $w_{d,l}$ in Figure 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 : $\sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j \ \& \ j \in 3G\} \approx \sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j \ \& \ j \in 4G\}$.

4 Results and Comparison with External Sources

The clear advantages 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.

4.1 Daily and weekly cycle, local and national variations.

Figure 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.

4.2 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 our estimates are to other high-quality population measures. We here choose two points of comparison : residents geolocalized at their tax address and day and nighttime population density estimation from [Batista e Silva et al., 2020].

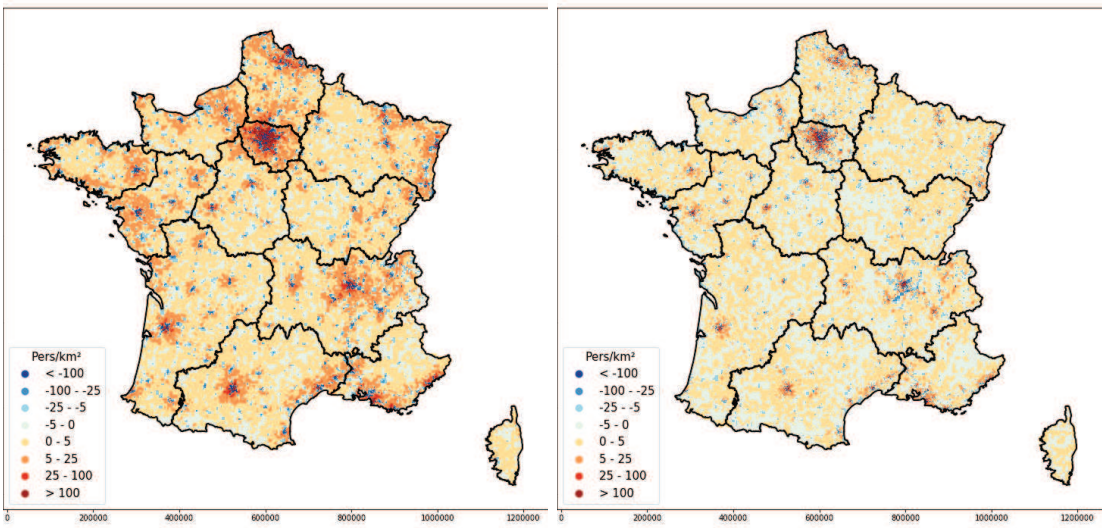
Our first comparison is with the resident populations $\mathbf{u^0}$ computed according to fiscal data from 2016. Note that we use this datasource has also been used to build our weights.

The second comparison source is the ENACT database that uses data fusion to map population at daytime and nighttime at the European level on a 1km grid. [18] [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.

We compare population densities in our dissemination grid. Figure 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 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 chose a static spatial mapping, that is, independent of t.

18. 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*

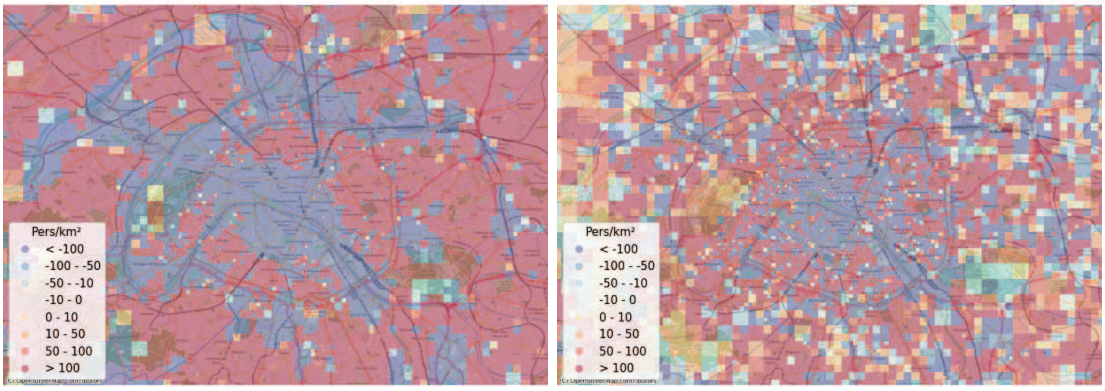
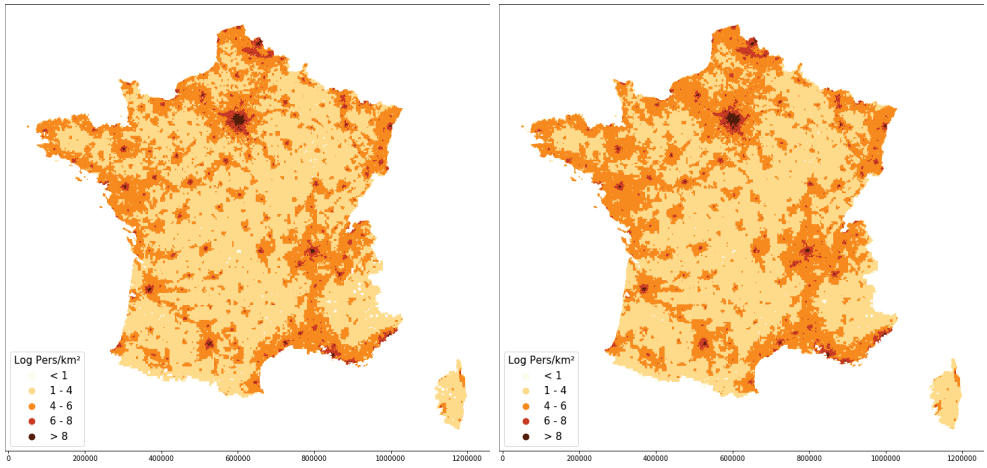


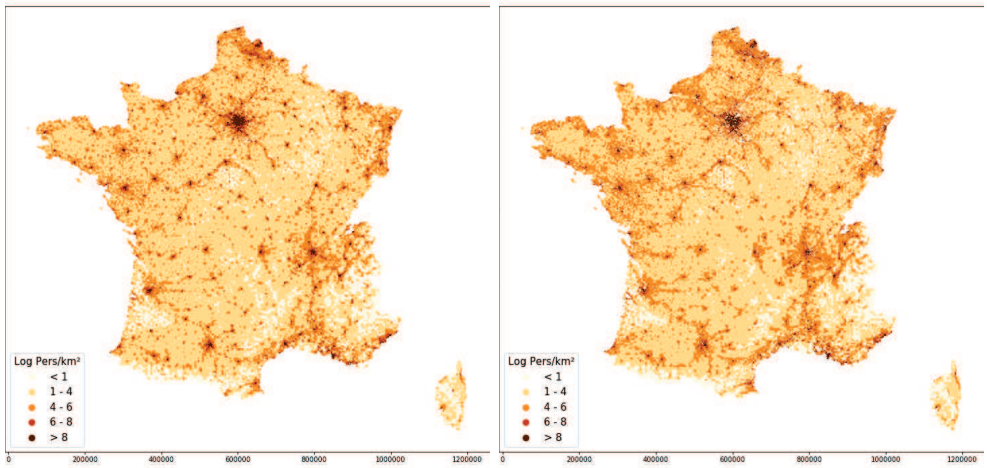
FIGURE 10 – **Variation of Present Population, within day and within week, at national level and in Paris.** *Note :* The 24-hour within-day variations are relative to the average present population on Wednesday, March the 20th. The first image corresponds to the time interval 0 to 1 a.m. The 7 days 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.

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. [19] 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

19. 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.



(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)*

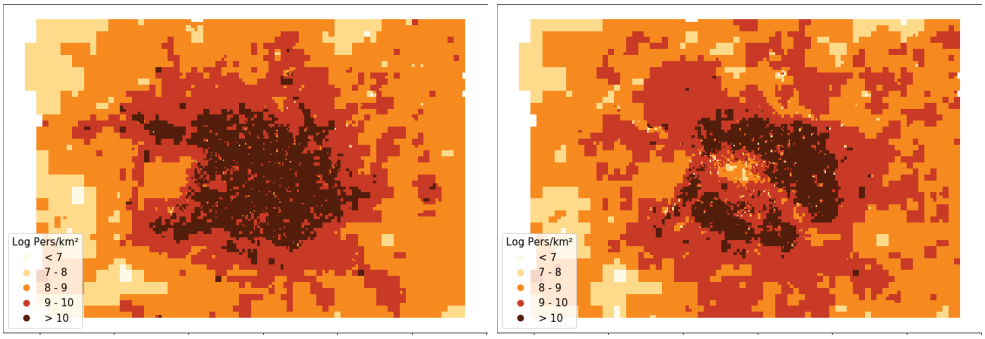
FIGURE 11 – **Population Densities in the Dissemination Grid.** Present Population : March 2019. ENACT : March 2011. Resident Population : 2016.

the *Seine* river and the *Défense* neighborhood attract population during the day, as predicted from activity-related presence with the ENACT methodology. [20]

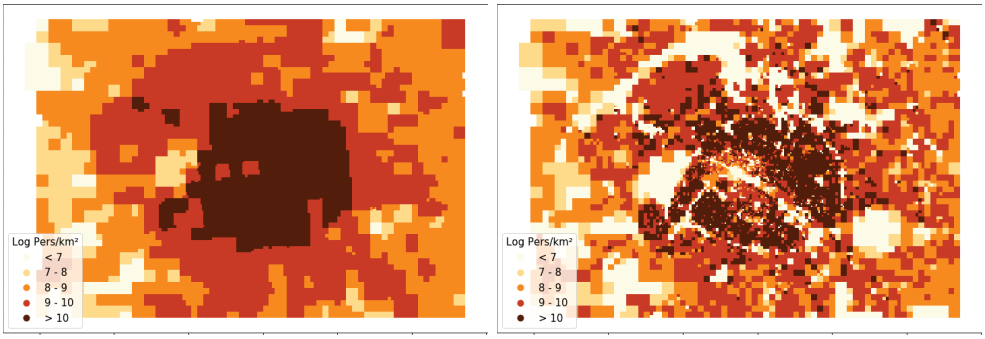
In addition, Table 3 reports comparisons along three metrics : correlation, rank correlation and allocation accuracy. Allocation accuracy can be interpreted as the percentage of population density allocated in the same tiles in both sources and is defined as :

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

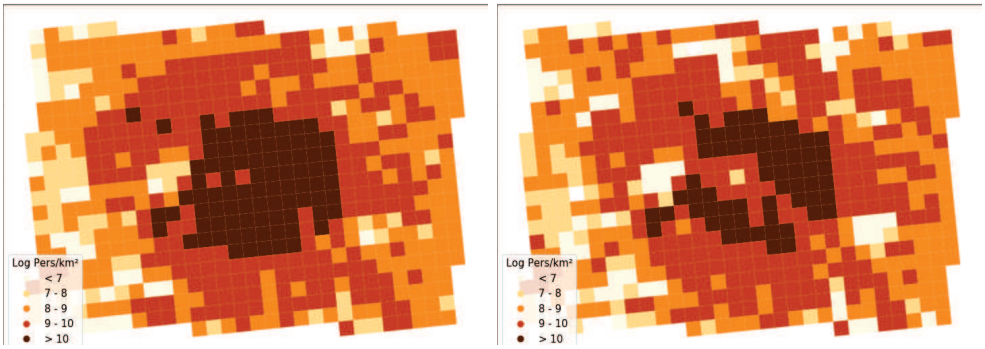
20. Figure 14 in appendix presents the differences between the present population statistics and both external sources over France. It makes clearer the tendency of mobile phone data estimates to create halos around dense areas. It shows that at night in dense areas such as Paris, except for particular places such as parks, the error ressembles 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.



(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 12 – **Population Densities in the Dissemination Grid (e-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 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

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 Discussion

This experimental present population estimate was built with knowledge and inspiration from a number of existing works notably [Salgado et al., 2021], [Statistics Netherlands, 2020] and [Ricciato et al., 2020]. However, we found that off-the-shelf solutions were never fully applicable to our case, and resorted to a number of simplifications which we discuss here. In particular, data fusion approach, merging high quality population datasets and mobile phone data could be considered and studied in the future to make the most of this promising data source in countries where these alternative exist.

[Salgado et al., 2021] proposes an ambitious bayesian general framework for producing statistics from mobile phone data, based on a Hidden Markov Model (HMM) modelisation for

device-level trajectories. In constrast, we do not rely in this work on any inference framework to quantify the uncertainties in our final estimates. Of course this is a downside, but the level of computational complexity entailed by resorting to this modelisation seemed prohibitive in our context. A static spatial mapping on cell-level aggregates, as opposed to the dynamical spatial mapping at the device level delivered by the HMM model, was considered to avoid computational burden. One significant advantage of our method is to be able to vary the spatial mapping after the computationally intensive aggregation step, as various spatial mapping have been show to provide quite different results. On a similar signaling dataset restricted to a large urban area, [Bonnetain et al., 2019] resort to a hidden markov chain modelisation for map-matching device trajectories on the transportation network but they simplify the cell spatial coverage information to a Voronoï tesselation. The computational complexity of resorting to simple temporal 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 in theory parallelizable at device level, the single device problem can be quickly high dimensional in space and time.

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 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. [21] 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 mostly for mobility analysis. The issues of data time sparsity and sensitivity to user behaviour are generic, and we show that they 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) due to the lack of access to longitudinal data for research. Throughout the lockdown period imposed in France during the Covid-19 crisis, some methods used by MNO showed 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 devices and therefore to unexpected large mechanical increases in present population estimates. Estimates based on reprocessed multi-MNO data were conducted by the French National Statistical Institute [Galiana et al., 2020b, Coudin et al., 2021]. In addition, interpolation has a positive effect on representativeness if the extent of network detection is correlated with socio-economic background : less active users (or users having access to a less performant mobile phone or local network) participate to aggregates more equally after interpolation relatively to more active users. Network usage has indeed be shown related to individual characteristics, such as gender [Jahani et al., 2017].

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 3 and 7). Imposing a constraint of equality at a local level between "usual" residents detected from mobile-phone data and actual residents as estimated from official sources appears milder than an alternative that would consist in equalizing resident population P_r to population present at night : $u_{r,t_0} = P_r$. A significant part of the population spends nights regularly outside of its main residency or works at night. These atypical location behaviours may be relatively

21. 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-grained mobilities).

more captured by MPD than by traditional sources. Figure 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, they could be based on as many characteristics as we can accurately recover at the device-level and for which we have an external population-level estimate. [22] This framework could be promising for future combinations of sources aiming at facing selectivity issues to derive representative statistics.

CBS (Statistics Netherlands) has recently published a report on its methodology [Statistics Netherlands, 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 heuristic. Their calibration step 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 as the ones we impose. One difference is that [Statistics Netherlands, 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 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 further work would be needed to integrate our minimal set of constraints and match our population totals.

We derived a prototype for an experimental present population statistics for France. Daily and weekly cycle, both at local and national level offer unprecedented insight on population dynamics over France. 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 at fine-grained level (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, if at all, in European countries.

22. See the discussion on pseudo-weights in [Beręsewicz et al., 2018]

Références

[Batista e Silva et al., 2020] 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 et al., 2018] 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 et al., 2019] 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.

[Chen et al., 2019] 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.

[Coudin et al., 2021] Coudin, E., Poulhes, M., and Castillo, M. S. (2021). The french official statistics strategy : Combining signaling data from various mobile network operators for documenting covid-19 crisis effects on population movements and economic outlook. *Data & Policy*, 3.

[Deville et al., 2014] 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.

[Fekih et al., 2021] Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A., and Galland, S. (2021). A data-driven approach for origin–destination matrix construction from cellular network signalling data : a case study of lyon region (france). *Transportation*, 48(4) :1671–1702.

[Galiana et al., 2020a] 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 et al., 2020b] 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).

[Hoteit et al., 2016] 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.

[Jahani et al., 2017] Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., Pentland, A. and de Montjoye, Y.-A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science*, 6(1) :3.

[Katsikouli et al., 2019] 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 and Daniel, 2018] Létroublon, C. and Daniel, C. (2018). Le travail en horaires atypiques : quels salariés pour quelle organisation du temps de travail ? *DARES ANALYSES*, 2018-30.

[Panczak et al., 2020] Panczak, R., Charles-Edwards, E., and Corcoran, J. (2020). Estimating temporary populations : a systematic review of the empirical literature. *Humanities and Social Sciences Communications*, 6(1) :1–10.

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

[Ricciato et al., 2020] 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.

[Ricciato et al., 2017] Ricciato, F., Widhalm, P., Pantisano, F., and Craglia, M. (2017). Beyond the "single-operator, cdr-only" paradigm : An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, 35 :65–82.

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

[Sakarovitch et al., 2018] 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 et al., 2021] Salgado, D., Sanguiao, L., Oancea, B., Barragán, S., and Necula, M. (2021). An end-to-end statistical process with mobile network data for official statistics. *EPJ Data Science*, 10(1) :1–46.

[Schiavina et al., 2020] Schiavina, M., Freire, S., Rosina, K., Ziemba, L., Marin Herrera, M., Craglia, M., Lavalle, C., Kemper, T., and Batista, F. (2020). Enact-pop r2020a-enact 2011 population grid. *European Commission, Joint Research Centre (JRC)*.

[Statistics Netherlands, 2020] Statistics Netherlands, C. (2020). Estimating hourly population flows in the netherlands. *Statistics Netherlands*.

[Tennekes et al., 2020] Tennekes, M., Gootzen, Y., and Shah, S. H. (2020). A bayesian approach to location estimation of mobile devices from mobile network operator data. In *CBDS Working Paper 06-20*.

[Vanhoof et al., 2018] 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.

6 Appendix

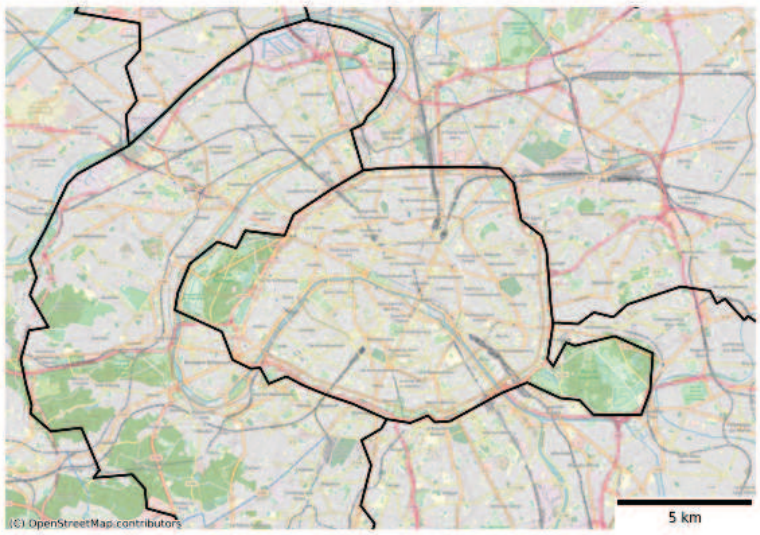
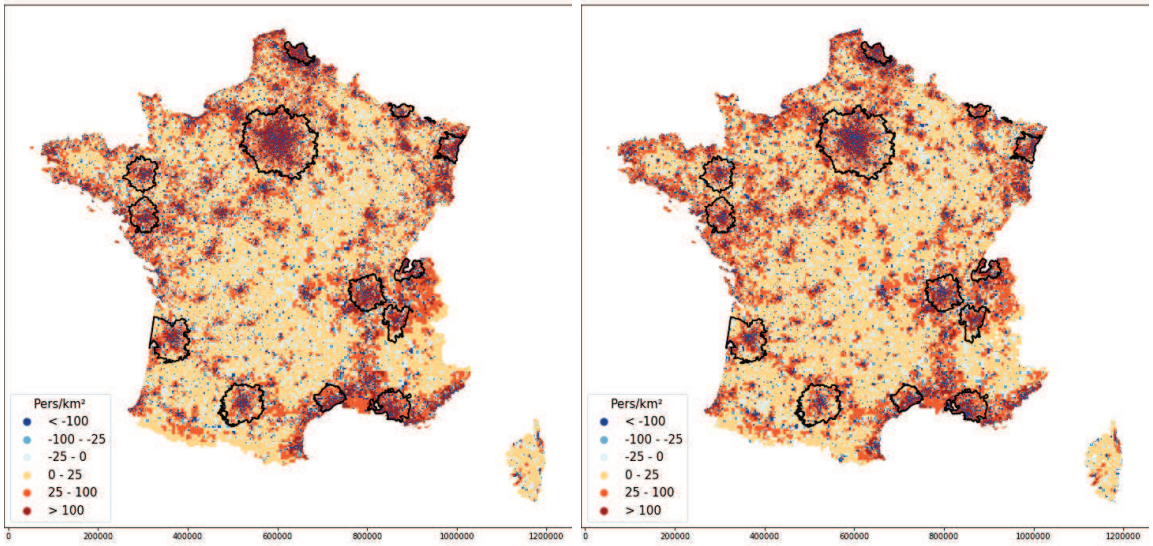
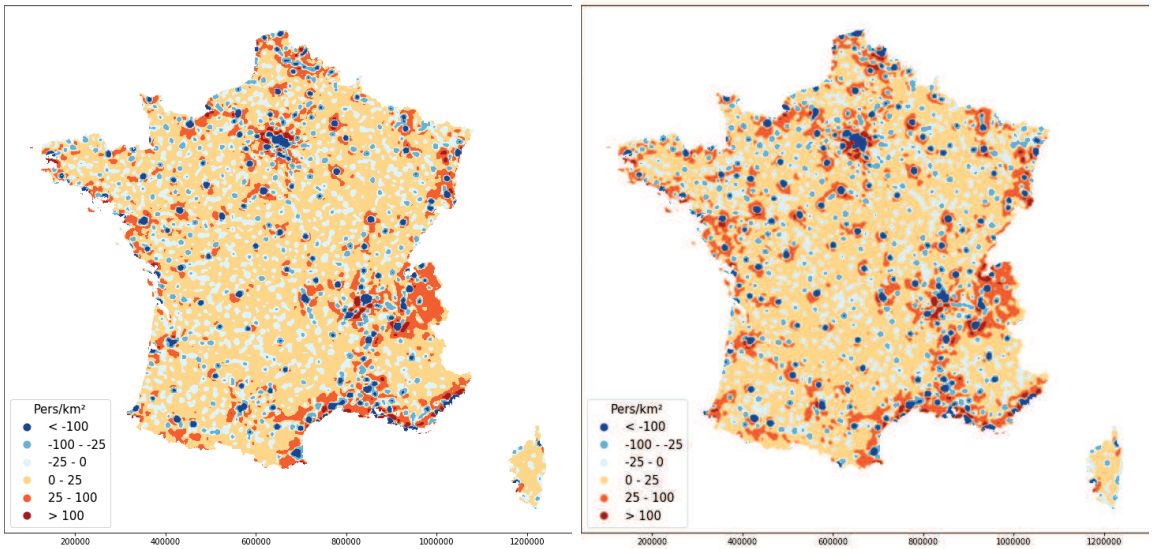


FIGURE 13 – Paris area context, with *Département* administrative boundaries

Night : Residents *Day : ENACT*



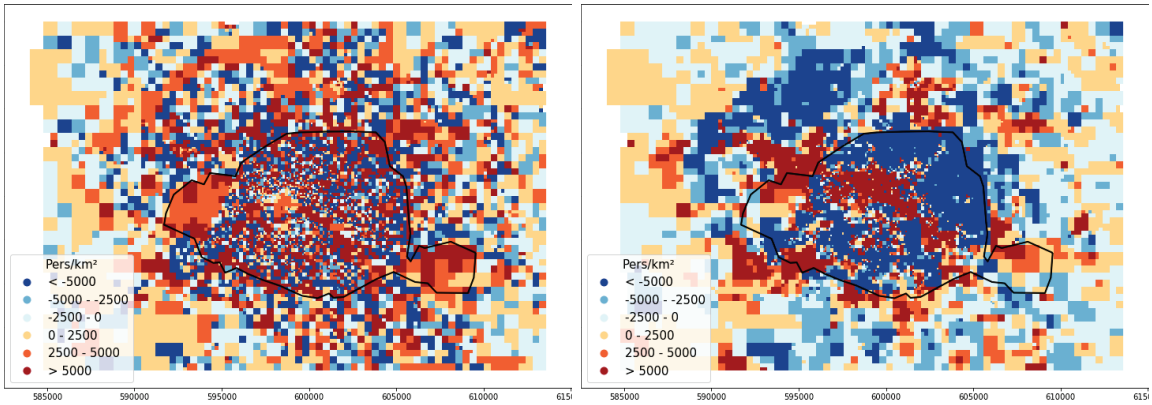
(a) Differences Present Population - External Source Population



(b) Smoothed Differences Present Population - External Source Population

FIGURE 14 – 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

Night : Residents *Day : ENACT*



(c) Differences Present Population - External Source Population

FIGURE 15 – Differences between Present Population at 3a.m. and Resident Population (Left) and between Present Population at 3p.m. and ENACT Day Time Population (Right).