Estimating urban mobility with mobile network geolocation data mining

Danya Bachir

▶ **To cite this version:**

Danya Bachir. Estimating urban mobility with mobile network geolocation data mining. Networking and Internet Architecture [cs.NI]. Université Paris Saclay (COmUE), 2019. English. NNT : 2019SACLL004 . tel-02046122

HAL Id: tel-02046122
https://theses.hal.science/tel-02046122v1

Submitted on 22 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

NNT : 2019SACLL004

Thèse de doctorat

Estimating Urban Mobility with Mobile Network Geolocation Data Mining

Thèse de doctorat de l'Université Paris-Saclay
préparée à IRT SystemX, Bouygues Telecom, Télécom SudParis

Ecole doctorale n°1 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Réseaux, information et communications

Thèse présentée et soutenue à Gif-sur-Yvette, le 25-01-2019, par

DANYA BACHIR

Composition du Jury :

Jakob PUCHINGER Professeur, LGI, Centrale-Supelec	Président
Latifa OUKHELLOU Directrice de Recherche, IFSTTAR, Université Paris-Est	Rapporteur
Andre-Luc BEYLOT Professeur, IRIT-INPT, ENSEEIHT	Rapporteur
Marco FIORE Professeur, CNR - IEIIT	Examinateur
Mounim El YACOUBI Professeur, Télécom SudParis	Directeur de thèse
Vincent GAUTHIER Maître de conférence, Télécom SudParis	Encadrant de thèse
Sylvain GOUSSOT Vice-Président Big Data Innovation, Bouygues Telecom	Invité
Nicolas GAUDE CTO - Data Scientist, prevision.io	Invité

Abstract

In the upcoming decades, traffic and travel times are expected to skyrocket, following tremendous population growth in urban territories. The increasing congestion on transport networks threatens cities efficiency at several levels such as citizens well-being, health, economy, tourism and pollution. Thus, local and national authorities are urged to promote urban planning innovation by adopting supportive policies leading to effective and radical measures. Prior to decision making processes, it is crucial to estimate, analyze and understand daily urban mobility. Traditionally, the information on population movements has been gathered through national and local reports such as census and surveys. Still, such materials are constrained by their important cost, inducing extremely low-update frequency and lack of temporal variability. On the meantime, information and communications technologies are providing an unprecedented quantity of up-to-date mobility data, across all categories of population. In particular, most individuals carry their mobile phone everywhere through their daily trips and activities.

In this thesis, we estimate urban mobility by mining mobile network data, which are collected in real-time by mobile phone providers at no extra-cost. Processing the raw data is non-trivial as one must deal with temporal sparsity, coarse spatial precision and complex spatial noise. The thesis addresses two problematics through a weakly supervised learning scheme (i.e., using few labeled data) combining several mobility data sources. First, we estimate population densities and number of visitors over time, at fine spatio-temporal resolutions. Second, we derive Origin-Destination matrices representing total travel flows over time, per transport modes. All estimates are exhaustively validated against external mobility data, with high correlations and small errors. Overall, the proposed models are robust to noise and sparse data yet the performance highly depends on the choice of the spatial resolution. In addition, reaching optimal model performance requires extra-calibration specific to the case study region and to the transportation mode. This step is necessary to account for the bias induced by the joined effect of heterogeneous urban density and user behavior. Our work is the first successful attempt to characterize total road and rail passenger flows over time, at the intra-region level. Although additional in-depth validation is required to strengthen this statement, our findings highlight the huge potential of mobile network data mining for urban planning applications.

Abstract (Français)

Dans les prochaines décennies, la circulation et les temps de trajets augmenteront drastiquement en raison du fort taux d'accroissement de la population urbaine. L'augmentation grandissante de la congestion sur les réseaux de transports menace le bon fonctionnement des villes à plusieurs niveaux, tels que le bien-être des citoyens, la santé, l'économie, le tourisme ou la pollution. Ainsi, il est urgent, pour les autorités locales et nationales, de promouvoir l'innovation pour la planification urbaine, à l'aide d'une politique de soutien à l'innovation et de prises de mesures radicales. Pour guider les processus de décisions, il est crucial d'estimer, analyser et comprendre la mobilité urbaine au quotidien. Traditionnellement, les informations sur les déplacements des populations étaient collectées via des rapports nationaux et locaux, tels que les recensements et les enquêtes. Toutefois, ces derniers ont un coût important, induisant une très faible fréquence de mise-à-jour, ainsi qu'une temporalité restreinte des données. En parallèle, les technologies de l'information et de la communication fournissent une quantité de données de mobilité sans précédent, au jour le jour, toutes catégories de population confondues. En particulier, les téléphones portables accompagnent désormais la majorité des citoyens lors de leurs déplacements et activités du quotidien.

Dans cette thèse, nous estimons la mobilité urbaine par l'exploration des données du réseau mobile, qui sont collectées en temps réel, sans coût additionnel, par les opérateurs télécom. Le traitement des données brutes est non-trivial en raison de leur nature sporadique et de la faible précision spatiale couplée à un bruit complexe. La thèse adresse deux problématiques via un schéma d'apprentissage faiblement supervisé (i.e., utilisant très peu de données labélisées) combinant plusieurs sources de données de mobilité. Dans un premier temps, nous estimons les densités de population et le nombre de visiteurs au cours du temps, à une échelle spatio-temporelle relativement fine. Dans un second temps, nous construisons les matrices Origine-Destination qui représentent les flux totaux de déplacements au cours du temps, pour différents modes de transports. Ces estimations sont validées par une comparaison avec des données de mobilité externes, avec lesquelles de fortes corrélations et de faibles erreurs sont obtenues. Les modèles proposés sont robustes au bruit et à la faible fréquence des données, bien que la performance des modèles soit fortement dépendante de l'échelle spatiale. Pour atteindre une performance optimale, la calibration des modèles doit également prendre en compte la zone d'étude et le mode de transport. Cette étape est nécessaire pour réduire les biais générés par une densité urbaine hétérogène et les différents comportements utilisateur. Ces travaux sont les premiers à estimer les flux totaux de voyageurs routiers et ferrés dans le temps, à l'échelle intra-régionale. Bien qu'une validation plus approfondie des modèles soit requise pour les renforcer, nos résultats mettent

en évidence l'énorme potentiel de la science des données de réseaux mobiles appliquées à la planification urbaine.

Acknowledgement

First I would like to thank Nicolas Gaude and Sylvain Goussot for initiating this thesis project and ensuring good conditions for my research in both industrial and academic environments. I also thank Eric Vachon, Jeremy Banidol and Gabriel Jouffrai from Bouygues Telecom Big Data Labs for their help on data engineering, domain knowledge on cellular networks and privacy concerns.
Second I thank my thesis director, Professor Mounim El Yacoubi, and thesis supervisor, Dr Vincent Gauthier, for their guidance. Mounim helped me regarding data mining and machine learning methodologies while Vincent advised me on telecommunication networks aspects.
Third I am grateful for conducting my research at IRT SystemX where I discussed with many researchers and engineers specialized in different domains. I acknowledge the help and support from Dr. Mostepha Khouadjia and Dr. Fabien Tschirhart on transport networks data processing. I also thank Professor Jakob Puchinger for his collaboration on our journal paper and discussions on transport applications of our work. In addition, I would like to thank my fellow Ph.D students Reza Vosooghi, Dr. Ouail Al Maghraoui and Abood Mourad for their great support and insights regarding transport research. In addition of their respective talents, Reza, Ouail and Abood are among the most helpful and kind persons I have ever met.
Many thanks to Dr. Ghazaleh Khodabandelou, I am deeply grateful for her collaboration on research papers for conferences and journal. She provided me detailed good quality feedback on methodology, structure and writing. Not only has she been a great co-worker, fully dedicated to her work, but she is also the kindest person whom I feel blessed to have as a friend.
At last, words are not enough to express my gratitude toward my family, lover and friends. They have consistently and unconditionally supported me besides of having always inspired me to become the best version of myself.

Table des matières

Table des figures

Liste des tableaux

Acronyms

MP Mobile Phone

CDR Call Detail Records

BTS Base Transceiver Station

2G Second-generation wireless digital phone technology used for voice telephony

3G Third-generation wireless phone technology used for voice telephony and internet access

4G Fourth-generation wireless phone technology used for voice telephony and fast internet access

GPS Global Positioning System

EGT Enquête Global Transport

OSM Open Street Map

OD Origin-Destination

LAC/LAU Location Area Code/Location Area Update

RMSE Root Mean Squared Error

1 Introduction

1.1 Context

Current forecasts for global population growth predict an increase, from the actual 7 billion people, to 9 billion individuals by 2050 and up to 12 billion by 2100 (LERNER et VAN AUDENHOVE, 2012 ; GERLAND et al., 2014). In urban territories, the residential population is expecting a rise from 50% to 70%, while workers are trusted to triple in city centers. Without effective measures, congestion and travel times will dramatically increase. This threatens citizens everyday life, including well-being, health and productivity at work. Cities economy is also at risk as it is closely linked to cities attractiveness regarding tourists and workers. In the race for digital transformation through artificial intelligence, smarter cities will be a key factor to retain, attract and train smart people. Thus, the need for efficient urban planning policies aimed at improving mobility is becoming everyday more crucial for the future.

So far, information on population mobility have been collected from census and surveys. Housing census records provide general insights on residential population meanwhile travel surveys are conducted to estimate statistics on modal share across areas and travel flows for a typical day. However these traditional data sources lack temporal depth, hence are unpractical to capture daily variations of urban mobility. In particular, by reason of their important cost, surveys suffer from low-update frequency, being generally once per decade, and from limited size of surveyed individuals, which may introduce sampling bias.

In parallel, the pervasive usage and the high penetration rates of mobile phones have made mobile network data the largest mobility data source ever. Call Detail Records (CDR) are collected at no additional cost by telecommunications operators for billing purposes. Several research works have described the potential of such data for mobility analysis (CHEN, J. MA et al., 2016 ; GADZIŃSKI, 2018 ; BLONDEL et al., 2015). Popular research topics are travel demand modeling (TOOLE et al., 2015 ; M.-H. WANG et al., 2013), itinerary reconstruction (ASGARI et al., 2016 ; BECKER et al., 2011), traveler behavior understanding (CALABRESE, DIAO et al., 2013 ; Yihong WANG et al., 2018 ; AHAS et al., 2010), population density estimation (BACHIR, GAUTHIER et al., 2017 ; KHODABANDELOU et al., 2016a), transport mode detection (H. WANG et al., 2010 ; BACHIR, KHODABANDELOU, GAUTHIER, EL YACOUBI et VACHON, 2018), traffic state estimation (DEMISSIE et al., 2013 ; DONG et al., 2015), passenger flow estimation (ZHONG et al., 2017), anomaly de-

tection (PANG et al., 2013), Origin-Destination (OD) matrices construction (ÇOLAK et al., 2015a ; ALEXANDER et al., 2015 ; TOOLE et al., 2015), mobility and activity patterns extraction (JIANG, FERREIRA et al., 2017 ; CHEN, BIAN et al., 2014) to name a few. Therefore, mobile network data represent an inexpensive and up-to-date supplement to traditional data, able to provide real-time large-scale mobility insights.

1.2 Objectives and Challenges

The broad objective of this thesis is to mine mobile network data to estimate daily urban mobility information. In particular, the present work addresses the following questions :

— What knowledge can mobile network data mining bring compared to traditional data sources for mobility analysis ?

— Which data pre-processing guarantees a good model performance ?

— How does the choice of the spatio-temporal scales affect model performance ?

— How many individuals visited a location given the mobile phones activity ?

— How to identify the dominant type of human activity (e.g., residential, business etc.) involved in an area given the mobile phone activity ?

— Which transportation mode is taken by traveling cellphone holders ? How many modes can be identified ?

— How many trips, in total, occur between different urban areas, at a time ?

— What are the travel patterns ? What are current mobility trends ?

In order to solve these problematics, several challenges need to be accounted during model conception. Although mobile network data are massively collected by mobile providers at no additional cost, the raw data pre-processing is non-trivial and time consuming by reason of the following constraints. First, mobile phones communication frequency shows important discontinuity and irregularity in time which leads to sparse geolocation. Secondly, the exact GPS coordinates of cellphones are unknown to mobile providers, as GPS require a different type of technology (i.e., satellites). Instead, cellphones are geolocated using the mobile network equipments. Devices are positioned inside mobile network cells, which are the coverage area of the signal (i.e., 2G, 3G, 4G) received by a phone from an antenna positioned on a base transceiver station. However, mobile networks are coarse-grained as mobile network cells are wide from several hundreds meters to kilometers, depending on the urban density of the area. In other words, the mobile network geolocation precision of cellphones has an uncertainty of the order of 1 km, against a few meters (up to 300m in dense areas) for raw GPS

traces. A few studies have used triangulated mobile network geolocation, using data collected in the U.S. (ALEXANDER et al., 2015 ; H. WANG et al., 2010). Still, triangulation is currently not authorized for large-scale analysis in France, to protect users privacy, except for police investigations and emergency calls. Third, mobile network geolocation suffers from complex spatial noise induced by load balancing effects, when devices are at the boundaries of several nearby cells, with overlapping signals. Consequently, mobile network geolocation is sparse, coarse and noisy hence providing approximate and partial knowledge of true users' trajectories.
In addition, mobile network data are unlabeled regarding the mobility of cellphone users. For instance, it is not straightforward to associate a geolocated record with a human activity, a transportation mode, or to a specific point of interest by reason of the spatial imprecision. Labels annotation requires expert knowledge and is a costly task, hence is unpractical for large datasets. Therefore, new unsupervised, or at least semi-supervised, approaches are needed to tackle this issue. Still, unsupervised models require a validation step, equivalent to the testing phase in supervised approaches. During validation, the estimates are compared to the most relevant external data, used as ground truth. However, recent validation data are extremely difficult to acquire for large population. Meanwhile, the data legislation imposes a one year retention period for the mobile network data. Thus, the performance evaluation of unsupervised models for mobility analysis is another challenge to overcome in this thesis.

1.3 Contributions

This thesis provides several contributions to research on urban mobility analysis with mobile phone data, which are listed below.

— This is the first work combining five different types of real datasets for mobility, collected from multiple sources, over long periods. The datasets involve hundreds of millions mobile network trajectories over two months, multi-modal transport networks, census data, detailed travel survey information and one month travel card data. The case study is the Greater Paris region, which is a 12000 km^2 wide area with more than a thousand towns. In addition, the densities of Greater Paris population and transport networks are among the highest worldwide, with heterogeneous densities between Paris and its suburb. Thus our models are generalizable to both high density and low density areas.

— Records collected by mobile operators are mainly generated by active cellphones. Turned-off or inactive devices and non-subscribers are thus undetected. Hence, despite important market share, a substantial part of the total population is missing. We address this problematic by implementing a dynamic rescaling method using Call Detail Records. Population densities and absolute numbers

of visitors are derived, for each 15 minutes time-slots, at several spatial resolutions, from census blocks to zipcode areas. In particular, we provide explicit interpretation of model parameters and results.

— In addition, this thesis presents the first approach for the estimation of Origin-Destination flows per transport mode. The model outputs the total flows per hour, for rail and road modes, traveling between all the 1276 zipcode areas in the region. The model requires a minimum of two locations to determine the transport mode from any trajectory. A transport probability is calculated and updated with each record without the need of the exact complete itinerary. Such model is resilient to noise and sparsity. OD flows are upscaled to total population using expansion factors calculated from the census and the travel survey.

— Several mobility datasets are combined through a learning scheme which is weakly supervised as we rely on small subsets of labeled data at some point in our approach. For performance evaluation, we conduct extensive validation tests against external data. All estimates are validated with high correlations (above 0.95) and small errors (below 10%).

1.4 Thesis Structure

Following this introduction chapter, the thesis is organized as follows.

Chapter 2 reviews the state-of-the-art with related work on mobility analysis using mobile phone data. In turn, we describe the mobile network data which has been used in the literature, the traditional population estimation models and the most relevant studies on travel flow mining.

Chapter 3 presents our case study and the different datasets used for this thesis. We successively introduce the mobile network data, the traditional population data and the transport data, which are compared in terms of pros and cons for urban mobility analysis. The chapter ends with an overview of the methodology, including the unsupervised and semi-supervised learning schemes, used in the thesis.

Chapter 4 describes the dynamic rescaling model to estimate population densities and total numbers of visitors. Following the introduction of such a rescaling problematic, we provide detailed methodology and results, including validation, ending the chapter with a discussion.

Chapter 5 presents the model for the estimation of OD matrices per transport mode. The structure of this chapter is similar to chapter 4.

Chapter 6 concludes on the presented work. We outline the contributions, obtained from the resolution of each problematic, and the general findings. Eventually, the thesis stresses the limitations of the presented work and provides future work recommendations.

2 Literature Review

In this chapter, we review related works on mobile network data mining applied to urban mobility analysis. First, we describe the different mobile network data types used in the literature, namely Call Detail Records and passive records, followed by the traditional data pre-processing techniques, such as noise reduction, smoothing and segmentation. Then, we report the state of the art methods for traditional urban mobility research. Related studies on population densities estimation and travel flow mining are successively presented. In particular, we focus on Origin-Destination matrix construction, itinerary reconstruction, transport mode detection, passenger flow estimation and mobility pattern mining.

2.1 Mobile Network Data

Different types of mobile network data have been used in the literature. The mobile network data can be classified into two main categories, namely Call Detail Records (CDR) and passive network records. Although these two data types share a common structure, they possess different spatio-temporal properties.

2.1.1 Call Detail Records

Several studies have used Call Detail Records (CDR) (JÄRV et al., 2014 ; CALABRESE, DIAO et al., 2013 ; DONG et al., 2015). However the description of the content of CDR in the literature often lacks clarity. In particular the “call” denomination is ambiguous as one may only think of a conversation on the phone. Phone calls are mobile network communications between two phones such as voice communications, including unanswered calling attempts, and text messages. In past studies it is unclear whether data logs (i.e., 3G-4G connections) are included in CDR. Mobile operators collect CDR for customer billing, profiling, market segmentation and for quality of service (QoS) optimization, to name a few. Each record contains the anonymized ID (aimsi) of the caller, and sometimes of the callee, a timestamp with a duration, the ID of the telecommunication equipment (cell ID) connected to the device and the type of call i.e. incoming, outcoming, voice or text. The sampling rate of CDR can vary from a few minutes to several hours. An example of CDR is given in Table 2.1.

2.1.2 Passive Records

Passive network records, sometimes referred to as sightings, have emerged in the literature more recently (Calabrese, Colonna et al., 2011 ; Chen, J. Ma et al., 2016 ; F. Wang et Chen, 2018). Passive records and CDR share the same structure. However, compared to CDR, passive records have a higher volume by reason of their higher collection frequency. The term passive records is a generic denomination, yet there can be different types of passive records collection. In particular, the sampling rate can vary depending on operators needs, material resources and legislations. Passive records can be generated from any interaction between a device and the network. Several positions can be recorded during a call or when the phone is not being used. This is not the case for CDR which require an active phone usage for calls and sms. Passive records are more convenient than CDR for mobile phone tracking as more frequent. The highest information level for passive records are the interactions between one device and several nearby cells at a time, which are used for triangulation.

2.1.3 Geolocation Precision

The area covered by the signal range of an antenna, positioned on a Base Transceiver Station (BTS), is called a cell, with radii varying from hundred meters to several kilometers. When a device receives a signal from the mobile network, a connexion is established between the device and one particular cell (2G, 3G, or 4G). If one device is located within the signal range of several equipments, it generally connects to the nearest cell available. Mobile phone positions are commonly approximated at the cellular scale which is coarse. In past studies, cells are traditionally represented as voronoï areas centered on BTS. Still, mobile network geolocation can have a finer spatial resolution through triangulation (Calabrese, Colonna et al., 2011 ; Calabrese, Diao et al., 2013 ; F. Wang et Chen, 2018 ; Alexander et al., 2015 ; Jiang, Ferreira et al., 2017). Triangulation requires signal frequencies of at least three nearby antennas to estimate a refined position for a device. In addition of being resource expensive, triangulation usage is severely restricted in several countries to protect users' privacy. Consequently triangulation remains a limited practice worldwide. An example of triangulated passive records is given in Table 2.2.

Table 2.1: Example rows from CDR

User ID	Timestamp	Duration (s)	Cell ID	Type
9221959679262440000	2018-06-01 20 :49 :01	59	123456	Voice
9221959679262440000	2018-06-01 21 :34 :30	0	234567	Text

TABLE 2.2: Example rows from triangulated passive records

User ID	Timestamp	Longitude	Latitude
9221959679262440000	2018-06-01 20 :49 :01	2.351516	48.865550
9221959679262440000	2018-06-01 20 :50 :00	2.355850	48.865522
	...		
9221959679262440000	2018-06-01 21 :34 :30	2.353532	48.865804

Our study is conducted with mobile network data from France where massive triangulation is prohibited. Our data contains both CDR and a low level of passive records. Further details on our data are provided in Chapter 3.

2.1.4 Pre-processing

Raw mobile network geolocation is coarse, noisy and sparse. Thus, it has limited spatio-temporal precision. The first condition required to observe meaningful results from mobility mining is massive data collection followed by careful data pre-processing. The first pre-processing step is commonly noise reduction. The two main steps of this process are oscillations filtering and positions smoothing.

Noise Reduction with Oscillations Filtering

Raw mobile network geolocation is impacted by an oscillation phenomenon, also called ping-pong effect. Oscillations are caused by network load balancing. A mobile phone connection can be transfered several times between nearby antennas, e.g. a phone covered by two antennas A and B can have a connexion pattern 'ABAB'. This phenomenon occurs in order to balance and optimize the network traffic charge. As a result a non-moving phone might appear as switching position. Oscillations generate false movements (i.e., noise) and increase the error for spatio-temporal measurements such as number of trips, distances, travel times and speeds. A few works addressed the task of oscillations removal. According to Bayir et al (BAYIR et al., 2009), a minimum of 3 switches is required to identify an oscillation pair. In order to remove oscillations, the authors applied a majority voting over the locations. As an example, for two locations (e.g., two cells) noted 'A' and 'B', in case we have an oscillation 'ABA' then the majority location, which is 'A', is retained. However this method does not address more complicated patterns such as 'ABAB' or 'ABCABC', having 'C' a third location. Çolak et al (ÇOLAK et al., 2015b) generated groups of consecutive records according to a distance threshold. A group is constituted of spatially close positions. For each group, the medoid position i.e., one minimizing the distance to others, is selected as the most representative location. Wu et al

(WU et al., 2014) used four heuristics to remove oscillations using notions of stable periods and jumps at impossible speed. A stable (or stay) period is defined as a device remaining at the same cell location, such as the period time duration is above a certain threshold (10 minutes in their study). First, if two consecutive stable periods occur at the same cell, such as the time interval is reasonably small, then the records between these two periods are considered as oscillations and removed. As an example, a sequence 'AAABCAAA' becomes 'AAAAAA' in case of a short elapsed time (e.g. 2 minute) between the two stable periods at 'A'. Second, if one record occurs after a stable period within a short time interval such as the next cell is too far away from the stable cell, then it is considered as an oscillation and is removed. Thirdly, the authors characterized oscillation patterns 'ABA' or 'ABC', between base stations such as 'A' and 'B' are separated by a long distance while 'A' and 'C' are spatially close. If such an oscillation occurs within a very short time it is considered as an impossible long jump and position 'B' is discarded. For the fourth heuristic, authors identified cycles, defined by sequences where the first and last record have the same cell, with at least one record from a distinct cell (e.g. 'ABCA'). Cycles occurring within short times are replaced by their most representative cell. Such a cell is found by calculating the highest score as the ratio of the cell frequency in the sequence and the average distance to cells in the sequence. Authors findings reveal that the fourth heuristic detects the highest number of oscillations compared to the three other heuristics. From their experiment on a 1 Tb CDR dataset, authors were able to identify that 6% of records were oscillations. Still, the aforementioned heuristics rely on time and distance conditions for which the authors fail to justify how to chose thresholds. In order to remove noise from triangulated sightings positions, Wang et al (F. WANG et CHEN, 2018) applied a time-window-based filtering method. For each trajectory, they recursively identified cycles, in a moving time window T_w. Different time window values were tested. Authors plotted the ratio of oscillation records over total records for each device, in function of T_w. The optimal T_w was found when the tangent to the curve remained stable beyond this value (elbow method). For their case study the optimal value was $T_w = 5$ minutes. In addition, authors identified patterns involving two switches, e.g. of the type 'ABAB'. In case at least one of the inter event time is shorter than T_w, the sequence is considered as oscillations. For each location of the oscillation sequence, the total spent time was calculated over the entire month. Then the oscillation sequence was replaced by the location with maximum duration.

Noise Reduction with Position Smoothing

Although oscillations filtering contributes in removing erroneous positions, CDR trajectories are still distant from real users trajectories. This phenomenon is stronger

in lower density areas where the network is coarser. In order to improve the geolocation precision, raw mobile network positions can be refined into more accurate positions by mean of trajectory smoothing techniques. Popular algorithms are mean, median, particle or Kalman filters (HORN et al., 2014), traditionally employed on GPS positioning. To the best of our knowledge, only one study provides a smoothing method for mobile network geolocation, based on a weighted moving average (CSÁJI et al., 2013). Let $\{x(1), ..., x(n)\}$ be a sequence of positions for a device, given some records with timestamps $\{t(1), ..., t(n)\}$. The smoothed positions noted $\{y(1), ..., y(n)\}$ are obtained as follows :

$$y(i) = \sum_{j \in B_\delta(i)} w(j)x(j) \tag{2.1}$$

$$B_\delta(i) = \{j : |t(j) - t(i)| \leq \delta\} \tag{2.2}$$

$$w(j) = 1 - \frac{|t(j) - t(i)|}{\delta} \tag{2.3}$$

where i is the index of the smoothed position, $B_\delta(i)$ is the set of indexes j for records occurring within a time window δ before or after record i, and $w(j)$ are the associated weights. Records separated by smaller inter-event time have higher weights, i.e. are more influenced by each other. Still, this study lacks a method to determine the optimal δ parameter. A too large or too small δ might impact the quality of the smoothing. Instead, the authors arbitrarily chose $\delta = 30$ minutes.
Trajectory smoothing enables to improve spatial precision by readjusting each location point with regard to the previous and next positions within a well chosen time window. It also reduces noise by smoothing oscillations. The underlying assumption is that whenever a device oscillates between several nearby antennas, this reveals that the real device position is probably between the oscillating cells.

Trajectory Segmentation

The second pre-processing step is trajectory segmentation which consists in identifying the different trips from the sequence of records made by a cellphone. One segment is defined as a pair of consecutive positions. By default, segments are represented by linear path for visualization. Segmentation is a crucial task which impacts the final results in mobility mining as it determines the number of trips. Most works on segmentation concerned GPS locations or triangulated CDR. Wang et al (H. WANG et al., 2010) grouped triangulated positions by proximity using a distance threshold $\Delta S = 1$ km. For each group of points within this threshold, a centroid is calculated. These centroids are used as sequences of Origins and Destinations (OD) locations. However this approach does not distinguish stay points and pass-by points from actual origin and destination of trips. More recent works identified stay points and moving points to differentiate stops from trips. These approaches

represent trips as sequences of moving points separated by stationary periods, which are sequences of stay points. Wang et al (M.-H. WANG et al., 2013) used a condition on duration to determine if a sequence is an actual trip or a stay period. For a given area, if the elapsed time between the first and last records is smaller than some threshold, the device is considered as traveling. The threshold depends on the area and corresponds to the average travel time expected. Jiang et al (JIANG, G. A. FIORE et al., 2013) and Toole et al (TOOLE et al., 2015) identified stay points by grouping locations according to a distance threshold $\Delta r < \delta$ and a temporal threshold $\Delta t > \tau$, where $\tau = 10$ minutes. An agglomerative grid-based clustering is then applied on each group of stay points candidates to remove noise and find the final stay locations.

2.2 Population Estimation

Estimating population densities and crowd size is another important information for transport planning and more globally for urban planning. Past research aimed at estimating the total number of individuals present in an area using mobile network data. In what follows, we review state of the art techniques and literature main findings.

2.2.1 Static model

Past research unveiled a relation between population and cellphone usage. A power relationship was identified between population density ρ and mobile phone activity σ (DEVILLE et al., 2014 ; DOUGLASS et al., 2015) :

$$\rho = \alpha\sigma^{\beta} \tag{2.4}$$

$$log(\rho) = log(\sigma) * \beta + log(\alpha) \tag{2.5}$$

where α is the proportionality coefficient between ρ and σ. Parameter β illustrates the superlinear effect of density estimates regarding mobile phone data (DEVILLE et al., 2014). The variable ρ is traditionally collected from census and represents the number of residents living in each area. This aforementioned relation is static as census population is constant in time. Past studies used three different ways to calculate σ in eq. 2.4. The mobile phone *activity* was derived as the number of calls, Short Message Services (SMS) and internet logs per time slot, aggregated per cell. Krings et al (KRINGS et al., 2009) obtained a positive correlation between cellphone *activity* and population density. Douglas et al (DOUGLASS et al., 2015) reached a Pearson correlation coefficient of 0.68 at 10-11AM, for calls. The mobile phone *activity density* is the *activity*, e.g. number of calls, normalized by the cells area.

Traditionally the area is calculated for the voronoï cell of the base station. Devile et al (DEVILLE et al., 2014) reached 0.8 correlations using night time *activity density*. The mobile phone *presence density* corresponds to the number of distinct phones per cell, normalized by the cell area. Khodabandelou et al (KHODABANDELOU et al., 2016b) obtained correlations between population density and *presence density* in the range $0.80 - 0.87$. The highest correlation was observed between 4-5AM. Model calibration, which aims to find optimal parameters (α, β), is proper to each case study region. Several studies concerning different regions have reported β to be slightly below 1 with little variation in time.

2.2.2 Population Mapping

Cellular and census data are gathered at different spatial scales. On the one hand, census uses geographic areas sometimes called census tracks, blocks or sectors. On the other hand mobile phone operators collect raw data at the base station, or cell, level. Therefore, prior to apply Eq. 2.4, one must conciliate the spatial scale for ρ and σ. Previous works, such as the one of Mennis (MENNIS, 2003), employed dasymetric models to map census densities at different resolutions using areal weighting. Deville et al (DEVILLE et al., 2014) performed spatial mapping using areal weighting. They calculated mobile phone counts σ_{c_j} per voronoï cell, noted c_j. Counts were mapped on census blocks, noted b_i, that intersect c_j. The resulting mobile phone counts per block σ_{b_i} is :

$$\sigma_{b_i} = \frac{1}{A_{b_i}} \sum_j \sigma_{c_j} A_{b_i \cap c_j} \tag{2.6}$$

where A_{b_i} is the area of block b_i and $A_{b_i \cap c_j}$ is the intersection area of b_i and voronoï cell c_j. Eq. 5.14 enables to distribute raw mobile phone counts in order to match census scale.

2.2.3 Clustering Urban Areas per Activity Type

The static model can be trained over a complete region (DEVILLE et al., 2014) and integrate land-use information (DOUGLASS et al., 2015). In addition, the training can be performed over specific groups of urban areas per activity type such as residential, business, leisure etc. (KHODABANDELOU et al., 2016b ; F. XU et al., 2016). One approach consists in training the static model exclusively on residential areas (KHODABANDELOU et al., 2016a ; KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018). The identification of areas activity type was performed by clustering the urban areas according to the mobile phone activity (FURNO, STANICA et al., 2015 ; FURNO, M. FIORE et al., 2017). The features correspond to the Median Week

Signature (MWS) derived from phone calls and texts. Each feature is derived from the median mobile phone activity, for a given time slot of a given day of the week (i.e. Monday to Sunday). Each signature feature is noted s_a and is calculated as follows :

$$s_a(mon, t) = \nu_{\frac{1}{2}}(v_a(d, t)|d \in d^{mon}) \tag{2.7}$$

Here, mon refers to the day of week (i.e., monday), t is the time slot, $\nu_{\frac{1}{2}}$ is the median value of the mobile phone activity $v_a(d, t)$ for the set of days d and the same time slots t.

2.2.4 Dynamic model

The population in an area is a time variant variable which is constituted of both residents that are staying at their home, workers and other recreational visitors. As census population corresponds to the total number of residents, it is therefore incorrect to use it as ground truth for a dynamic model calibration. Yet researchers and practitioners found there exists no substantial ground truth to solve this issue. To the best of our knowledge, only two approach have been proposed for dynamic populations. The first dynamic rescaling method consists in adding a rescaling term being the ratio of the total census population P over the total static population $\widehat{P}$. This method has been applied with Call Detail Records (DEVILLE et al., 2014) and passive records (F. XU et al., 2016). The dynamic model equation becomes :

$$\widehat{\rho_i} = R \cdot \alpha \sigma_i^{\beta} \tag{2.8}$$

$$R = \frac{P}{\widehat{P}} \tag{2.9}$$

The second method modifies the static parameters into generic dynamic parameters, noted $\widehat{\alpha}$ and $\widehat{\beta}$ (KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018) . An additional parameter λ (see Eq. 2.10) was introduced as the average activity per detected device over space (i.e., all census blocks).

$$\lambda_i^{call}(t) = \frac{1}{N} \sum_{i=0}^{N} \frac{\nu_i^{callin}(t) + \nu_i^{callout}(t)}{\sigma_i(t)} \tag{2.10}$$

where $\nu(t)_i$ is the number of records detected in cell i during time slot t. $\sigma_i(t)$ stands for the number of unique devices present in i during time slot t. The total number of blocks is noted N. During early morning hours, a linear relationship was observed between λ and $\widehat{\alpha}$ and between λ and $\widehat{\beta}$, for both calls and SMS. Eq. 2.4 was rewritten as :

$$\widehat{\rho_i} = (\widehat{a_\alpha}\lambda_i + \widehat{b_\alpha})\sigma_i^{(\widehat{a_\beta}\lambda_i + \widehat{b_\beta})} \tag{2.11}$$

2.3 Travel Flow Mining

2.3.1 Origin-Destination Matrices

A common representation for travel flows are Origin-Destination (OD) matrices. OD matrices are generally 2-dimensional arrays depicting the number of trips between OD pairs. Such matrices are widely employed in conventional transport planning models such as four steps and activity based models. So far, the transport community has employed traditional data sources to construct OD matrices yet several research works have shown that inexpensive and time variant matrices can be generated with mobile network data.

OD Matrices with Traditional Data

Traditional data sources for OD matrices are travel surveys and traffic link counts. The latter can be collected by mean of roadside monitoring using cameras, magnetic sensors etc. However gathering survey data is a long time process, which can take several years, hence leading to outdated information. In addition surveys are constrained by their important cost which restrains the number of surveyed individuals to only a small subset of the population (e.g., less than 1% residents). As a consequence surveys may suffer from sampling bias. Each individual answering the survey has to report every trip for only one day. Therefore, surveys fail to provide travel patterns for a period other than a typical day. Similarly, the cost of road sensors limits the number of roads equipped with sensors and restrains the size of traffic counts based matrices. Consequently, OD matrices derived from traditional data are unadapted for day-to-day large-scale mobility analysis.

OD Matrices with Mobile Phone Data

Mobile phone data have been used to derive time-variant and pervasive OD matrices over large populations. Common applications are the estimation of travel demand (M.-H. Wang et al., 2013; Toole et al., 2015), the evaluation and planning of traffic (Demissie et al., 2013; Dong et al., 2015), the identification of optimal locations for new transport routes (Berlingerio et al., 2013a), of trips purposes (Alexander et al., 2015) and of weekly travel patterns (Calabrese, Di Lorenzo et al., 2011) to name a few. Past studies on OD matrices construction with cellular data share a common methodology. The first step of the method is to determine mobile phones trips e.g., by mean of a segmentation algorithm. Then, one has to

identify the origin and destination of each trip. Still, in the literature, the determination of the origin and destination locations of trips is often unclear. Trips are grouped by same origin-destination locations and grouped by departure time to form the matrix. The aggregation of flows into matrices ensures privacy by preventing tracking of individual trips. The temporal and spatial scales are two important parameters to account for during the OD matrix construction. The last step is usually flow rescaling in order to convert the number of flows of detected mobile phones into flows of total population.

Several studies addressed OD matrix construction with mobile phone data for different case study regions. Caceres et al (CACERES et al., 2007) built an OD matrix in order to count vehicles on a highway. The conversion factor was calculated as the product of the global market share of the mobile provider and the probability of that phones inside vehicles are turned on. However, both CDR and vehicle traffic were simulated data for this study. Other studies used CDR from Boston and San Francisco, US (P. WANG et al., 2012) and Dahka, Bengladesh (IQBAL et al., 2014). For Dahka, a tower-to-tower (T2T) representation of OD matrix was used, a tower referring to a base station. The matrices were mapped to the road network to obtain a node-to-node OD matrix. When several roads are nearby a base station, the tower-to-node conversion rule attributes the road node of highest frequency over one month for a device. In our opinion, such a user specific method is, in practice, computationally expensive. In addition, because of mobile networks coarse granularity, the tower-to-node conversion becomes unpractical for regions having high density of transport networks. The final matrices were up-scaled using an optimization based algorithm minimizing the difference with observed traffic counts. Additional studies built OD matrices with triangulated sightings from Boston, US (ÇOLAK et al., 2015b ; ALEXANDER et al., 2015) and CDR from Singapore (JIANG, FERREIRA et al., 2017). The expansion factors were calculated as the ratio between the census population and the number of mobile phone subscribers living in this zone. Users' home locations were identified as the area of longest stay duration during night time.

2.3.2 Itinerary Reconstruction

Previous efforts on itinerary reconstruction mainly concerned trajectory mapping, also labeled as map-matching, using GPS traces. The objective is to find real users' itineraries on transport networks from sensors geolocation. Popular approaches used Kalman filters (HU et al., 2009), Expectation Maximization (HUNTER et al., 2011) and Hidden Markov Models (HMM) which are depicted as robust to noise and sparsity. HMM based approaches use the Viterbi decoding algorithm to find the most probable sequence of transport network nodes corresponding to the real path. Goh et al (GOH et al., 2012) used HMM on GPS traffic data. The optimal

path is the one with maximum likelihood over the markov chain. An SVM was used to learn the transition probability function. Thiagarajan et al (THIAGARAJAN et al., 2011) installed an android platform on devices to collect their geolocation with a one second frequency. An HMM was used to divide space into grid cells (hidden states) and determine the most likely sequence of pass-by grid cells. A second HMM was used in order to match the visited grid cell sequence to road segments. Asgari et al (ASGARI et al., 2016) applied an unsupervised HMM based approach to map raw mobile network trajectories on multi-modal transport networks. However, for CDR, map-matching highly depends on the number of records and performance might decrease for trajectories with low sample frequency. In addition, the algorithm suffers from important running times due to the density of transport networks in urban areas. For unsupervised approaches, validation is difficult as it requires a substantial amount of cellphone trajectories for which the real path is known. Another proposed strategy for itinerary reconstruction consists in generating O-D matrices and apply traffic assignment algorithms (TOOLE et al., 2015).

2.3.3 Transport Mode Detection

Few research has been conducted on transport mode detection with CDR. Previous methods have employed map-matching to infer the mode from the route (J. YUAN et al., 2010 ; ASGARI et al., 2016) and applied supervised learning algorithms (P. GONZALEZ et al., 2008 ; REDDY et al., 2010), which are both popular with GPS data. Contrary to GPS positioning, CDR geolocation is coarse, more noisy and sparse. Two consecutive records at two distinct antennas can be separated by long distances (from hundred meters to kilometers) or long time periods (from seconds to hours). Consequently, mobile network trajectories are an imprecise and incomplete representation of real users' paths. In addition of being computationally expensive, map-matching requires a substantive number of positions to find users' routes, therefore it is hardly generalizable to all CDR trajectories. Moreover, supervised models require training datasets with transport labels. Transport modes are either annotated manually, which is a costly task, or collected from mobile applications where users provide their travel information. Supervised models are thus constrained by the small number of labeled samples. Mobile network data is the biggest data source available, yet this data is unlabeled and requires unsupervised approaches. Biljecki et al (BILJECKI et al., 2013) calculated a transport score between consecutive GPS traces using boolean conditions on speed, distances to transport network and previous mode. Still, this work lacked a performance evaluation. Larijani et al (LARIJANI et al., 2015) and Aguiléra et al (AGUILÉRA et al., 2014) used base stations located inside Paris underground to identify underground flows from CDR. Additional modes were not addressed in these works. Wang et al (H. WANG et al., 2010) identified two transport modes, road and public transport, from CDR. Authors

estimated flows with triangulated positions and applied k-means clustering on travel times, followed by a comparison with Google travel times. Still, CDR low frequency induces important incertainty and delay on start and end travel times of CDR trips. Consequently a device may not be detected as traveling when the real trip begins and ends. Moreover the presented approach was applied on one unique Origin and Destination (OD) which is not sufficient to validate the method. In dense urban areas, travel times can be affected by traffic states (e.g., rush hours), transport incidents (e.g., delayed train), and can be identical for several modes, depending on the OD. Thus, additional features other than travel time should be investigated.

2.3.4 Passenger Flow Estimation

Few works addressed passenger flow estimation with mobile phone data. Aguiléra et al (AGUILÉRA et al., 2014) measured passenger flows inside Paris underground (RER A line) in order to evaluate the Quality of Service (QoS) of public transport. Passive records containing Location Area Updates were collected from base stations located inside the underground. In order to find the total flows, the raw flows were multiplied by rescaling coefficients. Such rescaling coefficients are the product of inverse market share and inverse probability of detecting mobile phones. This approach is similar to our rescaling method to estimate population size (BACHIR, GAUTHIER et al., 2017). The probability to detect a phone in a station was computed as $f = \frac{n_{s^+ \cap s \cap s^-}}{n_{s^+ \cap s^-}}$ where $n_{s^+ \cap s \cap s^-}$ is the number of phones detected both at the current station s, previous station s^- and next station s^+, and $n_{s^+ \cap s^-}$ is the number of phones detected both at previous and next station. The authors estimated the hourly O-D flows for mobile phones that were detected by underground base stations. In addition O-D flows were derived from individual travel card data for RER A stations. The two O-D matrices were compared on the travel times, number of flows and train occupancy for validation. However, in practice it is hardly possible to associate the mobile phones detected by underground base stations to one specific line when several lines transit between the same stations. For instance, downtown Paris, the RER A and underground line M1 share five stations in common. Moreover, travel card data provide the number of travelers entering the station but do not indicate which line is taken, neither the number of travelers inside passing by trains. Thus the differences existing between the two datasets make validation harder. In a more recent study, Zhong et al (ZHONG et al., 2017) characterized passenger flows for a transport hub in Shanghai, China. Passive records were collected from the base stations inside, or close to, the hub area. Flows were upscaled using the (inverse) market share. Passengers entering the hub were distinguished from passengers exiting. Although this work shows the potential of mobile phone data to estimate real-time travel patterns in transport hubs, the presented approach lacks a validation step (e.g., a comparison with travel card data).

Previous efforts on passenger flows estimation only concerned train mode for which the base stations are mostly indoor (i.e., inside the underground or the station). There exist many transport facilities that are overground (trains, tramway, bus, etc.) and do not benefit from indoor coverage. Thus, no previous works have solved the problem of large scale multi-mode identification with CDR.

2.3.5 Mobility Pattern Mining

Activity patterns

Past research represented mobility patterns by sequences of visited places, where individuals are involved in different types of activities. The three common activity types are 'home', 'work', and 'other' (SCHNEIDER et al., 2013 ; ALEXANDER et al., 2015). Activity patterns are helpful to analyze the behavior of distinct groups of population e.g., workers, tourists, students etc. Csáji et al (CSÁJI et al., 2013) identified the home-work locations of 100000 users from Portugal. A clustering was applied on users most frequent locations. The features of the locations are the hourly call activity, aggregated per week day. When at least three clusters were considered, the average call pattern of clusters matched the national statistics for time spent at home and work. Shneider et al (SCHNEIDER et al., 2013) observed the mobility networks for 40000 users from Paris. They found 17 unique motifs were sufficient to represent 90% of trips. This finding reveals the periodicity of human behavior. Alexander et al (ALEXANDER et al., 2015) identified the home location as the most visited place on week-ends and week-day night, in Boston. The work place is the stay location where users have the highest total travel distance d from their home : $max(d \cdot n)$, where n is the number of visits during day-time of the week. Widhalm et al (WIDHALM et al., 2015) identified activities by clustering the different stay locations, in Boston and Vienna. Features were derived from call activity and land-use information. Several activity types i.e., 'home', 'work', 'education', 'recreation', 'shopping' were inferred from the resulting clusters. The activity patterns were obtained using a relational markov network. The approach was validated against survey data. Jiang et al (JIANG, FERREIRA et al., 2017) identified the home location as the most visited place during night time from passive records from Singapore. Users mobility networks were identified and compared to classic activity-based model based on survey.

Trajectory patterns

Several studies addressed mobility pattern mining from users trajectories, in order to find similar travel behavior. Pattern mining has been used for outliers detection in trajectories (LEE et al., 2008), identification of traffic anomalies (PANG et al., 2013), to find travel companions (TANG et al., 2012) or users with same travel route (YE et al., 2009 ; ZHENG et XIE, 2011 ; Yilun WANG et al., 2014). Several approaches on trajectory clustering have been developed in the literature. Early studies found regression mixtures models with Expectation-Maximization (EM) algorithm outperformed K-means and standard gaussian mixtures (GAFFNEY et SMYTH, 1999 ; CADEZ et al., 2000). Such models are generative clustering able to handle different data types, with non-vector form and distinct lengths (e.g., time series, trajectories, videos etc.). Later, density based clustering algorithm such as DBSCAN became popular for trajectories by reason of their robustness to noise and outliers (ANDRIENKO et al., 2009 ; F. LIU et ZHANG, 2017). In trajectory clustering, the choice of a good similarity metric is of crucial importance. GPS trajectories are traditionally represented in an euclidean space with geolocation points of the form (x,y,t). GPS-based distances (e.g., Dynamic Time Warping, Frechet distance etc.) are not adapted for CDR trajectories which are noisy and sparse. Several works have represented CDR trajectories as strings representing the sequences of cells id. For instance a trajectory 'ABC' corresponds to successive records at cells 'A', 'B' and 'C'. In order to find clusters in CDR trips, Kang et al (KANG et al., 2009) modified the Longest Common Subsequence (LCSS) distance by the Common Visit Time Interval (CVTI) MAX LCSS distance. For each cell of the LCSS of two trajectories, the algorithm calculates the distance as the cumulative time spent at the locations of the LCSS. Yuan et al (Y. YUAN et RAUBAL, 2014) modified the Edit Distance, also named Damereau-Levenshtein, into a spatio-temporal cost function. The classic Edit Distance computes the number of operations to transform one string into another one. The possible operations are the removal, insertion or substitution of a letter. Let $T_1 = \text{`}UVW'$, $T_2 = \text{`}XYZ'$ and $T_3 = \text{`}XYYYYYYZ'$ be three CDR trajectories. For T_3, the device performed several records at Y. However, the distances are $ED(T1, T2) = 3$ and $ED(T2, T3) = 4$ leading to $ED(T1, T2) < ED(T2, T3)$ which is intuitively erroneous. This example illustrates the fact that the classic Edit Distance is not well suited to raw CDR because of duplicates. To solve the problem, Yuan et al (Y. YUAN et RAUBAL, 2014) removed successive duplicates within 30 min time slots. The centroid of each trajectory was calculated as the average record position. The new distance between trajectories was calculated as the sum of centroid displacement after each operation.

2.4 Summary

In this chapter, we reviewed the literature on mobile phone data mining applied to human mobility. Our work involves mobile network data, thus we first presented the different existing data types used in the literature (i.e., CDR and passive records) and common pre-processing algorithms for noise reduction and segmentation.
Second, we have reviewed state-of-the-art on population densities estimation. Such methods addressed the problem of rescaling active mobile phone users to total population, depending on time of day and area.
In addition, a comparison between traditional OD matrices and mobile phone based matrices is provided, followed by popular techniques for matrices construction. After pre-processing the raw cellular data, the common steps are the determination of origin and destination locations, spatio-temporal aggregation, and flows rescaling.
Past efforts on itinerary construction and transport mode detection mainly concerned GPS traces. The former techniques widely relied on map-matching while the latter were supervised learning algorithms. Both are unpractical for broad mobile network data which have lower precision, lower frequency, are noisy and unlabeled. Previous attempts on passenger flows estimation were restricted to underground networks for which base stations are indoor.
At last, we presented past research on mobility pattern mining, used to study travel behavior. Mobility patterns can be categorized into activity patterns, for which one groups users trips with same sequence of activities (e.g., home, work etc.), and trajectory patterns, for which one groups trips based on their spatio-temporal information.

3 Datasets & Methodology

3.1 Case Study Region

The models presented in this thesis focus on the case study of the Greater Paris region, which we briefly describe. This region spans over a 12000 km^2 area including more than 1200 cities (i.e., "communes") and 12 millions inhabitants. The Greater Paris is subdivided into administrative areas at different scales. The three coarser areas are the city center i.e. Paris, the first and the second ring in the suburb. The region contains 8 departments. Paris corresponds to one department, the first ring is constituted of departments 92, 93 and 94 while the second ring groups departments 77, 78, 91 and 95. In addition, the Greater Paris has 1276 communes, which are the smallest administrative territorial division in France. The region benefits from dense transport networks, including several public transport facilities and a high density of roads. In total there are 5 overground lines (RER), 16 underground lines (metro), 9 tramway lines and 8 train lines (transilien). The road network spans over more than 1300 km, including 450 km of highspeed roads.

3.2 Mobile Network Data

3.2.1 Content & Legislation

The main data used for this thesis are mobile network data (see Table 3.1) representing billions of rows each day (Terabytes). In order to enable mobile phones to interact, the mobile network transmits signal to each phone. A record is generated from incoming or outgoing calls, short message services (sms) and data transmission (i.e. connections to 3G and 4G). Records are timestamped and geolocated. Their collection aims at producing customers billing and to optimize the mobile network's quality of service. Records are produced at the start and end of voice calls, and every time a message is sent or received. Data records are generated at the start and end of 3G and 4G connections (i.e., IMSI attach/detach) with the mobile network. If the device has changed LAC, a data record is generated from a Location Area Update (LAU). At last, periodic location updates are recorded each

TABLE 3.1: Example rows from mobile network records

AIMSI	Timestamp	LAC	CellID	Techno	Type
#.....	2018-09-01 12 :06 :29	57301	X1	4G	Data
#.....	2018-09-01 12 :36 :29	57301	X1	4G	Data
#.....	2018-09-01 12 :50 :47	23022	X2	3G	Start Voice
#.....	2018-09-01 12 :52 :35	23022	X3	3G	End Voice
#.....	2018-09-01 12 :59 :33	22018	X4	2G	SMS
#.....	2018-09-01 12 :59 :38	57301	X5	4G	Data
#.....	2018-09-01 13 :29 :38	57301	X5	4G	Data

30 minutes. The latter enables to optimize the speed of signal transmission from the network in case of a new communication.

Mobile network providers must ensure that the data is compliant with data legislations stated by the french Commission nationale de l'informatique et des libertés (CNIL) and the European Union (i.e., GDPR). First, the contents of conversations exchanged by calls or sms and information on internet navigation are never recorded by the mobile phone provider. Second, the data is rigorously anonymized. The international mobile subscriber identity (imsi) is encrypted into an anonymized-imsi (aimsi). Third, the geolocation data must be aggregated in order to group at least 10 individuals having a similar behavior. As a result, data aggregation strengthens anonymization. Fourth, the data retention period must not exceed one year. Although the aforementioned three first points have no impact on the model performance, the fourth condition increases the difficulty for model validation (see. Sec 3.6.5). Indeed, during validation we need to compare results derived from mobile phone data for the ongoing year to some external data, yet the latter are generally outdated. The data used for this thesis are compliant with the current legislation.

3.2.2 Geolocation Precision

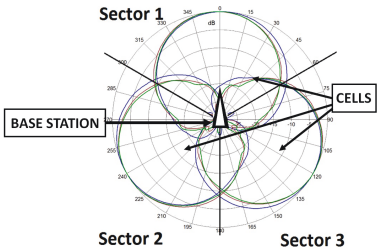

FIGURE 3.1: Tri-sector BTS with cells

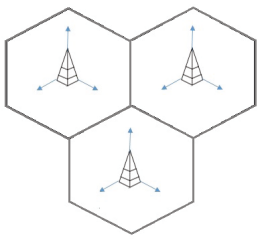

FIGURE 3.2: BTS polygons

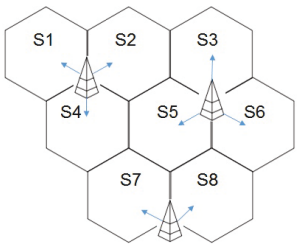

FIGURE 3.3: Sectors polygons

Mobile phone providers have no access to GPS coordinates of cellphones. Instead, each record is geolocated on the mobile network, using the position of the nearest telecommunication equipment connected to a device. In France, triangulation is

unauthorized except for authorities demands, as it is considered to bypass user consent and privacy. Raw mobile phone geolocation corresponds to mobile network cells which are signal areas (see Fig. 3.1). Although mobile phones are located near base stations, it is extremely rare to encounter devices positioned exactly at the base station. Mobile phones can be anywhere inside the cells. The latter are represented by circular shapes. Their radii range from a hundred meters in congested areas up to several kilometers in low density areas. Each base station is equipped with several antennas projecting several cells toward different directions. Mobile network signals span over a multitude of overlapping cells. For this work, we pre-process the raw cellular scale in order to merge overlaps. The mobile network is partitioned using the angle of the signal emitted by the different antennas attached to each base station. The resulting subdivisions are called mobile network sectors. The default configuration is tri-sectors base stations with sectors covering approximately 120°, as described in (RATTI et al., 2006). Each record is thus associated to its corresponding sector position. Traditionally, mobile network areas are represented as voronoï polygons centered on base stations. On Fig. 3.2, we display the mobile network areas using hexagonal polygons. This representation can be coarser than sectors as the average number of sectors per base station is three for our case study. Therefore, we use sector centroids (i.e., the barycenter of cells centroids from the same sector) to create voronoï polygons (see Fig. 3.3) in order to improve the spatial precision. Greater Paris sectors have a median area of 38 m^2, an average area of 386 m^2 and a standard deviation of 2570 m^2. Further down-scaling of mobile phone geolocations to smaller areas might introduce important location errors.

3.2.3 Sampling Frequency

Our records are enriched with Location Area Updates (LAU) to increase the sampling frequency. With classic CDR, several hours can pass between two consecutive records. With the LAU, the time-interval decreases from hours to a few minutes, having a 5 min median and a 30 min average for the Greater Paris region. The hourly box-plots for average number of records per cellphone are given in Fig 3.4. In average, cellphones generate between 1 and 10 records per hour, with a median close to 4 records per hour during the day. During the night, the median is maintained near 2 records per hour, resulting from the periodic LAU.
The literature have mentioned passively generated records which can be collected even when devices are not in use. Traditionally, passive records have been used for position tracking through triangulation. The highest level of detail of passive records can be generated from any interaction between a device and its nearest mobile network cells. In our case, the data connections resulting from location updates are passively generated i.e., a record is produced even if the device is not calling. Still, the passive records at hand are not generated frequently enough to detect all

turned-on devices anytime. In case a mobile phone does not call or text and does not change their Location Area Code (LAC) during a period shorter than 30 min, then no records are produced. Therefore, the data remains sparse in time and we must account for undetected mobile phones to study the mobility patterns of the total population.

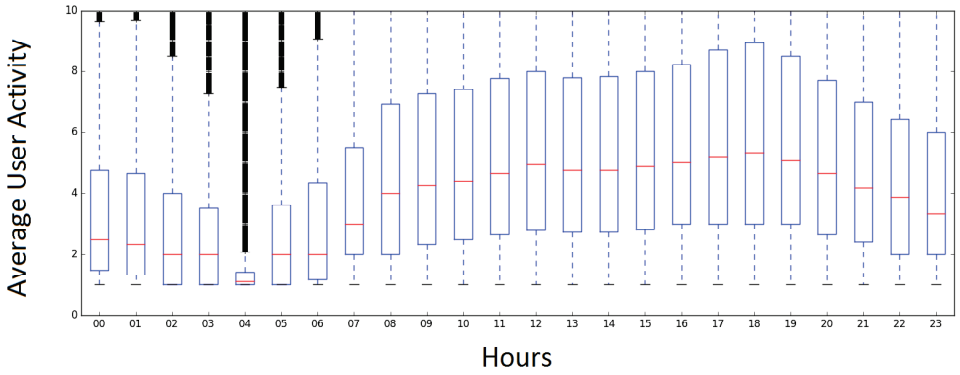

FIGURE 3.4: Average mobile phone activity per device and per hour for a typical day

3.2.4 Data Pre-processing

In this paragraph, we describe the general pre-processing steps applied by the mobile phone provider on anonymized raw data prior to the processing conducted for this thesis. First, a noise reduction strategy is applied following the approach in (CSÁJI et al., 2013). Trajectories are smoothed using a cumulative weighted moving average on a time window δ. The window is chosen according to the Nyquist criterion :

$$f \leq 2B \tag{3.1}$$

$$f_N = \frac{f}{2} \tag{3.2}$$

where f (Hz) is the frequency of the signal and B (Hz) is the size of the bandwidth. The Nyquist frequency f_N is defined as the minimum value for which a signal can be sampled without noise. In our case the data collection frequency f corresponds to the median of the inter-event time across cellphones, which is 8 min = 480 s. Consequently the smoothing window is set to $\delta = 240$ s. For each record, the raw position is smoothed according to the positions of previous and next records such as $\Delta t \leq \delta$. The smoothing step is followed by trajectory segmentation, based on two conditions on speed and time. Stay points are grouped according to a speed threshold $\Delta v < 10$ km/h and an elapsed time threshold $\Delta t > 15$ min. Thus, a device is considered as non moving if the elapsed time between the first and last stay points is at least 15 min, with a low speed. Records not fulfilling this condition are categorized as moving points. As noise reduction and trip segmentation are applied prior to the thesis contributions, these steps are not further detailed. After segmen-

tation, moving points are grouped together to form a trajectory corresponding to one trip.

3.2.5 Statistical Analysis of Trajectories

Prior to designing urban mobility models, we explore the properties of mobile network trajectories in order to analyze human mobility characteristics and check for possible noise and outliers. The following statistical analysis is based on the complete set of Greater Paris trajectories, for more than 2 million cellphones, during one month. In turn, we analyse trajectory duration, jump length, speed, gyration radius, inter event and number of positions. The boxplots of these features are given in Fig. 3.2.5 below.

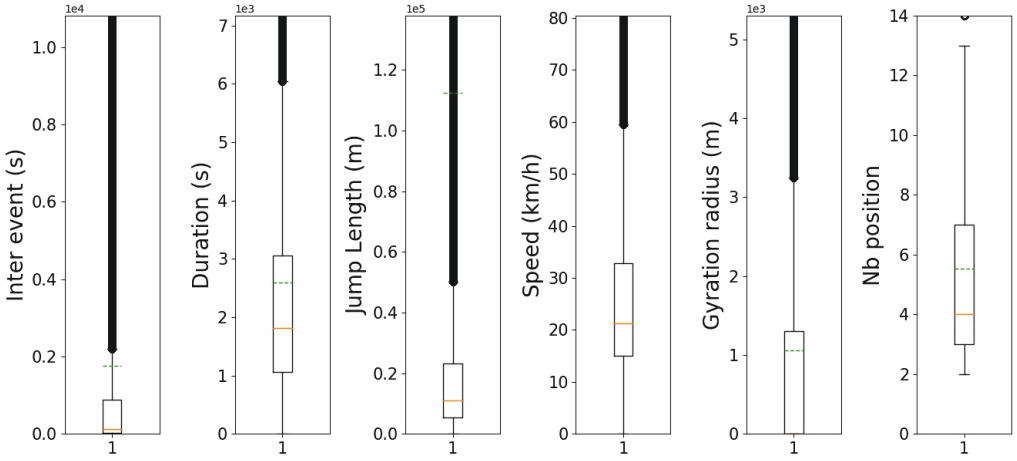

FIGURE 3.5: Boxplot of attributes for one month trajectories in the Greater Paris

Trajectory Duration

The trajectory duration is calculated as the time difference between last and first trajectory record : $\Delta t = t_{last} - t_{first}$. Here the duration is calculated for each trajectory (i.e., a sequence of moving points) and does not account for temporary stops (i.e., intermediate stay points). Hence, duration of trajectories can differ from the total travel time of individuals. Statistics on duration are given in Tab. 3.2.5. Crossing the complete Greater Paris region takes approximately up to three hours, with good traffic conditions. The upper bound of trajectory duration is approximately 5 hours.

TABLE 3.2: Statistics for trajectory duration

MIN	P99^{th}	AVG	MED	VAR	STD
2 s	4h45	45min	30min	$1.09\ 10^6$ s	55 min

Trajectory Distance

The trajectory distance, also called jump length, corresponds to the traveled distance during a trip. According to Brockmann et al (BROCKMANN et al., 2006), the jump length Δr follows a power-law distribution :

$$P(\Delta r) \sim (\Delta r)^{-(1+\beta)} \tag{3.3}$$

where $\beta < 2$. This finding reveals that people usually have short length trips and fewer long distance travels. More recently, the jump length was described as following a truncated power-law distribution (M. C. GONZALEZ et al., 2008).

$$P(\Delta r) \sim (\Delta r + \Delta r_0)^{-\beta} exp(\frac{\Delta r}{k}) \tag{3.4}$$

with $\beta = 1.75 \pm 0.15$, $\Delta r_0 = 1.5$ km and k is a cut-off value depending on the dataset. Jump length statistics are given in Tab. 3.3. In our study, the median and an average distance are respectively 10.2 km and 129.5 km. The minimum value is 4 m while the 99^{th} percentile is around 45 km. For the truncated power law distribution, our cut-off value is $k = 1000$ km.

TABLE 3.3: Statistics for Jump Length (Km)

MIN	P99^{th}	AVG	MED	VAR	STD
$4.16 \cdot 10^{-3}$	45.7	129.5	10.2	$6.82 \cdot 10^8$	825

The obtained jump length distribution in Fig. 3.2.5 confirms that the jump length follows a truncated power-law distribution. Parameters $\beta = 1.6$, $\Delta r_0 = 0$ and $k = 1000$ km are the best fit for our jump length distribution.

Trajectory Speed

The speed of a trajectory T is calculated considering the smoothed segments of trajectories. It is the ratio of the cumulative smoothed distance over cumulative duration.

$$\Delta v(T) = \frac{\sum_{i=0}^{n-1} \sqrt{(x_{i+1}^{smooth} - x_i^{smooth})^2 + (y_{i+1}^{smooth} - y_i^{smooth})^2}}{\sum_{i=0}^{n-1}(t_{i+1} - t_i)} \tag{3.5}$$

Speed statistics are given in Tab. 3.4. The minimum speeds are 10 km/h as this is the minimum threshold to consider a user in movement. The 99^{th} percentile for speed reaches 114 km/h. The last percentile has extreme speeds, up to 10^3 km/h and could hypothetically include trips by plane, high-speed trains or noise. High-speed trains and planes have speed approximating respectively 300 km/h and 900 km/h.

TABLE 3.4: Statistics for travel speed (km/h)

MIN	P99^{th}	AVG	MED	VAR	STD
10.0	$2.26 \cdot 10^3$	98.1	21.2	$3.99 \cdot 10^5$	631.4

Radius of Gyration

The radius of gyration represents the deviation of users positions from their centroid position (i.e., the center of mass). Yan et al (YAN et al., 2010) found that the radius of gyration is a constant depending on the time, first increasing quickly then slowly before convergence. For a device observed at n positions noted $r_i^{\vec{(t)}}$, $i \in [1, n]$, the radius of gyration is given by :

$$\begin{aligned} r_g\vec{(t)} &= \sqrt{\frac{1}{n(t)} \sum_{i=1}^{n(t)} (r_i^{\vec{(t)}} - r_{cm}^{\vec{(t)}})^2} \\ \text{where } r_{cm}^{\vec{(t)}} &= \frac{1}{n(t)} \sum_{i=1}^{n(t)} r_i^{\vec{(t)}} \text{is the centroid} \end{aligned} \tag{3.6}$$

The corresponding distribution can be approximated by a truncated power-law :

$$P(r_g) = (r_g + r_g^0)^{-\beta_r} exp(\frac{-r_g}{k}) \tag{3.7}$$

where $\beta_r = 1.65 \pm 0.15$, $r_g^0 = 5.8$ km and $k = 350$ km are the parameters found in (M. C. GONZALEZ et al., 2008). The value k is the cut-off value corresponding to the maximal gyration value possible. In our case the maximum radius is near 300

km which is close to the value found in (M. C. GONZALEZ et al., 2008). Gyration radius statistics are given in Tab. 3.5. The median and average gyration radius are respectively around 3 km and 9 km.

TABLE 3.5: Statistics for gyration radius (m)

MIN	MAX	P99^{th}	AVG	MED	VAR	STD
0	296801	14575	9387	2946	$1.19\ 10^7$	446

Number of positions per trip

Eventually we calculate the number of distinct positions per trajectory. The corresponding statistics are given in Tab. 3.6. Trajectories have at least 2 positions and at most hundreds of location points, with a median and an average around 4 and 5 locations.

TABLE 3.6: Statistics for number of recorded positions per trip

MIN	MAX	P99^{th}	AVG	MED	VAR	STD
2.0	232	22.0	5.5	4.0	19.8	4.45

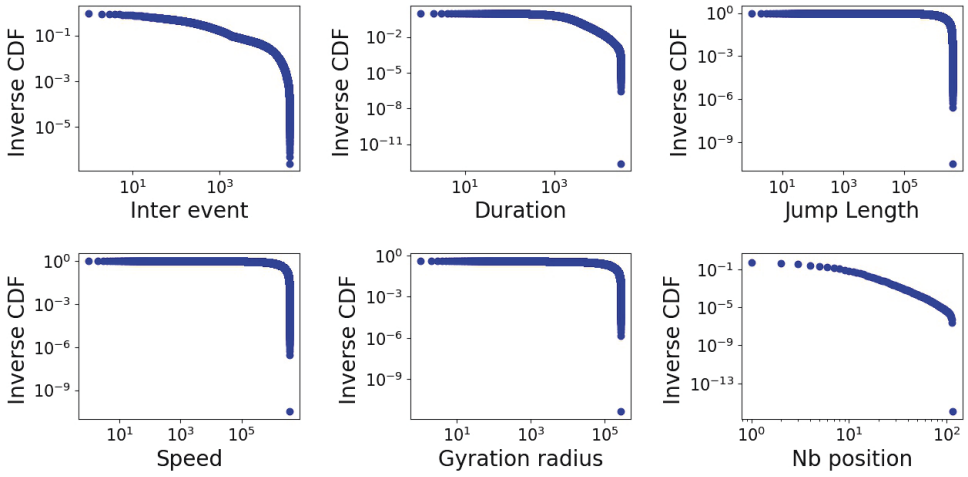

FIGURE 3.6: Inverse Cumulative Distributions of Trajectory attributes

3.3 Population Data

3.3.1 Housing Census

A housing census is conducted by state institutes in order to collect socio economic information from population (i.e., age, sex, work, status etc.). Such census counts

the total number of individuals in populations and the residential population in each area. In France, the housing census is conducted by the INSEE (Institut National de la Statistique et des Etudes Economiques). For the census, the territory is partitioned into several areas with varying spatial scales. Such areas are generally referred to as census sectors or blocks. For the french census, there exists two common scales. The coarser scale is the Commune level. Communes are identified with a unique INSEE code of 5 digits. The later is different from zipcodes, yet there exist a correspondence between the two codes. In France, the smallest census blocks are called IRIS, which is an acronym for 'aggregated units for statistical information'. An IRIS corresponds to a 2000 residents area. There are about 50000 IRIS in France. The Greater Paris contains 5261 IRIS which have irregular shapes, ranging from 8620 m^2 to 167.9 km^2, with a mean and median area of respectively 2.3 km^2 and 0.3 km^2. IRIS codes are constituted of 9 digits, the first 5 digits being the INSEE code, and the 4 last digits referring to the IRIS code.

3.3.2 Travel Survey

Transport authorities conduct travel surveys to collect information on individual mobility behavior. Surveys gather details on individuals' trips for a given day, such as origin, destination, time, duration, travel purpose and transport mode. The Enquête Global Transport (EGT) is the latest travel survey for the Greater Paris, conducted in 2010 by OMNIL. About 43000 residents, among 12 million, were surveyed about their travels during one day, outside holidays. The survey indicates the number of day trips per person for several transport modes. Transport modes are separated in two categories, motorized modes including public transport, cars and motorbikes, and unmotorized modes i.e. walk and bike. Survey samples are collected at different spatial resolution. Residents are aggregated by home location areas, at three different scales : the ring scale, the department scale and the canton scale (4 first digits of INSEE code).

3.4 Transport Data

3.4.1 Transport Networks

The different existing transport networks infrastructures are collected from Open data sources. First, we retrieve rail lines for underground, overground, tramway and train stations shared by Île-de-France Mobilité (STIF, 2018) (See Fig. 3.7). Second, we use OpenStreetMap (OSM, 2018) to retrieve roads (See Fig. 3.8) and high-speed rails. Roads are categorized by traffic importance. Residential roads

have the highest density and the lowest traffic. Therefore we filtered residential roads in order to reduce the computation cost.

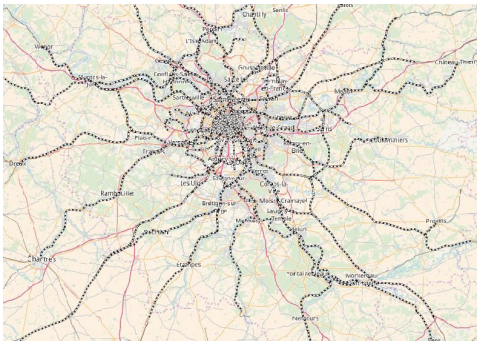

FIGURE 3.7: Greater Paris rail networks for underground, overground, tramway and highspeed rails

FIGURE 3.8: Greater Paris road networks for motorway, trunk, primary and secondary roads

3.4.2 Travel Cards Data

Their exist two public transport operators in the Greater Paris, supervised by one transport authority. The later provides an open access to daily travel cards data for public transport stations. When entering public transports, Greater Paris commuters have to swipe their travel cards, yet they are not usually required to swipe a second time when exiting the transport system. For our work, we collected one month travel cards data. The dataset is constituted of daily entry counts inside stations. When changing lines between two operators, travelers might need to swipe their smart cards more than once. There exist 249 train stations for the overground, 383 stations for underground and 186 stations for the tramway.

3.5 Comparison of Mobility Data

In this section, the different mobility data available in the Greater Paris are compared to understand how they can be used to study human mobility. First, travel surveys are open source data providing detailed information on surveyed individuals, yet they are not appropriate to study spatio-temporal variability of urban flows. Second, travel cards are advantageous to estimate flows inside the public transport system. Some cities (e.g., London, Singapore) benefit from entry and exit taps, which enable to build O-D flows. However, only entries are collected in the Greater Paris, thus requiring more complex models to reconstruct real flows. Third, although no GPS data have been used for this study, we describe how they could be collected and what are their strengths and weaknesses. GPS grant the highest spatio-temporal precision. The position is generally expressed in latitude-longitude, with a collection

frequency of a few seconds and a location error of a few meters. Still, access to GPS data is non trivial. The EU and French data legislation authorize GPS data collection only if their processing is used to deliver a service to a user through mobile apps (e.g., itinerary recommendation). The second requirement is to obtain explicit user consent to share their real-time position (Opt-in) when using such apps. In case such conditions are fulfilled for GPS data collection, users regularly turn-off their GPS when closing the apps and to reduce battery consumption. Consequently, GPS positions are collected during irregular time periods and do not capture all users movements. A second limitation with GPS is that navigation apps tend to be mainly used by drivers and walkers. At last, GPS signal quality decreases in indoor environment such as the underground. Therefore, GPS are well suited to study individual trajectories in outdoor environments, for apps users activating the GPS. Such data are not allowing to track global urban flows over time.
The characteristics of cellular network data, GPS data, surveys and travel cards are summarized in table 3.7 below.

TABLE 3.7: Comparison of multiple source data

Data Type	CDR	GPS	Surveys	Travel Cards
Source	Mobile Operators	Mobile Apps	Governmental Organizations	Transport Operators
Technology	Mobile Networks	Satellites	None	Gates Sensors
Cost	Low	Low	High	Moderate
Data Volume	High (Terabytes)	Depends on nb. of users	Small (Megabytes)	Moderate (Gigabytes)
Update Frequency	High (minutes, hours)	High (seconds)	Low (years)	High (minutes)
Granularity	Coarse	Fine	Coarse	Moderate
Access	Private	Private	Open	Partially Open
Transport Modes	Motorized	All possible yet mostly car, walk, bike	All	Public Transport
Limitations	Spatial noise Sparsity Coarseness	Spatial noise Low Indoor Precision Battery Consumption	Bias from self-reported data, Sampling Bias	Unknown itinerary Fraud Entry & Pass-by Overlaps
Main Strengths	Representative Population Flows	Highest Precision & Frequency	Individual Details e.g. modes	Representative Entry Flows at stations

3.6 Methodology

The broad problematic of estimating the mobility of a total population is generally seen as an unsupervised problem rather than a supervised problem, by reason of a lack of ground truth for training. Although some ground truth can be extracted from census and surveys, real population densities and travel flows are not available on a day-to-day basis. Therefore, regression models with traditional data sources are prone to overfitting, hence not sufficient. In addition, cellphones trajectories are unlabeled as it is extremely complex and time-consuming to manually identify the real transport mode with such a sparse and coarse geolocation. Thus, it is not possible to apply supervised classification models on mobile network trajectories for our purpose of transport mode identification. Consequently, we chose to tackle the lack of ground truth with a semi-supervised learning methodology, i.e, relying with small amounts of labeled data.
In this section, we describe our general strategy to solve our two problems i.e., the estimation of dynamic population densities and of OD matrices per transport mode. Each model can be decomposed into three main steps. The first step is data pre-processing. The second step is the learning phase which is constituted of several sub-steps, later described in details in Chapter 4 and Chapter 5. In particular, this section describes the different techniques, namely regression, clustering and bayesian inference, used in the thesis. The third step is performance evaluation, including validation with external data. Eventually we provide interpretation on the model and the results. In what follows, we present the techniques being used during the learning phase and for performance evaluation.

3.6.1 Regression

In this thesis, we use linear regression models to calibrate our estimates, using mobile network records, with external data (i.e., census, survey, travel-cards). A linear regression estimates an output variable Y from an input vector X such as Y is a linear combination of the parameters (β_0,β_1) : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i \in [1, n]$. Linear regression models are used both during the learning phase for population densities (see Chapter 4), to calibrate the OD flows estimates and for validation (see Chapter 5). The goodness of fit of a linear regression model is evaluated with the coefficient of determination $R^2 = \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y})^2}$, where y_i are the ground truth values having a mean value $\bar{y}$ and $\hat{y_i}$ are the predicted values.

3.6.2 Clustering

Through the thesis we also make use of several clustering methods. Clustering is a vast set of techniques used to explore data to unveil the underlying structure. Clustering has been applied for information retrieval and data filtering (e.g., identification of outliers). The data are grouped into clusters according to their similarity. A broad choice of clustering algorithms have been presented in the literature, yet choosing one particular method is not straightforward and is problem specific. For our work, we benchmark several algorithms in order to find an optimal partitions of our datasets. Three clustering algorithms are tested, namely k-means, Hierarchical clustering and DBSCAN (ESTER et al., 1996). Such clustering are depicted as scalable with large datasets and high dimensional data. Each of the aforementioned clustering is described below.

K-means

The k-means algorithm is among the most popular method and has been used over 60 years. The k-means starts with a random initialization of clusters centers. Each sample is assigned to the cluster having the closest center. The new center is calculated as the centroid, i.e. mean, of the cluster. In order to find the optimal clusters centers, the k-means minimizes the squared error between the points of a cluster and the center. The algorithm updates recursively the centers until it minimizes the sum of squared error over all clusters : $\sum_{k=1}^{K} \sum_{x_i \in k} \|x_i - \nu_k\|^2$. One reason of the wide use of k-means is the fast computation time, having a $\circ(N)$ complexity. However the clusters produced by k-means are convex shaped, which is not always an appropriate partition of the data.

Hierarchical Clustering

The hierarchical clustering is used to find clusters with an underlying hierarchical structure, i.e. parents and children clusters. The agglomerative hierarchical clustering (AHC) merges the closest pair while the divisive splits the farthest pair. The time complexity is $\circ(N^2 \cdot log(N))$. The AHC starts with N clusters of size 1, N being the number of samples. Each point is recursively merged with its closest neighbor according to a linkage criterion and a distance function. We test the algorithm with linkage types “ward", “complete", “average" and distance functions euclidean (l2), manhattan (l1) and Pearson correlation. Complete linkage minimizes the maximal distance between two points from two clusters. Average linkage minimizes the average distance between clusters points. Ward linkage, with euclidean distance,

minimizes the sum of squared error $ESS = \sum_{k,i,j} |X_{ijk} - \bar{x}_{kj}|^2$, where X_{ijk} is the sample value for sector i, feature j and cluster k ; $\bar{x}_{kj}$ is the mean value of feature j for cluster k. The clustering applies until all data points are merged into a single cluster of size N.

DBSCAN

The DBSCAN acronym stands for Density Based Spatial Clustering of Application with Noise. The algorithm recursively merges a set of point belonging to the same neighborhood, according to a density threshold (eps). Outliers are detected as points in lowest density regions. The DBSCAN is well adapted to datasets having clusters of similar density separated by lower density regions. It has shown good performance on clusters of various shapes and on spatial datasets. The average run time complexity is $\circ(N \cdot log(N))$

3.6.3 Optimal Number of Clusters

The determination of the optimal number of cluster k is a non trivial task, which is often dependent on the specific nature of the data and of the problem. Several internal evaluation metrics are used to assess the clustering performance and to identify the optimal cluster number. For our study, we start by checking each metric individually. By default we select the optimal value, being either the minimal or the maximal depending on the metric. In case of an abrupt variation, we rather use the elbow method to get the value at the intersection of the asymptotes. Different metrics might produce different optimal values therefore we look for a number of cluster representing the best trade-off between metrics. Several strategies are possible such as selecting the k returned by the highest number of metrics. A visual projection of clusters in a lower dimension space can strengthen the assumption for k. In what follows we describe the different evaluation metrics techniques used in the thesis.

Silhouette

The Silhouette (S) is used to evaluate clusters separability (KAUFMAN et ROUSSEEUW, 2009).

$$s_{ik} = \frac{b(i) - a(i)}{max(b(i), a(i))} \tag{3.8}$$

$$S_k = \frac{1}{N_k} \sum_{i=1}^{N_k} s_{ik} \tag{3.9}$$

$$S = \frac{1}{N} \sum_{k} S_k \tag{3.10}$$

where $a(i)$ is the average intra cluster distances for sample i and $b(i)$ is the lowest value among average inter-cluster distances. Here N_k stands for the size of cluster k. The number of samples equals N. The optimal number of clusters maximizes the silhouette (Y. LIU et al., 2010).

Calinski-Harabasz

The Calinski-Harabasz (CH) index (CALIŃSKI et HARABASZ, 1974) can be defined as the ratio of between clusters over within clusters dispersion. Let B and W be respectively the between and within clusters scatter matrices. The CH index is computed as follows.

$$CH = \frac{Trace(B)}{K-1} \cdot \frac{N-K}{Trace(W)} \tag{3.11}$$

$$Trace(B) = \sum_{k=1}^{K} n_k \|\nu_k - \nu\|^2 \tag{3.12}$$

$$Trace(W) = \sum_{k=1}^{K} n_k \sum_{i=1}^{n_k} \|x_i - \nu_k\|^2 \tag{3.13}$$

Having N the number of samples, K the number of cluster, ν the center of the dataset, n_k the number of points in cluster k and ν_k its centroid. The maximal CH value gives the optimal number of cluster.

S_{dbw} validity index

In addition we use the S_{dbw} validity index which is a trade-off between clusters densities and variances.

$$S_{dbw}(k) = Scat(k) + Dens_{db}(k) \tag{3.14}$$

$$Scat(k) = \frac{1}{k}\sum_{i=1}^{k}\frac{\sigma(\nu_i)}{\sigma(D)} \tag{3.15}$$

$$Dens_{db}(k) = \frac{1}{k(k-1)}\sum_{i,j=1}^{k}\frac{dens(u_{ij})}{max(dens(v_i), dens(v_j))} \tag{3.16}$$

where ν_i denotes centroid of cluster i and u_{ij} is the middle point between clusters i and j, i.e. at mid distance from the two centroids (ν_i,ν_j). The scattering index $Scat$ is used to estimate the intra-cluster compactness based on standard deviations σ of clusters over total dataset D. The term $Dens_{db}$ represents clusters densities. It calculates the average ratio of clusters middle point densities over clusters centers densities. The underlying assumption is that well defined clusters are denser around their centroids than at their mid distance. This index has been depicted as the most performing among internal clustering evaluation metrics (HALKIDI et VAZIRGIANNIS, 2001 ; Y. LIU et al., 2010). A small S_{dbw} grants smallest clusters dispersions and highest density of points around clusters centroids. The optimal number of cluster is found when the index reaches its minimum.

3.6.4 Bayesian Inference

Bayesian inference is used for the transport mode estimation model, described in Chapter 5. This technique is a statistical inference based on the Bayes theorem : $P(Y|X) = \frac{P(X|Y)\cdot P(Y)}{P(X)}$. Here, the posterior probability $P(Y|X)$ is the conditional probability of a variable Y given an observed variable X. The posterior probability is derived from an initial hypothesis updated by real observations. The former is the prior probability noted $P(Y)$ while the later is formed by the likelihood $P(X|Y)$ and the marginal likelihood noted $P(X)$.

3.6.5 Validation

Validation is the crucial final step asserting whether the model is able to predict correctly. Classic supervised learning models have a training phase followed by a test phase. The training and test sets are both generated from the same initial data source. The predicted labels from the test set are confronted to the real labels to asses the performance of the supervised model. For unsupervised learning tasks, the data is unlabeled. The performance evaluation generally consists in evaluating the separability of the output, e.g. the clusters. Still, the final objective of our work is to produce valid estimates of the number of visitors and travelers in urban areas.

Therefore, we leverage all other existing mobility data available in order to construct external test sets. The validation step thus consists in comparing two datasets produced from different sources, such as the first data source is used to build the model while the second data source is held out for validation.

3.7 Summary

In this chapter, we have presented the different datasets used for the thesis. First, we have collected mobile network data, composed of both Call Detail Records and passive records. We have successively described the spatio-temporal properties and the initial pre-processing steps of the raw data, followed by a statistical analysis of cellphone trajectories. The second data type used for the thesis are population information from the national household census and the regional travel survey. Our third data source is constituted of two public transport datasets : transport networks and travel-cards counts. Then, we have compared the different existing mobility data to highlight their respective characteristics, strengths and limitations. At last we have introduced the unsupervised learning scheme applied in this thesis. The later is a combination of four main steps : data pre-processing, unsupervised classification (e.g., clustering), performance evaluation with model interpretation and finally validation. In the next chapters, we present our models combining the three aforementioned data sources.

4 Dynamic Population Attendances in Urban Areas

4.1 Introduction

In 2017, mobile networks have connected more than 5 billion individuals owning more than 7 billion SIM cards worldwide. Although this represents 67% of the global population, most developed countries have reached 100% penetration rate. The aggregation of mobile phone traces thus represents a huge potential for research studying the size and behavior of urban populations. In particular, the estimation of fine-grained daily population dynamics could bring valuable knowledge for several urban planning applications, such as travel demand, anomalous event detection and location based services. Past research has used mobile network data to identify and study seasonal variations of visitor rates (GIRARDIN et al., 2009), human mobility (BERLINGERIO et al., 2013a ; JAHANGIRI et RAKHA, 2015 ; IQBAL et al., 2014), socio-economic characteristics (SOTO et al., 2011), disease transmission (WESOLOWSKI et al., 2012), pollution rates (SHANG et al., 2014), attendances rates (BOTTA et al., 2015), emergencies and disasters (BAGROW et al., 2011), and to classify urban areas according to human activity (FURNO, M. FIORE et al., 2017), Traditional data sources on population are surveys and censuses which have several limitations. A survey is restrained to a small subset of population due to cost constraints. Meanwhile a census traditionally reports numbers of residents and is not accounting for daily numbers of visitors. In addition, both surveys and censuses suffer from low update frequency. Mobile network data, by contrast, provide ubiquitous and up-to-date information on a representative and larger sample of the population. The potential of mobile network data to study population densities was first highlighted by Ratti et al (RATTI et al., 2006). Later, several works demonstrated a power law relationship between census population size and mobile phone activity (DEVILLE et al., 2014 ; DOUGLASS et al., 2015). Recent works have combined census and cellular data with land-use information to improve model performance (DOUGLASS et al., 2015 ; F. XU et al., 2016 ; KHODABANDELOU et al., 2016a). Past studies used coarse administrative level resolution (DEVILLE et al., 2014), cellular scale (KHODABANDELOU et al., 2016a), land-use regions (F. XU et al., 2016) and grid-square scale (DOUGLASS et al., 2015 ; DEVILLE et al., 2014). The case study regions concerned France and Portugal (DEVILLE et al., 2014), Shanghai (F. XU et

al., 2016) and Italy (BOTTA et al., 2015 ; DOUGLASS et al., 2015 ; KHODABANDELOU et al., 2016b), for which open data is available (TIM, 2014). Since raw mobile network locations are coarse, sparse and noisy, it is essential to carefully pre-process the data. Still the nature of the data (e.g., sampling rate) and the pre-processing (e.g., noise reduction) of each study remain unclear or partly confidential. Although Deville et al. (DEVILLE et al., 2014) proposed an approach for dynamic population estimation using Call Detail Records from France, the presented estimates have been confronted to census data exclusively and no day-time validation study was performed.

The following chapter is an extension of the article entitled “Using Mobile Phone Data Analysis for the Estimation of Daily Urban Dynamics” published in the proceedings of IEEE ITSC 2017 (BACHIR, GAUTHIER et al., 2017). The objective of this work is to estimate the dynamic population densities and the total number of individuals in a given area, at any time of the day. To solve this problematic, we need to address several issues. First, with mobile network data, one can only detect the active cellphones from customers of a given mobile provider. In addition, devices are initially detected at the mobile network cell resolution which is coarse and not representative of urban areas. In order to produce meaningful results for urban planning, one has to change the raw spatial scale to associate populations with specific areas (e.g., towns, postcode areas, census blocks, grids etc.), while minimizing the errors. At last, it is difficult to calibrate and validate a model by reason of a lack of ground truth at fine spatio-temporal resolutions.

In what follows, we present our model for the estimation of time variant population rates using mobile network data from the Greater Paris. We provide estimates at fine-grained resolution, corresponding to the smallest census blocks, for the first time in this region. Initially, raw anonymized cellphone positions are pre-processed and aggregated at the cellular level. In order to dispatch phones on census blocks, we compare our mapping method to the state of the art (DEVILLE et al., 2014). The identification of the population scaling factors is two-fold. First, the static parameters are derived using multiple loglinear regression models on mobile phone presence densities to estimate the census population densities, given each 15 min time slot. Since census population is constant in time, this model is called ‘static’. Census reports residential populations yet other visitors can be present in a given area. Therefore, prior to the static calibration, a clustering is applied on census blocks in order to filter non-residential areas, using the median week signature of blocks (FURNO, STANICA et al., 2015). The static model performance is evaluated at several resolutions using a granularity ranking metric. A detailed interpretation of each static parameters is provided for the first time. The β parameter represents the inverse penetration rate. Meanwhile, the α parameter is inversely proportional to both market share and mobile phone usage frequency. Second, according to our interpretation, we modify the static parameters to account for the variation of mobile phone activity during the day. During early morning hours we observe a linear relation between

inverse mean mobile activity and static coefficients. The dynamic parameters are obtained from fitting the inverse mean mobile activity to the static parameters. In addition, we rescale our estimates a second time to fit the total population of the region. Thus, we provide a novel dynamic model to estimate both daily population densities and counts by exploiting two state-of-the-art techniques. Our dynamic model is validated against sport events attendances from two stadiums. The results show that spectators numbers are estimated with smaller errors compared to past approaches. Section 4.2 presents our model and Section 4.3 describes the main results. A conclusion and a summary of the chapter are provided in Section 4.4 and Section 4.5.

4.2 Method

4.2.1 Overview

This section presents the methodology for the estimation of time variant population densities, using mobile network data. In order to rescale the number of detected mobile phones to the total population present in a given area, we estimate population scaling factors. The only candidate ground truth for calibrating the static model is the census population. Such static model relies on the hypothesis that the inhabitant density approximates the total population density. This hypothesis is assumed reliable for residential areas during night time as users are expected to stay home. In order to improve the static model performance, we identify and filter non-residential blocks to reduce noise from our training data. Then, we train several loglinear regression models to fit the mobile phone density with the population density according to Eq 2.5. One couple of parameters is generated for each time-slot. Since census population is the number of inhabitants for a given year, this variable does not change over time (e.g., days, hours etc.), then such model is 'static' and not dynamic. The static parameters are not sufficient to account for the fluctuations of urban flows during day time. Therefore, we propose a method aimed at correcting the static parameters into dynamic ones, in order to estimate both population densities and number of visitors anytime. Eventually the dynamic scaling factors are validated against the attendances in two stadiums. The steps of our methodology are summarized below.

1. Processing mobile phone data :
   - Calculation of mobile phone presence for each cell and each time slot ($t = 15$ min) (Sec. 4.2.2).
   - Spatial mapping of mobile phone presence from cells to census blocks (Sec. 4.2.3).

— Filtering of non-residential areas by clustering blocks according to mobile phone activity (see Sec. 4.2.4).

— Normalization of census population and mobile phone presence by block area to obtain densities, prior to applying Eq. 2.4.

2. Static model : for each time-slot, we fit a loglinear regression between presence density and population density, according to Eq. 2.5, and extract the static scaling factors $\hat{\alpha}$ and $\hat{\beta}$. Model performance is evaluated using several metrics, defined in Sec. 4.2.5.

3. Dynamic model : scaling factors are tuned to account for time variant population densities.

4. Validation : the dynamic model is extrapolated to estimate visitor counts in stadiums.

Next, we describe the processing of the mobile phone presence density, the spatial scaling methods, the clustering of blocks and the evaluation metrics used to assess the model.

4.2.2 Mobile Phone Presence

For this study, about 10 billions records, geolocated in the Greater Paris, have been collected, spanning over a period of 61 days. The raw data is pre-processed as follows. For each 15 min time interval, we retrieve the last observed position of a user. If no communication occurs during the interval, we assume that the user is undetected and his presence is not accounted. Then, the mobile phone presence is calculated by counting the number of cellphones per cell and per time slot.

Input: list of mobile network cells X ;
list of call detail records ;
list of cellphones u ;
list of timeslots t ;
Output: list of mobile phone presence P ;
foreach u **do**
  foreach t **do** **if** u *is active* **then**
    get X from last record ;
    $P_X \mathrel{+}= 1$
  end
  ;
end

Algorithm 1: Mobile phone presence

4.2.3 Spatial Mapping

Two mapping methods are applied to project mobile phone presence from mobile network cells to census blocks.

Areal Weighting (AW)

The first method is the state-of-the-art areal weighting defined in Eq. 5.14. Following the definition given by DEVILLE et al., 2014, the areal weight of a census block b_i which intersect a mobile network cell c_j is :

$$w^{(1)}_{b_i,c_j} = \frac{A_{b_i \cap c_j}}{A_{c_j}} \tag{4.1}$$

where A_{c_j} is the area of cell c_j and $A_{b_i \cap c_j}$ is the area of the intersection of block b_i with cell c_j. Such weight equals 1 in case a cell j is fully included in the block i. Yet, in many cases, mobile network cells are wider than census blocks. The median area of a cell is 0.5 km^2 while it is 0.3 km^2 for blocks. The weight defined in eq. 4.1 satisfies the constraint :

$$\sum_i w^{(1)}_{b_i,c_j} = 1 \tag{4.2}$$

Consequently each block inherits from a weighted subset of the mobile phone counts from the cell. Such weighs spread cellphones proportionally to the area of the intersection between a block and a cell.

Areal Population Weighting (APW)

In reality, census block population is not proportional to block area. This statement holds true for any type of population (residents, workers, visitors etc.). In other words, a wide block can have a small number of visitors (e.g., a block including parts of a forest) and reciprocally. Consequently, in case several blocks are covered by the same network cell, larger blocks are not necessarily containing more users. In order to correct this bias we add a second weight term, noted $w^{(2)}$, to account for the census population.

$$w^{(2)}_{b_i,c_j} = \frac{\rho_{b_i}}{\sum_k \rho_{b_k}} \quad \forall j,\ i \in k \tag{4.3}$$

where ρ_{b_i} is the residential population density of census block b_i and $\sum_k \rho_{b_k}$ is the total population densities of all blocks intersecting cell c_j. This weight satisfies the constraint :

$$\sum_i w^{(2)}_{b_i,c_j} = 1 \tag{4.4}$$

Eq. 5.14 is accordingly adapted.

$$\sigma_{b_i} = \sum_j a.w^{(1)}_{b_i,c_j}.w^{(2)}_{b_i,c_j}\sigma_{c_j} \tag{4.5}$$

where constant a is added for normalization purpose to satisfy the following constraint.

$$\sum_i a.w^{(1)}_{b_i,c_j}.w^{(2)}_{b_i,c_j} = 1 \tag{4.6}$$

Mobile phones are distributed from cells to blocks using both area and census densities of blocks. Both weights are calculated considering all neighboring blocks covered by the same cell. Therefore our method combines the areal weighting with a population weighting into our novel Areal Population Weight (APW) mapping.

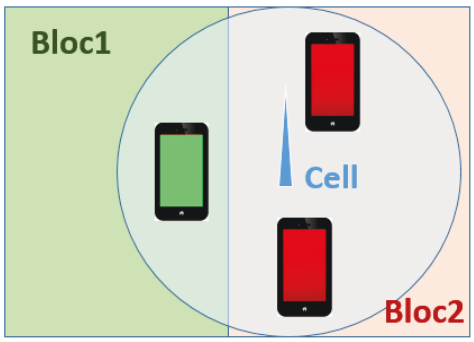

FIGURE 4.1: AW mapping : mobile phones are mapped according to the weights of the areas i.e. the ratio of the intersection area over total block area.

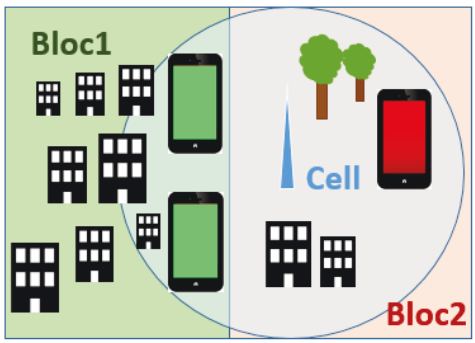

FIGURE 4.2: APW mapping : mobile phones are mapped according to the weights of the areas and of population densities.

4.2.4 Data Filtering with Mobile Activity

The static model uses the census population by default due to a lack of existing ground truth. Non-residential areas contain different types of visitors, in addition of the residents. Such areas need to be identified and removed from the training set in order to reduce noise. One way to obtain the main type of visitors in an area is to collect the land-use information, which can be obtained from open data repositories. However, land-use are based on environmental and material facilities (i.e., buildings, forest, river etc.) and are not well adapted to represent human activities. Censuses are another source of information. The latter provide a classification of census blocks according to the number of residents and workers. However, census activity types are based on arbitrary rules. For instance, the activity type of Greater Paris blocks is based on the following rule : "For a given block, if the number of workers is at least twice the number of residents then label it as a business area" (INSEE, 2013). According to the census, the Greater Paris has 78% residential blocks, 2%

business blocks and 20% blocks labeled as 'other'. Such classification is not reliable enough for our work, therefore we generate our own partitioning of the territory.
Following the method of Furno et al. FURNO, STANICA et al., 2015, we cluster blocks on their Median Week Signature using calls and texts. In the work of Furno et al., several normalization and distance metrics have been tested. As a result, one technique generally outperformed the others. The later combines the standard scaler with the Pearson correlation coefficient as distance metric. The standard scaler normalization is based on the z-score : $z = \frac{x-\nu}{\sigma}$, where x is the initial data with mean ν and standard deviation σ. The second best performing approach combines feature normalization by total mobile phone activity of the day, coupled with the euclidean distance. In their study the AHC is used with average linkage.
For our study, we test the two normalization, i.e. standard scaler and daily normalization on our MWS dataset. We benchmark several clustering algorithms : k-means, DBSCAN and AHC. For the AHC we test different linkage : ward, average, complete. For DBSCAN and AHC we test several distance metrics : euclidean, Manhattan and Pearson correlation. In order to determine the clustering with best performance and determine the optimal number of clusters, we calculate the Silhouette score, the Calinski-Harabasz index, the S_{dbw} validity index and check for the dendrogram of the AHC. Then, once the region is partitioned into optimal activity clusters, we consider three criteria to identify the candidate cluster for the residential activity type. First we calculate the average signature of the mobile phone activity per cluster. The residential area clusters are expected to obtain lower activity than business area clusters. Second, we refer to the size of the clusters in order to discard preferably smaller clusters, as the residential areas are expected to constitute a substantial part of the region (nearly 80% according to the census). Third, for each cluster, we calculate the Pearson correlation coefficient between the mobile phone presence density and the census population density during night time (between 0-6AM). Thus, we assess that the candidate clusters for training have high correlation with the ground truth.

4.2.5 Evaluation Metrics

Pearson Correlation Coefficient

The Pearson correlation coefficient, noted r, is used to gauge the correlation between population prediction, noted $\hat{\rho}$, with the census population ground truth noted ρ.

$$r_{\rho,\hat{\rho}} = \frac{COV(\rho, \hat{\rho})}{\sigma_{\hat{\rho}} \cdot \sigma_{\rho}} \tag{4.7}$$

where $COV(\hat{\rho}, \rho)$ is the covariance between variables $\hat{\rho}$ and ρ having standard deviations $\sigma_{\hat{\rho}}$ and σ_{ρ}.

R-squared Coefficient

In order to assess the performance of the regression model during training and evaluation, we use the coefficient of determination R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\rho_i - \hat{\rho}_i)^2}{\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2} \tag{4.8}$$

where N is the total number of blocks, ρ_i and $\hat{\rho}_i$ are respectively the census population and the estimated population of block i and $\bar{\rho}$ is the average population across blocks.

Normalized Root Mean Squared Error (NRMSE)

The normalized root mean squared error ($NRMSE$) is calculated to evaluate the correctness of the predicted populations :

$$NRMSE = \frac{1}{\rho_{max} - \rho_{min}} \sqrt{\frac{\sum_{i=1}^{N}(\rho_i - \hat{\rho}_i)^2}{N}} \tag{4.9}$$

where ρ_{max} and ρ_{min} are the maximal and minimal ground truth. For the static model, ρ is the census population and N is the number of block. For the validation study, ρ is the stadium attendances and N is the number of sport games.

Mean Absolute Error (MAE)

In addition of the $NRMSE$, we calculate the mean absolute error between stadium attendances ρ and predicted attendances $\hat{\rho}$.

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |\rho_i - \hat{\rho}_i| \tag{4.10}$$

where N is the number of sport games.

Mean Absolute Percentage Error (MAPE)

Similarly, we use the mean absolute percentage error which gives a percentage value of the MAE.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\rho_i - \hat{\rho}_i|}{\rho} \cdot 100 \tag{4.11}$$

Spatial Rank

At last, we use a spatial ranking to assess model performance at different spatial scales. One block is considered to represent the spatial rank 0. Then, for a given block b_i, we determine its neighboring blocks having at least one common border. The aggregation of one block with its first rank neighbors is considered to be the spatial aggregation at rank 1. Neighbors of blocks from rank 1 are aggregated to form the rank 2 etc. The spatial scales range from rank 1 to rank 5 aggregates of blocks and the last scale is the 'Commune'.

4.3 Results

4.3.1 Filtering Blocks Activity Types

Test with StandardScaler Normalization

The best approach of FURNO, M. FIORE et al., 2017 combines a standard scaler with the correlation distance and an Agglomerative Hierarchical Clustering (AHC) with average linkage. Thus, we test the standard scaler and compare several clustering algorithms with several distance functions. The t-sne projections are displayed in Fig. 4.3. From the visualization it appears indeed that the AHC with correlation distance provides the best clustering. Both average and complete linkage are acceptable candidates.

In order to determine the best clustering among the two, we compare their evaluation metrics for values of k varying between 2 and 10 (see Fig. 4.4). The average linkage generally outperforms the complete linkage regarding the three metrics. A cut-off value can be observed for $k = 4$ for both Calinski-Harabasz and S_{dbw} index. In addition, this value of k is also a local maximum for the Silhouette. Therefore, we assume the optimal number of cluster is $k = 4$ according to these metrics.

The output of the clustering, with $k = 4$, is then projected on the Greater Paris map (see Fig. 4.5). As a result, the clustering fails to separate blocks inside Paris.

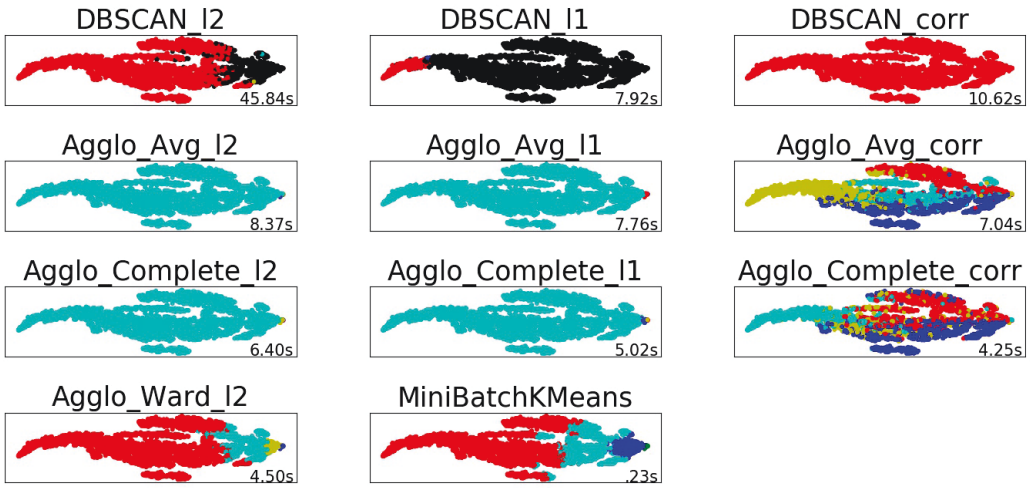

FIGURE 4.3: Benchmark of clustering with daily normalization. t-SNE projections are displayed using $K = 4$. t-SNE Parameters : $perplexity = 30$, $learning\_rate = 300$, $n\_iter = 2000$.

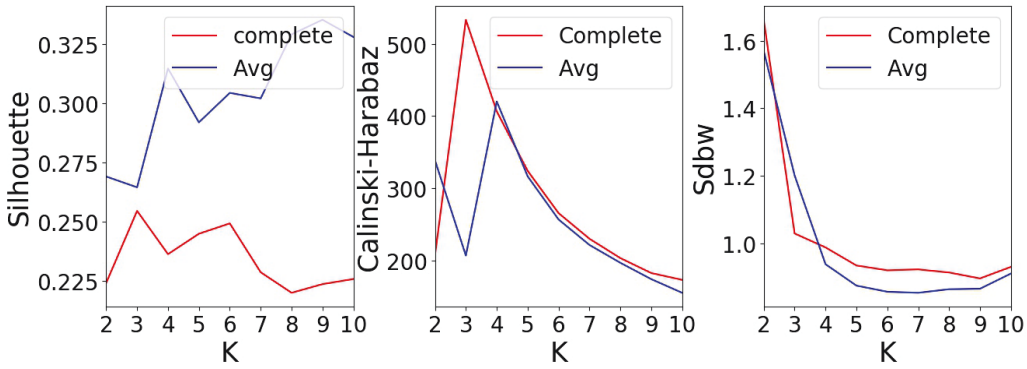

FIGURE 4.4: Clustering evaluation metrics for the standard scaler normalization, correlation distance and hierarchical clustering.

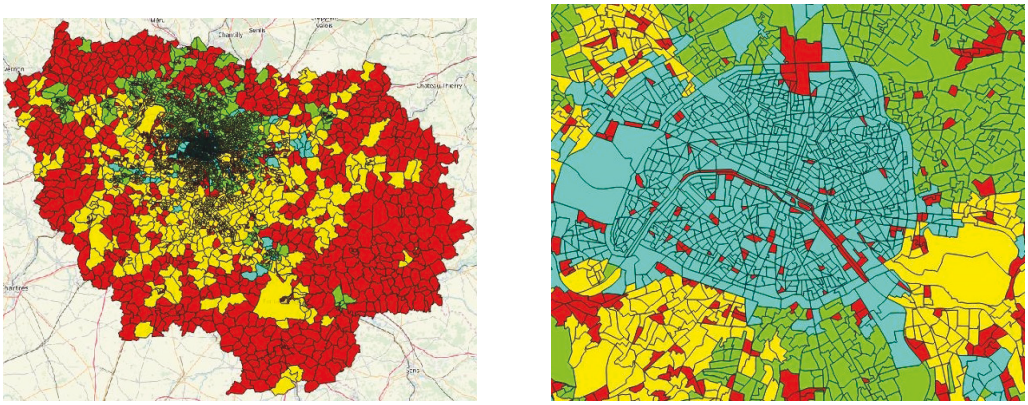

FIGURE 4.5: Census blocks in the Greater Paris colored according to their MWS clusters, displayed for $k = 4$. The clustering for MWS with Standard Scaler fails to separate blocks in Paris city center (Right), for any value of k.

This observation still holds for higher values of k. On possible explanation is that FURNO, M. FIORE et al., 2017 model has been applied on (big) city centers, such as Paris which has a 10^2km^2 area, instead of complete regions. In our case we

consider the whole Greater Paris region having a 10^4 km^2 area and including both Paris and the suburb. The region has a much more heterogeneous mobile phone activity compared to the city center. Therefore, although the standard scaler has shown competitive results on Paris, it seems not the best normalization when the scale of the data is extended to the Greater Paris.

Test with Daily Normalization

Consequently we apply the daily normalization on the MWS and benchmark the clustering algorithms. This time, clusters are much harder to visualize with the t-SNE. Three candidate algorithms produce clusters of comparable sizes and are thus retained for comparative analysis. The three algorithms are k-means, ward AHC with euclidean distance and complete linkage AHC with correlation distance. The evaluation metrics are calculated to assess the performance of the clustering (see Fig. 4.7). The Silhouette score reveals that the k-means fails to separate the data

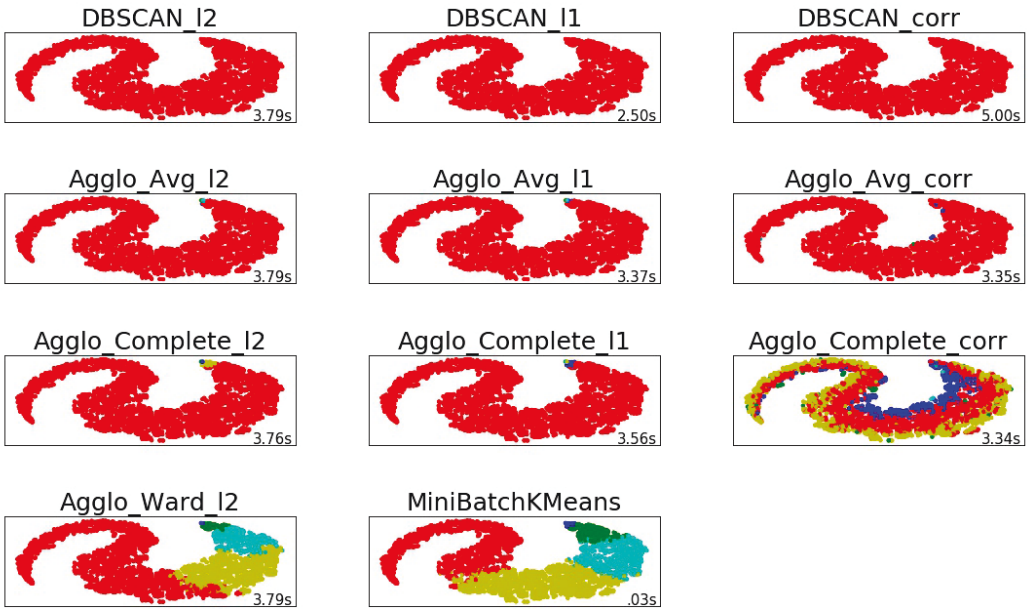

FIGURE 4.6: t-SNE projection for the $K = 5$ clusters. Parameters : $perplexity = 30$, $learning\_rate = 300$, $n\_iter = 2000$.

while the Calinski-Harabasz index discards the complete linkage AHC. The ward AHC is retained as the best performing algorithm regarding the three metrics. Still, we note that the performance is lower regarding the S_{sbw}. This could be explained by the fact that the inter-cluster densities remain high compared to the intra-cluster densities, as hinted visually with t-SNE.

For the ward clustering, the highest Silhouette is obtained for $K = 3$, corresponding to the best separability. Yet, $K = 5$ grants a Silhouette cut-off value. The highest Calinski-Harabasz score is obtained for $K = 5$, for which there is also a cut-off.

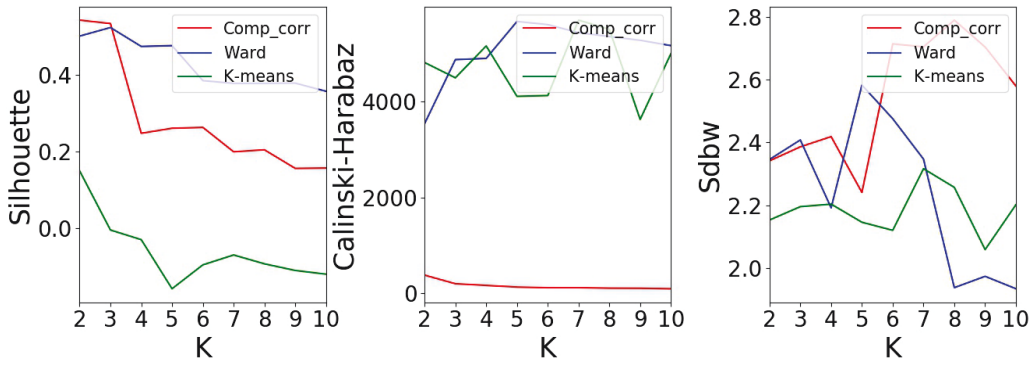

FIGURE 4.7: Clustering evaluation metrics for the daily normalization. Candidate algorithms are k-means, AHC ward and AHC with complete linkage and correlation distance.

Consequently we retain $K = 5$ as the number of clusters for Greater Paris blocks. Compared to the clustering with standard scaler normalization, both silhouette and calinski-harabasz are higher, confirming that both clusters separability and dispersion are more optimal, despite the inter-cluster compactness.

The 5 clusters are projected using t-SNE on Fig. 4.6. A projection of the clusters on the Greater Paris map is depicted on Fig. 4.8. This time both the suburb and city centers blocks are separated. The mean signature of each cluster is plotted on Fig. 4.9. For each cluster, we report its size and the correlation between mobile phone presence and census population, in Tab. 4.1. The largest cluster in size is cluster C_0, with 62% of blocks, represented in red. This cluster obtained the lowest mobile phone activity signature and a reasonably high Pearson correlation coefficient of 0.81 in median. The second largest cluster is C_3 with 25% of blocks (yellow). Such cluster has the second lowest activity and a high median correlation (0.86). Although both C_0 and C_3 are good candidate clusters for our training set, grouping C_0 and C_3 leads to a drop of the correlation with census population. Cluster C_0 is thus retained for model training. In addition, the clustering unveils important business areas concentrating important mobile phone activity. With 0.1% of blocks, cluster $C1$ (dark blue) is the smallest in the size yet has the highest activity. $C1$ is constituted of the major business and touristic areas of the region, namely La Défense and Disneyland and contains the most frequented commuting zones such as CDG airport and Châtelet. Interestingly, $C1$ obtained the highest correlation with the census population (0.94 in median). One possible interpretation is that $C1$ areas are extremely frequented at anytime thus the number of visitors, workers and residents has small fluctuations. Consequently, as the census population is constant, the correlation remains high. Eventually, the lowest correlations are obtained for clusters $C2$ (cyan) and $C4$ (green) which are constituted of several recreational and business areas. Therefore, clusters $C1$, $C2$, $C3$ and $C4$ are filtered out from the training set.

TABLE 4.1: Statistics per cluster. Pearson Correlation Coefficient are calculated between MP density and census population of blocks, for each time interval of 15 min. Maximal, average, median and minimal values are reported.

Cluster	$Size$(%)	r_{max}	r_{avg}	r_{med}	r_{min}
ALL	100	0.79	0.60	0.59	0.47
C_0	62	0.85	0.79	0.81	0.56
C_1	0.1	0.99	0.93	0.94	0.69
C_2	11	0.93	0.75	0.76	0.52
C_3	25	0.92	0.82	0.86	0.54
C_4	2	0.88	0.71	0.71	0.50
$C_0 + C_3$	87	0.82	0.74	0.77	0.53

Input: List of time interval $t \in$ 00-06 AM ;
List of days D ;
Number of clusters C ;
List of census blocks i
with population density ρ_i
and mobile phone presence σ_i ;
Output: List of correlations r_c for each cluster c
foreach $c \in C$ **do**
  foreach $t \in T$ **do**
    foreach $i \in c$ **do**
      get the median of presence density over D :
      $\sigma_{t,c}(i) = median_D(\sigma_c(i))$;
      get the population density $\rho_c(i)$
    end
    Calculate the correlation over all blocks in c during t :
    $r_{t,c} = corr(\sigma_{t,c}, \rho_{t,c})$;
  end
  Calculate the min, max, mean, median value :
  $r_{min} = min(r_{t,c})$; $r_{max} = max(r_{t,c})$;
  $r_{avg} = avg(r_{t,c})$; $r_{med} = med(r_{t,c})$;
end
return $r_c = \{r_{min}, r_{max}, r_{avg}, r_{med}\}$

Algorithm 2: Calculation of Pearson correlation per cluster between mobile phone presence density and population density

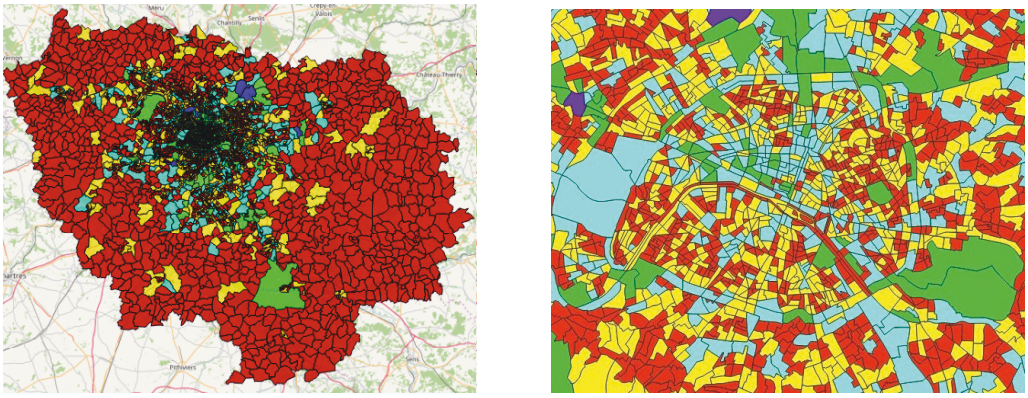

FIGURE 4.8: Census blocks in the Greater Paris colored according to their mobile activity clusters, obtained with daily normalization and $k = 5$. The red cluster is selected for training the static model.

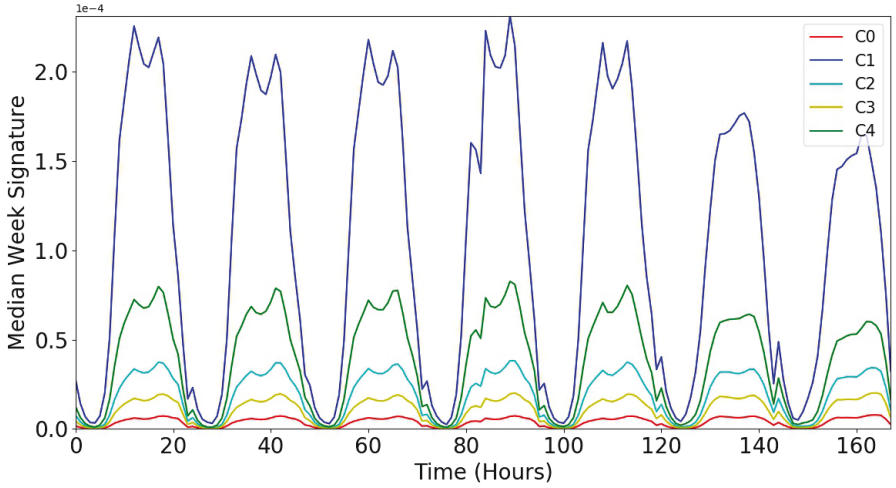

FIGURE 4.9: Median week signature of mobile phone activity averaged per cluster.

4.3.2 Comparison of Mapping Methods

The performance of the two mapping methods, i.e. areal and APW, is evaluated. Two static models are applied in order to test both mapping on the mobile phone data. The estimated mobile phone presence densities are compared to census population densities per blocks. Similarly, the estimated mobile phone counts are compared to the absolute census population per blocks. The evaluation metric is the Pearson correlation coefficients. Blocks are aggregated at different spatial scale according to our ranking metric (see Tab. 4.2).

First, the general model performance exhibits two similarities for both mapping techniques. Indeed, the performance gradually increases with the ranking (see Tab. 4.2). Correlations are the lowest at the block scale. The coarser scale, i.e. Commune, obtains the highest correlations, above 0.9 for both mapping. This reveals

TABLE 4.2: Pearson Correlation Coefficient between census population and mobile phone, for densities and absolute counts. Results are given for different spatial scales i.e. blocks neighbors rank

	Areal Weight (AW)		Areal Pop Weight (APW)	
Scale	$r_{density}$	r_{pop}	$r_{density}$	r_{pop}
Block (IRIS)	0.58	0.21	0.65	0.40
Rank 1	0.73	0.44	0.78	0.49
Rank 2	0.78	0.56	0.82	0.58
Rank 3	0.81	0.65	0.85	0.66
Rank 4	0.83	0.71	0.86	0.72
Rank 5	0.85	0.75	0.88	0.77
Commune	0.91	0.90	0.93	0.91

that changing the raw cellphone positions to a fine-grained scale introduces position errors. In this case, cellphones are not accurately placed on their true blocks. Still, aggregating blocks recursively with their neighboring blocks, from rank 1 to 5, reduces the errors. This shows that the true blocks of misplaced cellphones are in fact among the neighboring blocks. One possible cause of the positioning errors comes from the hypothetical areas of mobile network cells. The later are represented as circular areas in our case (another existing classic representation is voronoi areas), yet in reality, the signal area of an antenna is more similar to a tear-drop shape. Thus, the existing inaccuracy in cells area affects the mapping. Another limitation of the mapping comes from an heterogeneous repartition of populations in blocks, as the later do not occupy the full surface of blocks. As a result, both cells and blocks areas possibly introduce a bias in the mapping.

A second general observation is the correlation gap between densities and absolute counts. For both models the predictions on population densities outperformed the absolute population counts. As census population and blocks areas are heteroscedastic, predicting the absolute populations are more difficult. Such gap is more pronounced at the block scale yet all correlations converged around 0.9 at the Commune scale.

From the results, the APW generally outperforms the state-of-the-art AW mapping. This is particularly notable at the block level, for which the APW correlations are respectively 6% higher for density and 19% higher for population counts. The APW method is therefore retained for the static model.

4.3.3 Static Model

Performance Evaluation

The next step of the approach is to fit the static model between mobile phone densities and population densities of census blocks. In order to optimize the model performance and prevent underfitting or overfitting, it is crucial to train our model on a reasonably large enough observation period and on the largest scale possible. Thus, we retain one month observation period. The census population is characteristic of a residential activity type which has been identified for C_0. The ground truth is not reliable for other activity types, hence other clusters, which contain visitors. Therefore we train our model on Greater Paris blocks belonging to cluster C_0. In addition the model is trained during night time (0AM-06AM) i.e., when residents are expected at home. One model is trained for each 15 min time slots (see Fig. 4.11).

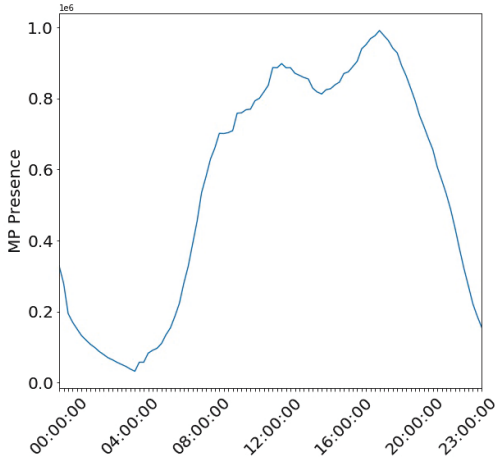

FIGURE 4.10: Total raw mobile phone presence summed over all Greater Paris blocks (IRIS) and averaged per hour.

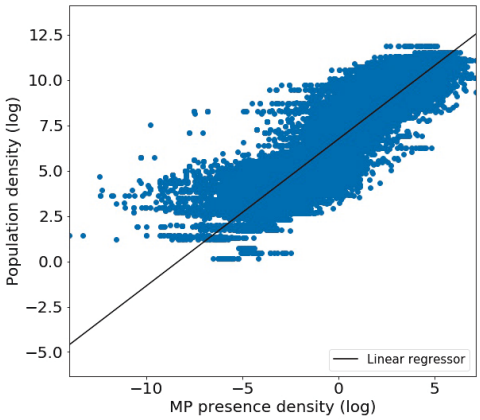

FIGURE 4.11: Linear regression between mobile phone presence at 03 AM and population densities for census blocks at logarithmic scale.

In order to assess the performance of the regression, several cross-validation (CV) strategies are conducted and different number of folds k are tested :

— Classic CV (CV) : census blocks and days are chosen at random. This assesses the model generalization in case of randomly missing data.

— Temporal CV (T-CV) : each fold corresponds to $\frac{1}{k} \cdot 30$ days randomly chosen among a total of thirty days. This assesses the model generalization in case training and predictions correspond to different sets of days.

— Spatial CV (S-CV) : the data is geographically split such as each fold contains $\frac{1}{k}$ blocks. This assesses the model generalization in case training and predictions are performed on different groups of blocks.

TABLE 4.3: Static model performance over a 30 days period and blocks from cluster C_0. Results are shown for training over the complete month and night hours.

$\bar{\alpha}$	$CI_{95}(\alpha)$	$\bar{\beta}$	$CI_{95}(\beta)$	R^2	$NRMSE$
297	$[225, 370]$	0.80	$[0.79, 0.81]$	0.84	0.080

TABLE 4.4: Static model performance over a 30 days period and blocks from cluster C_0. Results are shown for different cross-validation strategies over the complete month and night hours. Both R^2 and $NRMSE$ are averaged over all time-slots.

Strategy	$\hat{R^2}$	$CI_{95}(R^2)$	$\hat{NRMSE}$	$CI_{95}(NRMSE)$
3-CV	0.84	$[0.82, 0.86]$	0.084	$[0.080, 0.088]$
5-CV	0.85	$[0.82, 0.86]$	0.084	$[0.080, 0.088]$
3-T-CV	0.84	$[0.81, 0.86]$	0.085	$[0.080, 0.089]$
5-T-CV	0.84	$[0.81, 0.86]$	0.084	$[0.080, 0.089]$
3-S-CV	0.84	$[0.82, 0.86]$	0.084	$[0.080, 0.088]$
5-S-CV	0.84	$[0.82, 0.86]$	0.084	$[0.080, 0.088]$

Results for training and cross-validation are shown in Tab. 4.3 and Tab. 4.4. For training, the R^2 shows that the proportion of the variance of the predicted population densities explains 84% of the variance of census population densities. For the cross-validation, the obtained R^2 coefficients are above 0.81, showing the goodness of fit of the model. The $NRMSE$ are below 0.09, demonstrating that the errors on predictions are reasonably small. As a comparison, the model performance reported respectively for Rome, Milan and Turin was a R^2 in the range $[0.8, 0.87]$ and a $NRMSE$ in the range $[0.073, 0.1]$ (KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018).

Interpretation

The static coefficients are plotted in function of time in Fig. 4.12. So far the static parameters α and β have not been clearly interpreted in the literature, yet it is a critical condition to understand the relation between mobile phone usage and populations, stated in eq. 2.4. As the β parameter is an exponent, this reveals that the relation between the density of detected cellphones and the population density is not linear across space. In other words, the number of mobile phone per person grows sub-linearly when the urban density increases. In developed countries, we can expect that $\beta \leq 1$, while developing countries could observe a higher β in case of lower penetration rates. In our case β remains close to 0.8. This is consistent with the penetration rate in France being 120% in 2014 (ARCEP, 2014).Note that

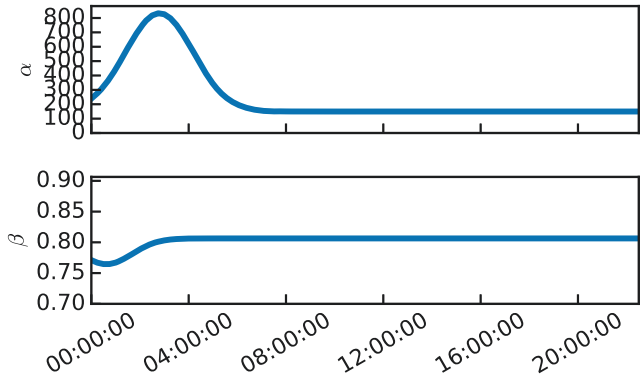

FIGURE 4.12: Distributions of Static parameters smoothed with a Gaussian.

the value of β might also depend on data filtering on the type of mobile phone subscription (i.e., personal mobile, professional or company mobile etc.). For this study, all types of mobile phone subscriptions are selected. Consequently one individual can own more than one mobile phone and reciprocally one mobile phone is equivalent, in average, to ìess' than one person. This could explain why previous studies found a β close to 1 (KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018; DEVILLE et al., 2014). The influence of β can be considered negligible compared to the effect of α ($\beta << \alpha$). In an ideal case scenario, where all mobile phones are turned-on and active anytime, all cellphones are continuously generating geolocated records. Then, α would correspond to the inverse value of the global market share of an area. However in practice the mobile phone activity fluctuates through time and one must account for inactive cellphones. In the theoretical scenario of one unique mobile operator in a region, the market share equals 1. This time, the α coefficient would only be affected by the mobile phone activity intensity given a time of the day. Following our analysis, we interpret the α coefficient as the multiplication of three factors. The first is the market share which is the ratio between the number of subscribers and the total population. The second term is the probability that a mobile phone is turned-on. The third term is the probability that a turned-on device emits a record during a time interval, in other words the probability that any device emits a call, text, updates its data session etc. These second and third terms are higher during the day as mobile devices are more used (see Fig. 4.13). Higher cellphone presences are detected during the day hence requiring a smaller α. Reciprocally, lowest activity rates occur during the night, thus requiring highest α (see Fig. 4.12). The eq. 2.4 is rewritten accordingly.

$$\rho(t) = \alpha(t)\sigma(t)^{\beta(t)} \tag{4.12}$$

$$\alpha(t) = \frac{1}{f} \cdot \frac{1}{p^{ON}(t)} \cdot \frac{1}{p^{record}(t)} \tag{4.13}$$

$$\beta(t) = \frac{1}{\tau(t)} \tag{4.14}$$

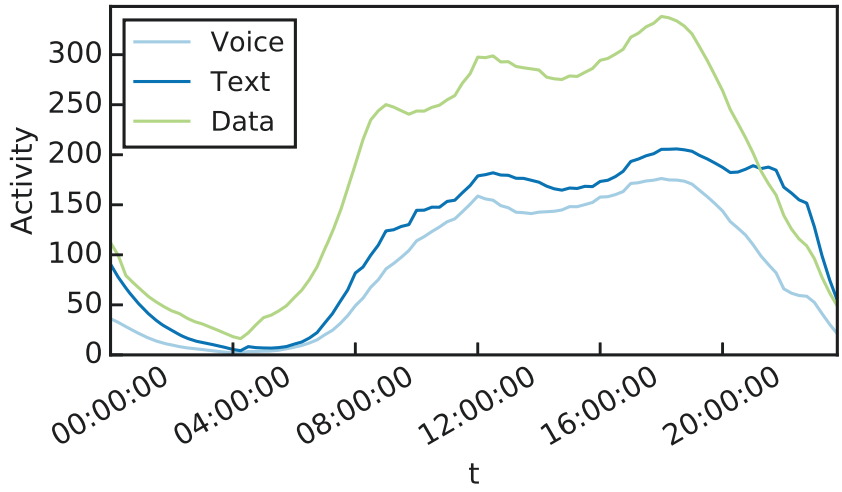

FIGURE 4.13: Mobile phone activity (calls, SMS and data) averaged per hour for all blocks

where $0 < f < 1$ is the market share factor of the case study region during the observation period. Note that the market share can change over long enough time periods (e.g., several weeks) as users may churn or change home location. In addition, $0 < p^{ON}(t) < 1$ is the probability of a device to be turned-on during time-slot t and $0 < p^{record}(t) < 1$ is the probability to detect the phone during t. Here, τ refers to the penetration rate of the region i.e. the number of phones being used over the population. It is unclear if this variable should be considered a constant. By default we initially define β as time varying although its variations are small.
At last we compare our obtained (α,β) values to the one obtained in previous studies. One study reported β being between 0.84 and 0.9, considering all regions in France and individual mobile subscriptions, for a different mobile phone operator (DEVILLE et al., 2014). For our case study the average value for β is 0.8 for the Greater Paris, all mobile subscriptions being considered. The obtained β parameter is consistent with previous findings, stipulating that β is close to 1 or slightly below. However we note a higher variability for α depending on the case study data and the temporal resolution. The order of magnitude for our α, on our case study, is 10^2, for 15 min time-slots. In a previous study (DEVILLE et al., 2014), the reported range for α for France is $[50, 100]$. Still, in such study, the static models were calibrated over the complete night time, i.e., several hours. In addition, studies using Italy data (TIM, 2014) reported α to be between 2 and 4 for cities of Rome, Milan and Turin, with time-slots of 15 min. Here it seems that for the Italian datasets, the α represents the market share term only. This is possibly due to two main factors. First, we applied different initial hypothesis when generating the mobile phone presence. In our study, we do not account a presence when a phone is undetected and we do not assume the current position being the previous one during an anterior time-slot. Meanwhile, the TIM datasets rely on the assumption that a phone remained at the

same previous position by default even when no records have been observed for a long period. This assumption is equivalent to detecting phones anytime i.e., to have $\forall t\ p^{ON}(t) = p^{record}(t) = 1$. Such a pre-processing could introduce some position errors in case of low sampling frequency or long periods of inactivity. The second factor could be a different record frequency, in case the Italian mobile provider have a higher level of passive records or shorter periodic location updates. For our data the default periodicity is 30 minutes in case cellphones receive $4G$ signals. Meanwhile the considered time-slots are 15 minutes. Therefore we never detect all devices during a given time-slot and both $p^{ON}(t)$ and $p^{record}(t)$ are lower than 1.

4.3.4 Dynamic Model

Although the static model has demonstrated good performance, the available ground truth is restricted to census population. Such a model is assumed reliable to rescale mobile phone presence densities only during night time. Following our interpretation of the static parameters, we emit the hypothesis of an underlying relation between the dynamic parameters and the mobile phone activity. Thus, we aim to determine dynamic coefficients that remain valid at any time of the day. The mean cellphone activity, which depends on space and time, has been introduced as the parameter λ, see Eq. 3.7. The mean activity is calculated for each type of record separately and noted λ_{CALLS}, λ_{SMS}, and λ_{DATA}. For our study, we compare two ways of calculating λ. The first version is the mobile activity per user averaged over blocks which was

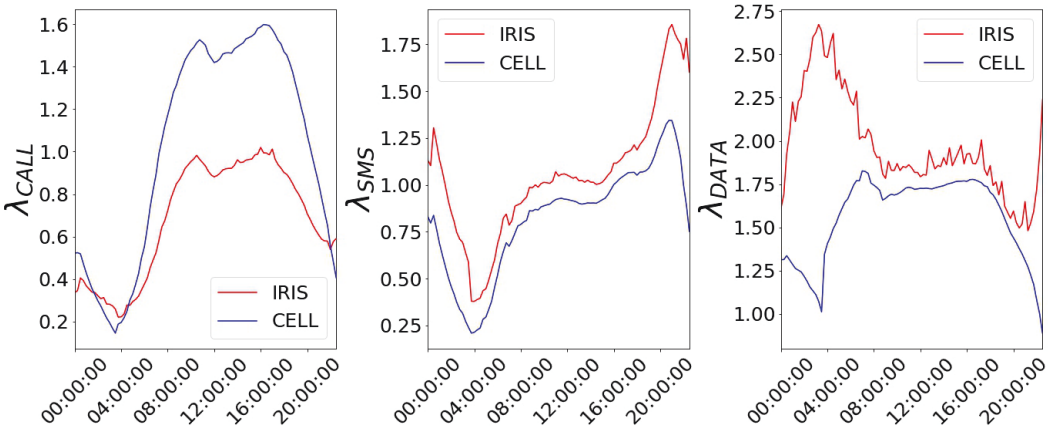

FIGURE 4.14: Evolution of λ over time, with mobile phone activity averaged over cells (blue) VS IRIS blocks (red).

used in (KHODABANDELOU et al., 2016a). Our second version of λ corresponds to the mobile phone activity per user calculated directly at the initial network cell level and averaged over all cells. Several differences can be observed from Fig. 4.14. For calls and texts, the variation of λ is similar, with Pearson correlations equal to

0.98 for calls and 0.87 for texts. Yet the average call activity is lower over blocks compared to cells, with a Median Absolute Percentage Error of 30%. Meanwhile the average text activity is higher over blocks, with a MAPE of 35%. The most striking difference is for the average data activity which has opposite variations during night time, with a negative correlation ($r = -0.28$). For blocks, the variation of the data activity is inconsistent with the expected user behavior to exhibit a progressive reduction of its activity during the night. A probable cause for this phenomenon is that calls and texts are recorded mainly by 2G and 3G cells which have larger radius (10^2 to 10^3 meters). Meanwhile the data sessions are recorded by 4G cells having smaller radius (10^1 to 10^2 meters). In the Greater Paris, $2G$ and $3G$ cells span over a 6 km^2 median area against 0.6 km^2 for $4G$ cells and 0.33 km^2 for blocks. These observations seem to unveil another limitation of the spatial mapping performance at block scale. Consequently, we decide to retain λ calculated over cells. The mean user activity is averaged over mobile network cells rather than census blocks, in order to avoid the areal bias generated by the spatial mapping.
The relationships for (α, λ) and (β,λ) are displayed in Fig. 4.15. The time range considered is $05 - 08$ AM, where the ground truth is assumed reliable. Here, both static parameters α and β exhibit a linear relation with $\frac{1}{\lambda}$.

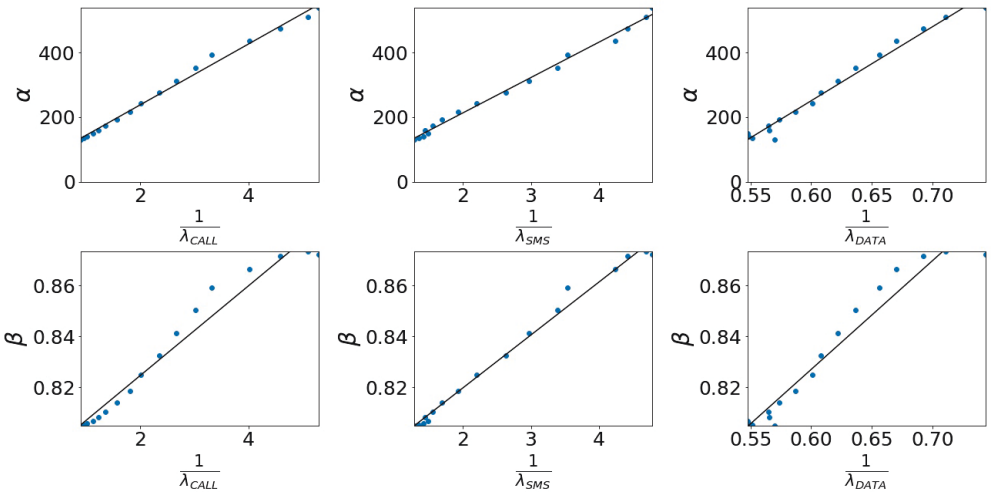

FIGURE 4.15: Linear regression for (α,$\frac{1}{\lambda}$) and (β,$\frac{1}{\lambda}$) for calls, SMS and data, between 05-08AM

The regression parameters are given in Tab. 4.5. For all activity types we observe a good fitting performance, having $R^2 \in [0.94, 0.99]$. Phone calls obtain the highest R^2 and are hence retained to calculate the dynamic scaling factors using λ_{CALLS}. Looking at the regression parameters in Tab. 4.5, it seems that β is relatively invariant to $\frac{1}{\lambda}$ as $a_\beta \approx 0$. Therefore β can be assumed a constant independent of time and of mobile phone activity. Such assumption remains consistent with our interpretation for β and with the small fluctuations shown in Fig. 4.12. The main difference between our result compared to the one of (KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018) is that the authors found a linear relationship

for ($log(\alpha)$, λ) and (β, λ). Meanwhile in our case there is a linear relationship for variable pairs (α,$\frac{1}{\lambda}$) and (β,$\frac{1}{\lambda}$). Again this is possibly due to the different nature of the mobile phone presence data (i.e., different record frequency or different hypothesis for presence pre-processing).
Then our initial static coefficients are modified by introducing the parameter λ to account for the dynamic mobile phone activity. The resulting dynamic parameters are noted ($\alpha(\lambda,t)$,β). Eventually, we add a rescaling term $R(\lambda,t)$ in order to adjust the estimates to the total population in the region.

$$\hat{\rho}(\lambda,t) = R(\lambda,t)\cdot\alpha(\lambda,t)\cdot\sigma(t)^{\beta} \quad (4.15)$$

$$\alpha(\lambda,t) = \left(\hat{a_\alpha}\frac{1}{\lambda(t)} + \hat{b_\alpha}\right) \quad (4.16)$$

$$R(\lambda,t) = \frac{P+\epsilon}{Med_t(\widehat{P}(\lambda,t))} \quad (4.17)$$

where $\widehat{P}(t) = \sum_i \hat{\rho_i}(\lambda,t)\cdot A_i$ is the total dynamic population estimated as the sum of dynamic densities multiplied by block area A. The total number of inhabitants of the region equals P while ϵ is an approximation of the number of visitors per day and can be tuned depending on the case study. The Greater Paris have a 12 million residential population and is a major touristic area (more than 30 million tourists per year for Paris center, about 10 million for Disneyland etc.) therefore we assume ϵ is of the order of a few million for a typical day. The term $R(\lambda,t)$ is broadly equivalent to multiply by some total ground truth and divide by the median value of the predicted population for a time-slot t. Thus this enables to recalibrate the model and prevents from overfitting between 05-08AM. It is believed that $R(\lambda,t)$ can be fine-tuned to the target area in order to match any external ground truth, when available. For our case study, the confidence interval for $R(\lambda,t)$ is $[0.66, 0.91]$. The mean and median values are respectively 0.79 and 0.56. In the next sections we present our results on dynamic population.

TABLE 4.5: Regression coefficients between α and $\frac{1}{\lambda}$ and between β and $\frac{1}{\lambda}$

	α			β		
	$\hat{a_\alpha}$	$\hat{b_\alpha}$	R^2_α	$\hat{a_\beta}$	$\hat{b_\beta}$	R^2_β
CALLS	93.3	49.3	0.99	0.02	0.79	0.96
SMS	108.6	−6.2	0.99	0.02	0.78	0.98
DATA	2275.5	−1117.1	0.98	0.44	0.56	0.94

4.3.5 Analysis of Results at Block Resolution

After estimating the dynamic population at block scale, we assess the impact of the scaling factor $R(\lambda, t)$ on the model performance. Without this parameter the estimated total population ranges between 3 million and 36 million which is not realistic, considering there are 12 million residents. With $R(\lambda, t)$, the estimated Greater Paris population is calibrated to be constant. First, the performance evaluation of the dynamic model at block scale is assessed trough a comparison with Paris. The city-center population is expected to fluctuate during the day while the residential population of the city is the expected lower bound during the night. Considering our dynamic estimates for Paris, the minimal value obtained during night time is 1.55 million against 2.25 million residents. Therefore, the Paris population is underestimated by 31%, at the block level. This result unveils a well-known limitation of traditional models which underestimate populations in denser areas while overestimating in lower density areas (DEVILLE et al., 2014). One way to correct this error is by adding a different rescaling term for the city center ($R_{ctr} > 1$) and for the suburb ($R_{sub} < 1$).

$$\hat{\rho}_*(t) = R_*(\lambda, t) \cdot \alpha(\lambda, t) \cdot \sigma(t)^\beta \tag{4.18}$$

$$R_{ctr} = 1.44 \cdot R(\lambda, t) \tag{4.19}$$

$$R_{sub}(\lambda, t) = \frac{\widehat{P}_{sub}(t) - \widehat{P}_{ctr}(t)}{\widehat{P}_{sub}(t)} \cdot R(\lambda, t) \tag{4.20}$$

where $* \in \{\text{sub}, \text{ctr}\}$, P_{sub} and P_{ctr} are respectively the residential population of the Greater Paris region and of Paris. A second way to overcome this issue could be to train several regression models in order to extract different parameters (α_j, β_j) for different zones j. Still, the optimal boundaries and size of such zones have to be determined, as the spatial scale can impact the model. For instance, different models could be trained for each department, towns, blocks etc. The final estimated

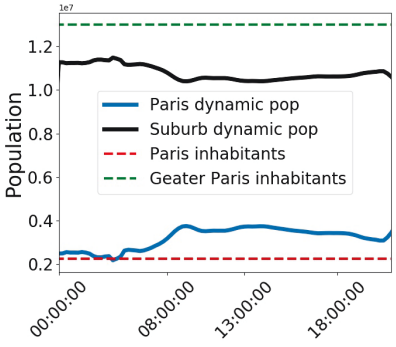

FIGURE 4.16: Final dynamic population in Paris over time, for a typical business day.

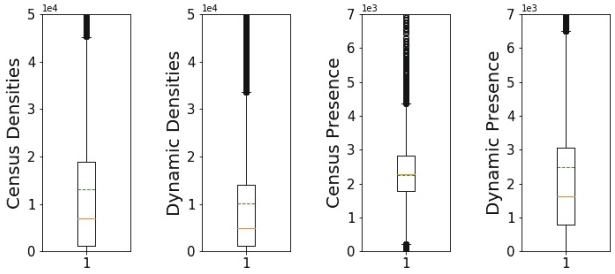

FIGURE 4.17: Boxplots for census vs. estimated population densities and absolute number of visitors for a typical business day.

attendance in Paris and the suburb are represented in Fig. 4.16.
Eventually, our final dynamic estimates at block scale are analyzed and compared with the census densities and absolute population (see Fig. 4.17 and Fig. 4.18). Looking at Fig. 4.17, we observe that the census and dynamic densities exhibit close mean (resp. $1.3 \cdot 10^4$ and $1.0 \cdot 10^4$) and median (resp. $7.0 \cdot 10^3$ and $5.0 \cdot 10^3$) with similar standard deviation (resp. $1.6 \cdot 10^4$ and $1.3 \cdot 10^4$). The main difference is that the range between first and third quantile looks more compressed for the dynamic densities, while the maximal dynamic densities (i.e., top 0.1 %) are much higher than for census. This phenomenon illustrates the movement of populations from numerous suburb blocks toward a smaller number of city-center blocks. Regarding absolute populations, the dynamic estimates spread over a larger range than census population, according to population movements during the day. For a typical

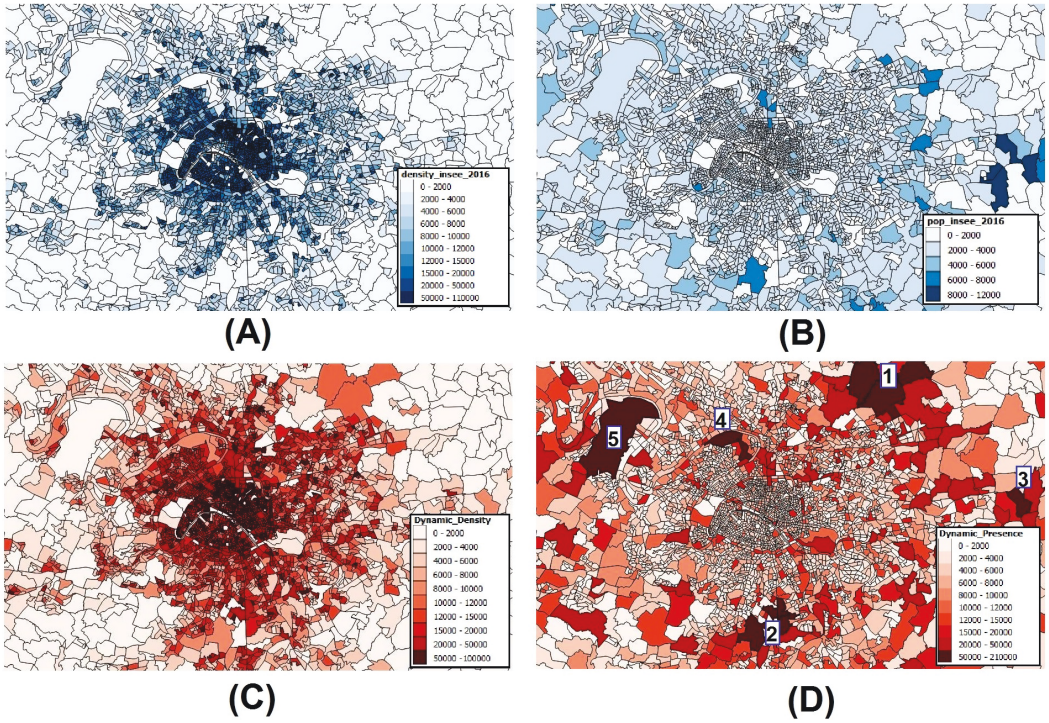

FIGURE 4.18: (A) Population densities INSEE 2016, (B) Absolute population INSEE 2016, (C) Median dynamic population densities and (D) Median absolute population at 5 AM for a typical business day in November. Top visited areas are 1 : Roissy, 2 : Orly, 3 : Disneyland, 4 : Gennevilliers (outlier), 5 : Saint-Germain Forest (outlier).

week day, the denser and most frequented blocks contain the two airports, namely Roissy Charles-de-Gaulle in the North-East corner and Orly located South, and also Disneyland in the East (numbers 1,2 and 3 on Fig. 4.18). Two additional areas, namely Gennevilliers and Saint-Germain also exhibit highest number of visitors (numbers 4 and 5 on Fig. 4.18). The former is an industrial area while the latter contains a forest. These two blocks are huge in size compared to their surrounding blocks therefore they completely drag the population toward them. This

is a consequence of the spatial mapping which depend of area weights and of the initial block scale being highly heterogeneous in block size. Consequently the estimates for these blocks are believed to be anomalies. Eventually, we can visually observe population variations through the day as shown in Fig. 4.19 for densities and Fig. 4.20 for absolute counts. As expected, the city center and the north-west area around la Défense progressively attract workers and visitors in the morning between 5 and 10 AM.

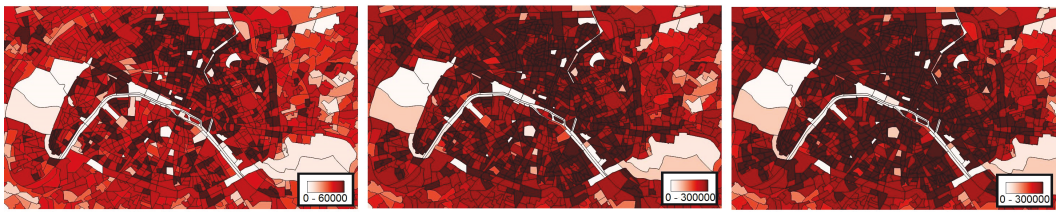

FIGURE 4.19: Dynamic presence densities per km^2 estimated for Paris at 5AM (left), 7AM (middle) and 10AM (right).

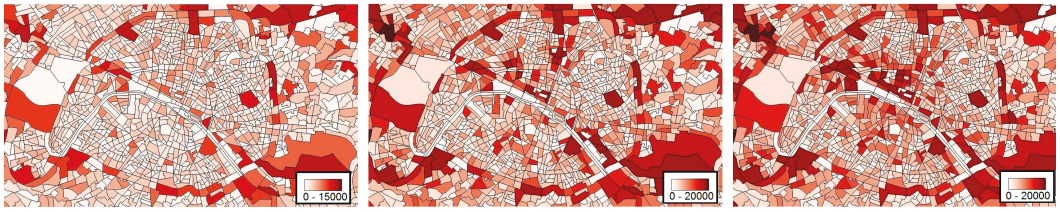

FIGURE 4.20: Absolute dynamic presence estimated for Paris at 5AM (left), 7AM (middle) and 10AM (right).

4.3.6 Day-time Validation on Stadium Attendances

In order to further validate our dynamic model during day time we collect the attendance rates from 9 sport events hosted in two stadiums, namely Stade de France and Parc des Princes. The population counts in the stadiums are estimated with our dynamic model and compared to the ground truth. Both stadiums benefit from indoor coverage as several base transceiver stations are built inside. The mobile phone presence is calculated at the cellular resolution. The validation steps are summarized as follows :

— Select the indoor cells in the stadium, noted s.

— Calculate the mobile phone presence density for each time slot t :
$\forall$t $\sigma_s(t) = \frac{1}{A_s} \sum_j V_j(t)$, where V_j is the number of unique cellphone present in cell j. A_s is the area of the stadium which is the real total coverage area.

— The attendance density is dynamically rescaled :

$$\rho_s(t) = \alpha(\lambda, t)(\sigma_s - \nu_s)(t)^\beta \tag{4.21}$$

where $\nu_s(t)$ is the median value of the cellphone density, during interval t. The later is calculated during the off-game period. This quantity is retrieved in order to avoid considering regular visitors, such as workers, as suggested by (KHODABANDELOU et al., 2016b). As the stadiums cells are indoor, the ν_s values are relatively small in our case (less than 10 individuals at night and less than 100 during the day). In order to assess the performance of stadiums dynamic estimates, we initially set $R = 1$ to compare the performance with and without this parameter.

— The total attendance is obtained after multiplying the attendance density by the stadium area : $V_s(t) = A_s.\rho_s$

The attendance is estimated during each 15 min time slot. The maximal and median attendance is calculated by considering the period between the beginning and the end of each sport game, see Tab. 4.6. In both case the obtained attendance estimates are close to the ground truth according to NRMSE and MAPE (see Tab. 4.7). Overall, the MAPE is less than 11% compared to the real attendance. A slightly different performance can be observed between international and national games, for which the MAPE differs from 2-3%. International events host international visitors

TABLE 4.6: Results on stadiums attendances. Variables MP_{Med} and MP_{Max} are resp. the median and maximal raw mobile phone counts in stadiums. The median and maximal predictions are resp. $\hat{P}_{Med}$ and $\hat{P}_{Max}$. The ground truth is P_{true}.

Day	T_{start}	International	MP_{Med}	MP_{Max}	$\hat{P}_{Med}$	$\hat{P}_{Max}$	P_{true}
06/11	08 :45 pm	NO	2097	3242	39289	46734	42002
11/11	09 :00 pm	YES	3307	5135	61301	79374	78000
19/11	05 :00 pm	NO	2474	3859	30266	42754	44258
19/11	09 :00 pm	YES	2871	4412	65463	79558	73700
26/11	09 :00 pm	YES	3620	5824	81952	96622	78500
30/11	09 :00 pm	NO	1996	2827	39282	44235	40597
06/12	08 :45 pm	NO	2743	3658	40155	46387	42650
11/12	08 :45 pm	NO	3662	4466	46152	59302	47665
14/12	09 :00 pm	NO	2788	3314	40544	47305	45183

TABLE 4.7: Scores between true and estimated attendances

	NRMSE		MAPE (%)	
	$\hat{P}_{Max}, P_{true}$	$\hat{P}_{Med}, P_{true}$	$\hat{P}_{Max}, P_{true}$	$\hat{P}_{Med}, P_{true}$
ALL	0.143	0.148	10.5	10.8
National	0.126	0.143	7.1	10.1
Internat	0.157	0.142	10.9	12.3

which have cellphones on roaming. Therefore, the bias induced by higher roaming rates can locally alter the market share. This factor could explain the slightly lower performance for international events. The maximum attendance predictions obtain a higher performance than the median attendance. This difference is not significant

overall, but is increased for national events for which there is a 3% gap in MAPE. The lowest error is 7% for national games. As sport games are particular events, the mobile phone usage pattern is potentially different (e.g., goals, inter-game period etc.), which could explain the subsiding error. From our results on sport games, the maximum attendance prediction better approximates the ground truth.
For the stadiums attendances, we observe that the rescaling term $R(\lambda, t)$ is not necessary to rescale the estimates and would even degrade the performance. For the Parc des Princes, which is at Paris border, multiplying the attendances by R_{ctr} results in overestimating the attendance by 30%. Similarly, for the Stade de France, which is the suburb, the attendances are underestimated by 30%. It seems that rescaling with $R(\lambda, t)$ is adapted to rescale estimates at the block resolution. For particular events such as sport games, this term can be tuned to better approximate the available ground truth being the real attendance.
This validation study asserts the validity of the dynamic model to estimate population counts in indoor environments. Although the dynamic parameters $(\alpha(\lambda, t), \beta(\lambda, t))$ are initially derived at the census blocks resolution, they remain valid to rescale the raw number of cellphones detected in stadium cells. Still it is unclear whether the dynamic parameters are valid for any spatial resolution, e.g. smaller areas. Additional ground truth is required to extend model validation to other areas, but such data is not accessible.

4.3.7 Comparative Evaluation

At last, we compare our results with three state-of-the-art approaches (DEVILLE et al., 2014; F. XU et al., 2016; KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018). There exist fundamental differences between our data and spatial resolution compared to the ones in previous approaches.
First, (KHODABANDELOU, GAUTHIER, M. FIORE et al., 2018) used datasets having a grid representation of non-overlapping mobile network areas (TIM, 2014). To avoid any confusion, we refer to such non-overlapping areas as cellular zones while the overlapping signal coverage are referred as cells. A cellular zone is believed to be an aggregate of several cells belonging to the same base stations. Instead of classic voronoi polygons, a grid representation is provided. The census population was down-scaled with AW mapping to match the cellular zones. Meanwhile we have performed an opposite rescaling scheme as we have down-scaled the raw mobile network presence from cells to census blocks. Our motivation is to produce fine-grained dynamic population estimates at a representative administrative scale. Their validation study calculates the attendance considering the zones intersecting stadiums in Milan, Rome and Turin. For our validation study, the census blocks scale results in a large underestimation of the attendances (by a factor 10^2). This finding is caused by an erroneous representation of the shapes of mobile network cells

covering the stadium. Theoretically, mobile network cells are represented by circular areas. In the case of indoor base stations, such representation is erroneous, as the signal is limited by physical boundaries (i.e., walls). With the default cells areas, a substantial part of attendees is mapped not only on the intersecting blocks but over additional blocks surrounding the stadium. Consequently we have switched back to the cell resolution to estimate stadium attendances and considered the stadium areas as the real coverage areas. Despite different data, case study and spatial scales, our dynamic model follows a similar methodology, with several adjustments of parameters. Their validation study is conducted on national football games for the stadiums of Milan, Turin and Rome. Therefore we also retain national games. The authors reported a NRMSE of 0.102 while our NRMSE is 0.126. In addition, the MAPE for Italy is 11.9% against 7% in our case. By reason of different observation periods and case studies, our results cannot be directly evaluated against. Still, our respective models grant satisfying estimates with relatively small errors.

Second, we compare our dynamic model with the one applied by (DEVILLE et al., 2014). The initial scale of their data is voronoi polygons generated per base station. Then the mobile phone presence is mapped over administrative areas ('Communes' for France) using the area weights. The dynamic presence is computed by adding a rescaling term $R = \frac{P}{\widehat{P}}$ to the static model : $\rho = R \cdot \alpha\sigma^{\beta}$. The first difference between their approach and ours is that the static model is trained considering all areas, without any data filtering over space. The second difference is that one couple of parameters is generated over the complete night-time, instead of several time intervals. As their study lacks a validation during day time we reproduce their model and assess the performance on stadium attendances against our estimates. The approach of Deville et al reaches a MAPE of 20.5% and a NRMSE of 0.30. In comparison our model grants a 50% gain in performance overall. On national games the MAPE is 13.5% which is still nearly 50% higher. Yet the difference is less significant for the NRMSE being 0.146.

Third, the approach of Xu et al (F. XU et al., 2016) shares similarities with the previous one (DEVILLE et al., 2014). The dynamic population is calculated as $\rho = R \cdot \alpha_j \sigma_j^{\beta}$. The difference of the approach is that several static models are trained for each land-use region j. Although our land-use regions are not generated with the same technique, we reproduce their approach by training a different model for each of our activity region. The second main difference of their study is that the static parameters are extracted at 7AM, for which they reported highest correlation with census population. Although their model was assessed through positive correlations with taxi flows, no validation was conducted to assess the performance of dynamic estimates against some real attendances. Therefore we reproduce their approach on our data. First, we select the time-slot for which the correlation between mobile phone presence density and population is the highest, which is achieved at 9PM, with $r = 0.72$. Then, a static model is trained on each cluster of blocks. Each stadium belongs to a different cluster, therefore the static coefficient are retrieved

accordingly. The factor $R = \frac{P}{\widehat{P}(t)}$ is calculated such as $\widehat{P}(t)$ matches the start time of the sport games. Finally we compare the results on stadiums. This leads to a poor performance, with MAPE above 65% and NRMSE of 0.75. Pearson correlations are calculated before and after the dynamic rescaling. The correlation between stadium raw mobile phone presence and real attendance initially equals 0.87 and drops to 0.36 after the dynamic rescaling using Xu et al method. With Deville et al approach the correlation is unchanged as $r = 0.87$. Meanwhile, with our dynamic model, the correlation reaches 0.96. Consequently, training several static models on several land-use does not improve the model and training should be performed solely on the residential areas.
Although additional ground truth should be used in order to extend the comparative evaluation on validation, our estimates tend to generally outperform past methods.

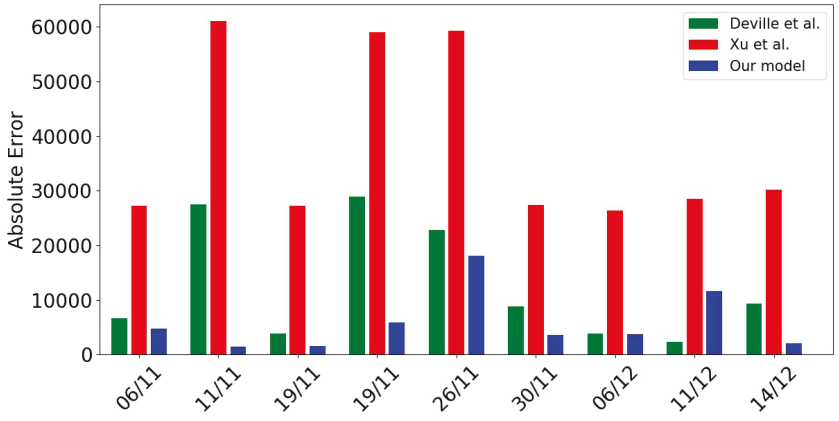

FIGURE 4.21: Absolute Errors for stadiums attendances estimated with our model and two state-of-the-art approaches

4.4 Discussion

From the results on the case study of the Greater Paris, several conclusions are suggested. First, we observe that the case study area impacts the results, as the heterogeneity of the population density increases with the size of the region. In the process of identifying the activity types of blocks, we find that the Median Week Signature of blocks should be normalized by the total activity of the day of week rather than with standard scaler, in order to find clusters in both city center and suburb for any k.
Second, the model performance is highly affected by the spatial scale and by the weights of the spatial mapping. Although the state-of-the-art AW mapping grants a good performance at coarser scales, such as Communes, our corrected APW mapping reduces the errors at the block resolution by 19% from increasing the weights of denser blocks. Still, by reason of unclear cells boundaries, a bias subsists when mapping cellphones to blocks. Down-scaling the coarse cellphone positions on the

mobile network to the fine-grained block scale using spatial mapping systematically introduces errors. One could overcome this issue by accessing a higher level of mobile network geolocation (e.g., passive records, triangulation), in case available and authorized. However, for the purpose of large-scale population densities, obtaining a higher frequency and precision from the raw cellphone geolocation requires important storage resources and increases computation cost.

Third, from the analyze of the relationship between mobile phone density and census population density, we provide a novel interpretation of the static model parameters. The β parameter seems to reflect the inverse cellphone share per individual, growing sub-linearly with the densities. The α parameter is the inverse product of the market share and the probability of detecting a mobile phone activity from a turned-on device. The census population used to calibrate the static model is assumed a valid ground truth during night time yet it is not representative of day time populations. During early morning hours, we have observed that the static parameters are inversely proportional to the mean activity parameter λ. Still, as β exhibit small variations, this parameter is set to a constant. The dynamic variable $\alpha(\lambda, t)$ allows to extrapolate the model to account for the fluctuations of the population. In addition we have introduced the parameter $R(\lambda, t)$ to rescale the dynamic population to the total population of the region. This parameter enables to calibrate the model despite the lack of ground truth during the day. A classic limitation of dynamic population models trained on large regions with heterogeneous densities, is an overestimated population in low-density areas while underestimating dense areas. Therefore we propose to correct the model by setting a different $R(\lambda, t)$ for the city-center and the suburb. Although such factor corrects the issue for the total population, it rescales uniformly all blocks, which might produce errors at the block resolution. For future works we recommend to train several models on subdivisions of the region to overcome the performance gap between city center and suburb. For instance, one model should focus on residential areas in the city center while a second model generates parameters for the suburb.

Fourth, our dynamic model is validated against sport game attendances in two stadiums, for which we obtain an absolute error of 10%. This finding reveals that although our dynamic model is primarily designed to estimate populations at the block scale, the dynamic parameters can be used to estimate visitors rates using the cellular scale in indoor environments. Our comparative evaluation with past studies reveals that our model outperforms the state-of-the-art regarding our validation study.

Still, several open issues subsist. First, the estimation of accurate visitor rates in outdoor areas remains an open challenge by reason of the lack of day-time ground truth for fine-grained calibration. Second, there is an important spatio-temporal variability of model parameters, i.e. parameters are different for other regions/cities and different time periods (i.e., season, week vs week-end, holidays etc.). Consequently several models are necessary to find the appropriate parameters for each

region, each season, etc. Third, despite good results on sport games, the later are particular events which are not representative of the broad spectrum of human activities. It is important to assess model performance on other types of particular events and specific types of populations (e.g., tourists etc.). In order to strengthen model validation and fine tune parameters, substantial ground truth on populations is necessary yet hardly accessible.
Eventually, our results on dynamic population estimation offer promising perspectives for urban planning. Our work can be useful to derive numbers of commuters in transportation hubs (e.g., train stations), number of visitors in shopping centers or in any areas that benefit from good indoor coverage. A second application for our work is anomalous event detection which could be revealed by abnormal variations in population counts.
In conclusion, we have demonstrated that mobile network data can efficiently capture the daily variations of urban populations. Our work has revealed that the choice of spatial resolution highly impacts the performance of the population estimates. When processing Call Detail Records, the best performance is obtained at the postcode resolution (e.g., town or Commune). Meanwhile, down-scaling the raw mobile phone counts to areas smaller than mobile network cells introduces positioning errors. Coarser scales are thus a safer choice to study urban mobility in outdoor spaces. Despite a moderate performance at the census block level, spatial mapping are still useful to map raw cellphone positions to well known urban areas, for which we look to estimate the number of visitors over time. Following our findings, we decide to build our next model at the postcode scale resolution (Commune).

4.5 Summary

In this chapter, we have presented an approach for the estimation of population densities and number of visitors at fine-grained spatio-temporal resolution, using mobile network data mining. Currently, urban planing authorities essentially rely on census and surveys to obtain information on population mobility behavior. Still the update frequency of such traditional data is, at best, annual-based. This work is motivated by the need for daily large-scale population insights, in order to analyze the daily visit frequency in urban areas. The presented model relies on two data sources, namely mobile network records and census. The spatio-temporal resolution of the study is 15 min time-slots and the census block scale (IRIS), being the smallest administrative subdivisions. During night time, mobile phone presence densities from residential blocks are calibrated with census population densities (static model) for each time-slots. The parameters are thus corrected to account for the variations of the mobile activity intensity during day time. Compared to the state-of-the-art, we perform several adjustments of the model and develop its interpretation. First,

for the spatial mapping of mobile phone presence, we introduce an additional weight accounting for blocks densities. Such mapping results in an increase of the performance of 19% at the block scale. Second, in order to filter non-residential blocks from the training set , we discover that the data normalization has a crucial impact on the clustering results. Third, an explicit definition of model parameters α and β is provided. Then, the dynamic population is calibrated such as the city center has its total census population being a lower-bound during night time, while the total population of the region is assumed constant through the day. Our validation study with sport attendances assesses the validity of our dynamic rescaling parameters for both block scale and cellular scale resolutions. To strengthen model validation, we encourage the use of additional data for future works.

5 Origin Destination Matrices by Transport Mode

5.1 Introduction

The expansion of urban population generates rising travel flows, increasing the need for efficient transport planning policies. Traditional transport planning models, such as four steps and activity based models, extensively rely on travel surveys (McNally, 2000 ; Bhat et Koppelman, 1999). However, surveys are constrained by their low-update frequency by reason of their important cost and may be prone to a sampling bias. In addition, surveys generally report one day of trips per individual, which is not sufficient to capture all the temporal variations of trips (e.g., seasonality, weekly patterns). In recent years, public transport operators have been collecting daily travel-card information (Pelletier et al., 2011 ; X. Ma et al., 2013 ; Munizaga et Palma, 2012). In most urban areas, multiple transport operators are in charge of public transport services and infrastructures. Each operator possesses travel data on its own transport network. Therefore, transport operators usually lack a global picture of the traffic state in the multimodal transport network. Such knowledge could be a valuable asset to evaluate and predict the impact of perturbations (e.g., congestion, public transport interruption, public transport strikes, road closure, meteorological events). In parallel, the pervasive use of mobile phones, along with their high penetration rates, have made mobile phone data the largest mobility data source. Call Detail Records (CDR) are collected at no additional cost by telecommunications operators for billing purposes. Several research works have described the potential of mobile phone data for mobility analysis (Chen, J. Ma et al., 2016 ; Gadziński, 2018 ; Blondel et al., 2015). Travel demand modeling (Toole et al., 2015 ; M.-H. Wang et al., 2013), itinerary reconstruction (Asgari et al., 2016 ; Becker et al., 2011), traveler behavior understanding (Calabrese, Diao et al., 2013 ; Yihong Wang et al., 2018 ; Ahas et al., 2010), population density estimation (Bachir, Gauthier et al., 2017 ; Khodabandelou et al., 2016a), transport mode detection (H. Wang et al., 2010 ; Bachir, Khodabandelou, Gauthier, El Yacoubi et Vachon, 2018), traffic state estimation (Demissie et al., 2013 ; Dong et al., 2015), passenger flow estimation (Zhong et al., 2017), anomaly detection (Pang et al., 2013), mobility and activity patterns extraction (Jiang, Ferreira et al., 2017 ; Chen, Bian et al., 2014) are among the most popular research areas. Mobile

phone data offer the possibility to build daily, or even hourly, Origin-Destination (OD) matrices of flows (Çolak et al., 2015a ; Iqbal et al., 2014 ; Alexander et al., 2015 ; Toole et al., 2015 ; H. Wang et al., 2010 ; Berlingerio et al., 2013b ; Di Lorenzo et al., 2016 ; Ni et al., 2018 ; Aguiléra et al., 2014 ; Calabrese, Di Lorenzo et al., 2011). Therefore, such data represent an inexpensive and up-to-date supplement to travel surveys and provide large-scale multimodal mobility information to complement data collected from travel-cards. Still, mining meaningful mobility insights from mobile phone geolocation raises new technical challenges such as computational efficiency, data processing, integration, evaluation, validation and user privacy.

The following chapter is an extension of two articles (Bachir, Khodabandelou, Gauthier, El Yacoubi et Vachon, 2018 ; Bachir, Khodabandelou, Gauthier, El Yacoubi et Puchinger, 2019). In this work, we present the first approach for the estimation of Origin-Destination matrices per transport mode using mobile network data from the Greater Paris region. In this perspective, we design an end-to-end model for road and rail passenger flows estimation using mobile network records. Hundreds of millions mobile phone trajectories from the Greater Paris region are collected for this study. Data pre-processing and transport mode inference are detailed in Section 5.2. The transport mode of trajectories is inferred through a two-steps model, such as the first step is semi-supervised and the second step is unsupervised. The model separates the trips into rail or road mode. Each trajectory is represented as a sequence of visited locations on the mobile network. During the first step, a clustering algorithm is applied to mobile network locations to determine their transport land-use. In the second step, we compute the Bayesian transport probabilities associated to trajectories. The OD matrices of flows are thus generated for both transport modes. As the number of mobile phones is limited by the operators' market share, mobile phone flows are rescaled to the total population using expansion factors, based on census and survey data. In Section 5.3 we summarize our main results. Finally, we perform a validation study in Section 5.4. Our estimates are validated against two external data sources : household travel survey and travel-card data.

5.2 Method

5.2.1 Overview

In this section, we present our method for travel flows estimation using mobile network data. Data processing steps are detailed in Section 5.2.2, 5.2.3, 5.2.4 and 5.2.5. The transport mode inference model is described in Section 5.2.6 and 5.2.7. In this study, we perform a bi-modal separation between road and rail trips to infer passenger flows. The mode inference is two-fold : we first perform clustering on mobile network sectors followed by a the inference of trajectory transport probabilities. The essence of the mode inference is that the mobile network trajectories are decomposed in order to learn the most probable transport mode from each record without the need of the complete real itinerary. Each time an event is recorded, one knows the location of a device on mobile network sectors. In the clustering phase, clusters of sectors grouped by transport use are generated. This step is equivalent to a transport land-use partitioning of the mobile network. Using a small labeled subset of sectors (e.g., antennas inside train stations, highways etc.) we derive transport mode probabilities per cluster. A transport mode probability is assigned to each sector, depending on its cluster. Then, we compute the bayesian probabilities of each anonymized trajectory. The prior transport mode probability is derived from the travel survey and each newly observed record updates the prior. For each trajectory, the posterior probability is computed and the mode with highest probability is retained. Once transport modes are obtained, we construct modal OD matrices of flows, rescaled to the total population (see Section 5.2.8). To evaluate the model performance we use several evaluation metrics, presented in Section 5.2.9. The model workflow is shown in Fig. 5.1.

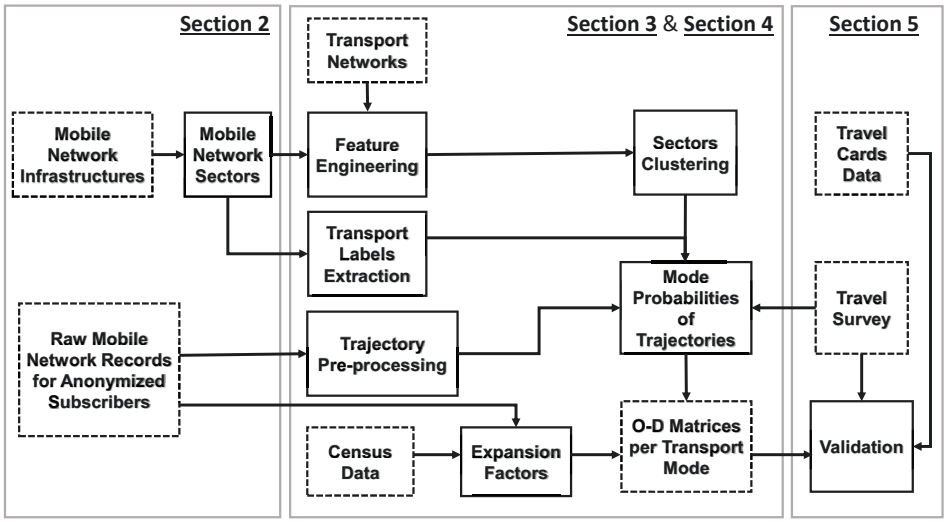

FIGURE 5.1: Workflow of the model for construction of OD matrices per transport mode

5.2.2 Trajectory Pre-Processing

In a previous chapter we describe the mobile network spatial scale (Sec. 3.2.2) and the general data pre-processing in details (Sec. 3.2.4). The mobile phone provider applies a smoothing technique to raw trajectories for noise reduction. Then stay points and moving points are identified. The resulting trajectories are the sequences of cellular locations, such as cellphones are moving with significant speed, according to some speed threshold, while not remaining in a stay position longer than a temporal threshold. For this study, each record is associated to a sector position on the mobile network, corresponding to a non-overlapping subdivision by base station and signal orientation.
For a moving device u, a trajectory is defined as a sequence of visited sectors locations : $T_j^u = \{(S_0, t_0), ..., (S_l, t_l)\}$, where j is the trajectory index, (S, t) is the position recorded at timestamp t and $S = (x, y)$ are the centroid coordinates of the visited sector. For this study, 360 millions trajectories are constructed from 2.4 million anonymized mobile phones during three months. Trajectories with at least 2 distinct moving positions are retained, since a single moving point could be noise. When comparing results with the household travel survey we select users living in the Greater Paris region using their home department (first two digits of their billing address postcode).

5.2.3 Feature Construction with Transport Networks

A trajectory is a sequence of sector locations for which we aim to find transport mode probabilities. Our idea is to first associate a transport land-use to each mobile network sector. In this perspective, we construct sector features based on related spatial information between mobile networks and transport network infrastructures. The road networks infrastructures are collected from OpenStreetMap (OSM, 2018). To reduce the computational cost, we apply a filtering on the highway type. We filter out residential type roads and only retain major road types (motorways, primary, secondary, tertiary). The rail infrastructures are retrieved for underground, overground, tramway and train stations from the STIF Open Data platform (STIF, 2018). In addition, high-speed rails are collected from OSM. The following sector features are then constructed :

— $d_{j,rail}$: shortest euclidean distance between the centroid of sector j and the rail network.

— $d_{j,station}$: euclidean distance between the centroid of sector j and the centroid of the closest train station.

— $N_{j,road}$: number of roads intersecting sector j.

— $N_{j,rail}$: number of rail lines intersecting sector j.

— $A_{j,station}$: area of train stations calculated as the sum of train stations areas intersecting sector j, such as $A_{j,station} = \sum_i A_{i \cap j}$, where i is a train station.

5.2.4 Feature Normalization

Feature normalization is a critical step of a learning model which can highly impact the performance. Although several well-established normalization procedures are used as default normalization strategies by most practitioners, the appropriate normalization solely depends on the dataset at hand. Several strategies can be considered to normalize our features. Common normalization techniques are the standard scaler and the MaxMin scaler.
Still, in the previous chapter we notice that such classic normalization fails to cope with highly heterogeneous urban areas. For the present case study, densities of transport networks and mobile networks are both heterogeneous. The city center benefits from a higher base station concentration with smaller sectors and denser transport networks. On the contrary suburbs have coarser sectors with transport infrastructures separated by longer distances. We normalize our features to reduce the bias induced by urban density over transport usage.

$$\widehat{d}_{j,m} = \frac{d_{j,m}}{\sum_i d_{j,i}} \in [0,1] \tag{5.1}$$

$$\widehat{N}_{j,m} = \frac{N_{j,m}}{\sum_i N_{j,i}} \in [0,1] \tag{5.2}$$

$$\widehat{A}_{j,station} = \frac{A_{j,station}}{A_j} \in [0,1] \tag{5.3}$$

where $d_{j,m} \in \{d_{j,road}, d_{j,rail}, d_{j,station}\}$ and $N_{j,m} \in \{N_{j,road}, N_{j,rail}\}$. The normalized features are noted $\widehat{d}_{j,m}$, $\widehat{N}_{j,m}$ and $\widehat{A}_{j,station}$, resulting from the normalization of features $d_{j,m}$, $N_{j,m}$ and $A_{j,station}$.

5.2.5 Label Extraction

A few base stations are built on transport network infrastructures, such as rail lines, train stations, highways or tunnels. This information is processed to construct labels for a small subset of base stations. For equipments located inside the underground and train stations, which represent 4% sectors, we attribute rail labels. For antennas inside tunnels (1% sectors) and highway antennas (11% sectors), we assign the road mode. In total we obtain 15% transport labels for sectors, represented in Fig. 5.2. Initially, we use categorical transport labels $\{road, rail\}$ on our subset of sectors. Still, categorical transport labels are not appropriate for most sectors, such as outdoor equipments. Indeed, in urban areas such as the Greater Paris, the classic scenario is to encounter several transport modes inside an outdoor

sector because of the mobile networks' coarse granularity. Equipments which are constructed at the border of roads or rail are not guaranteed to exclusively detect a single transport mode. When several transport routes are present in a sector, users could have taken any mode present in this sector. Yet sectors may have a dominant mode. Thus, we aim to find continuous transport probabilities $P \in [0, 1]$ for all sectors, based on the prior knowledge of the categorical transport labeled subset (see Section 5.2.6). For our case study area, maximal transport probabilities $P \in \{0, 1\}$ are restricted to indoor labeled sectors for base stations built inside the underground or tunnels.

FIGURE 5.2: The 15% labeled sectors projected on the Greater Paris area (1) with a zoom on Paris (2).

5.2.6 Mobile Network Sectors Clustering

In order to find groups of sectors with similar transport usage we use a clustering algorithm. Our goal is to find transport clusters of mobile network sectors with an underlying hierarchical structure, such as the highest hierarchy (for $k = 2$) ideally separates rail from road sectors. Consequently we apply the agglomerative hierarchical clustering (AHC). Three linkage types (ward, complete, average) and three distance functions (euclidean, manhattan and correlation) are tested. In addition we compare the performance of the AHC with K-means, which works with euclidean distance, and DBSCAN, tested with the aforementioned three distance metrics. The performance and the optimal number of cluster are determined by evaluating jointly the Silhouette score S, the Calinski-Harabasz CH index and the S_{dbw} validity index.

Then, for each cluster k we calculate the score $p_{k,m}$ of a given transport mode $m \in \{rail, road\}$.

$$p_{k,m} = \frac{L_{k,m}}{L_m} \tag{5.4}$$

$$P(m|S_{i,k}) = \frac{p_{k,m}}{\sum_j p_{k,j}} \tag{5.5}$$

where $L_{k,m}$ is the number of sectors having a mode label m in cluster k and L_m is the total number of sectors with label m in the dataset. $P(m|S_{i,k}) \in [0, 1]$ is the probability of using mode m given that users have visited a sector $S_{i,k}$ belonging to a cluster k. The probabilities satisfy the condition : $\sum_m P(m|S_{i,k}) = 1$. Unlabeled sectors obtain transport probabilities according to their cluster. In addition we update the probabilities of outdoor labeled sectors (i.e., rails and highways) using Eq. 5.5. Indoor labeled sectors have maximum (or minimum) transport probabilities in $\{0, 1\}$.

5.2.7 Inference of Trajectory Transport Mode

Bayes probabilities are used to determine the main transport mode associated to a mobile phone trajectory. The probability $P(m|T_j^u)$ to take a mode $m \in \{rail, road\}$ knowing the trajectory T_j^u is computed for each mobile phone trajectory. Trajectories are sequences of sectors $\{S_0, ..., S_l\}$ visited by mobile phone holders. Therefore, $P(T_j^u|m) = P(S_0, ..., S_l|m)$. In order to compute the term $P(S_0, ..., S_l|m)$, being the joint probabilities of visiting the sectors given the mode, we need to determine whether such probabilities are dependent or independent. In case of dependence we have :

$$\forall i \quad P(S_i, S_{i+1}|m) = P(S_{i+1}, S_i|m) \cdot P(S_i|m) \tag{5.6}$$

Meanwhile, the independence assumption gives :

$$\forall i \quad P(S_i, S_{i+1}|m) = P(S_i|m) \cdot P(S_{i+1}|m) \tag{5.7}$$

Consequently, we calculate :

$$\Delta(S_i \perp S_{i+1}) = |P(S_i, S_{i+1}|m) - P(S_i|m) \cdot P(S_{i+1}|m)| \tag{5.8}$$

The median and mean are respectively $2.1 \cdot 10^{-8}$ and $1.87 \cdot 10^{-7}$ with a standard deviation equal to $3.83 \cdot 10^{-6}$.
In addition, if two variables are independent then the covariance and correlation are null. Reciprocally if the covariance and correlation are not equal to zero, there is dependence. As a result $cov(P(S_i, S_{i+1}|m), P(S_i|m) \cdot P(S_{i+1}|m)) = 10^{-14}$ which tends to zero. Still, $corr(P(S_i, S_{i+1}|m), P(S_i|m) \cdot P(S_{i+1}|m)) = 0.17$. Despite non null values, both covariance and correlation remain small. Consequently we have a

weak dependence between sectors probabilities. In order to speed up the computation time of the probabilities of the millions of cellphones trajectories, we assume in what follows that Eq. 5.7 hold true. This assumption is assessed in the next sections on performance evaluation and validation. Thus, we have $P(T_j^u|m) = \prod_{i=0}^{l_j} P(S_i|m)$. The Bayes theorem is then recursively applied.

$$P(m|T_j^u) = \frac{P(T_j^u|m) * P(m)}{P(T_j^u)} = \frac{P(m)}{P(T_j^u)} \prod_{i=0}^{l_j} P(S_i|m) \tag{5.9}$$

Using Eq.5.5 we inject $P(m|S_i)$ to Eq. 5.9 :

$$P(m|T_j^u) = \frac{\prod_{i=0}^{l_j} P(S_i)}{P(T_j^u)} P(m)^{1-l} \prod_{i=0}^{l_j} P(m|S_i) \tag{5.10}$$

where l_j is the length of trajectory T_j^u. The term $\frac{\prod_{i=0}^{l_j} P(S_i)}{P(T_j^u)}$ does not influence the mode choice. The prior transport probability $P(m)$ can be seen as the initial guess for the distribution, before observing records. The prior is obtained from the travel survey and depends on users' home locations. From the survey, we calculate average trip counts per user to obtain the prior for each department D_i. The prior for rail mode can be rewritten as : $P(rail) = P(rail, D_i) = \frac{C_{rail}^{TS}(D_i)}{C_{rail}^{TS}(D_i)+C_{road}^{TS}(D_i)} \in [0, 1]$ and $P(rail, D_i) = 1 - P(road, D_i)$, where $C_{rail}^{TS}(D_i)$ and $C_{road}^{TS}(D_i)$ are the average rail and road trip counts in the travel survey (TS) for individuals living in department D_i.

Input: List of transport modes $m \in \{rail, road\}$;
A trajectory $T^u = \{S_0, ..., S_l\}$ for mobile phone u;
Survey transport probability $P(m)$ given the home location of u ;
Output: Transport probabilities $P(m|T^u)$;
Dominant transport mode m for T^u ;
foreach m **do**
  foreach $S_i \in T^u$ **do** get $P(m|S_i)$;
  Calculate joint sectors probabilities ;
  $P(m|T^u) \leftarrow \prod_{i=0}^{l} P(m|S_i)$;
  Update the trajectory probability ;
  $P(m|T^u) \leftarrow P(m)^{1-l}.P(m|T^u)$;
  Normalization ;
  $P(m|T^u) \leftarrow \frac{P(m|T^u)}{\sum_i P(m_i|T^u)}$;
end
m* = $\arg\max_m P(m|T^u)$

Algorithm 3: Transport Mode Inference

Finally we affect the mode obtaining the highest probability to each trajectory. The transport mode inference is summarized in Algorithm 3.

5.2.8 Origin-Destination Matrices

After modal inference, we construct OD matrices of flows which represent the total number of trips per mode. A matrix is a 3-dimensional array noted $F = (f_{ijt})$, such as an element f_{ijt} is the number of flows from origin location i to destination location j, for a given time-slot t. In particular we define respectively the total flows F^{tot}, total out-flows F^{out} and total in-flows F^{in} as follow :

$$F^{tot} = \sum_{\substack{i,j,t \\ i \neq j}} f_{ijt} \tag{5.11}$$

$$F_i^{out} = \sum_{\substack{j,t \\ i \neq j}} f_{ijt} \tag{5.12}$$

$$F_j^{in} = \sum_{\substack{i,t \\ i \neq j}} f_{ijt} \tag{5.13}$$

For each trip we find the origin and destination locations using the sector position of the first and last trip records. The choice of the spatial granularity is an important parameter which can affect the accuracy of OD matrices. First, there exists an uncertainty on the detected origin and destination positions. This uncertainty is caused by the potential delay between mobile phone use and the start or end of a trip. Noise in data can also contribute in inaccurate origin and destination positions. For our matrices, we chose two levels of spatial aggregations : departments and postcodes. These scales are coarse enough to reduce the error on the origin and destination. The Greater Paris has a $12000\ km^2$ area partitioned into 8 departments and 1382 postcode areas. The department OD matrix has 128 flow counts per mode and per day, considering the two ways of travel. The postcode OD matrix has several million T_{ijt} entries, for both transport modes.
In addition, our mobile phone data corresponds to users from one mobile phone operator, among a total of four in the region. Therefore, we rescale flows up to the total population, using expansion factors (ALEXANDER et al., 2015) per department and per postcode area. Population counts are obtained from the most recent census. The expansion factors are the inverse market share per area, calculated as the ratio of the total number of residents divided by the number of mobile phone users of the operator, living in the same area. The mean and median expansion factors are respectively 9.9 and 8.6 for Greater Paris departments. For postcode scale, mean and median expansion factors are 31.6 and 14.7.

5.2.9 Evaluation Metrics

In order to assess model performance, we use several evaluation metrics. First, we assess the separability between transport mode probabilities, using confidence intervals. Second, we propose a new metric, the transport mode Balance Index, to evaluate transport behaviors for round-trips. Third, Pearson correlation coefficients and normalized root mean square error (NRMSE) are used during validation to compare our results to external data.

Confidence Interval

In order to measure the separability between rail and road modes, the confidence interval $z^* \subset [0,1]$ of the corresponding distributions is estimated. The transport mode of a trajectory is considered as uncertain when transport probabilities are highly similar, the extreme case being a trip with identical probabilities (e.g., $P(rail) = P(road) = 0.5$). Uncertain mode trips have their probabilities falling into a certain range $q \subset [0,1]$. The confidence interval of the transport probabilities distributions is $z^* = [0,1] \setminus q$. With $N(P \in q)$ the number of uncertain trips and $N(P \in [0,1])$ the total number of trips, we calculate the ratio α of uncertain trips over total trips : $\alpha = \frac{N(P \in q)}{N(P \in [0,1])}$. Then, q is found when $1 - \alpha = 0.95$.

Transport Mode Balance Index

In this study, we define a new metric : the transport mode Balance Index. This metric assesses whether travelers performing round-trips take the same mode during both ways of their trip (e.g., leaving by road in the morning and coming back by road in the evening, for a pair of locations). This index constitutes a coherence indicator for the estimated transport modes. OD flows are filtered such as only mobile phones that traveled in both ways during the same day are retained. This index indicates whether traveling devices used the same transport mode every day for round trips. Let A and B be a pair of locations, such as $A \neq B$. For each transport mode m, a certain amount of mobile phones traveled from location A to location B, noted $N_{A \to B,m}$. The amount of mobile phones that came back from B to A is noted $N_{B \to A,m}$. Thus, the transport mode Balance Index is defined as follows.

$$\Delta_{BI}(A,B) = \frac{N_{A \to B,m}}{max(N_{B \to A,m}, 1)} - \frac{N_{B \to A,m}}{max(N_{A \to B,m}, 1)} \in [-1,1] \tag{5.14}$$

where $\Delta_{BI} = 0$ iff all phones have taken the same mode for both ways, $\Delta_{BI} = 1$ iff all phones have switched from rail to road and $\Delta_{BI} = -1$ iff all phones have switched from road to rail.

Correlation with external data

The Pearson correlation coefficient r is used to calculate the correlation between mobility variables (e.g., number of flows) extracted from cellphone data, noted x, and from external data, noted y : $r_{x,y} = \frac{COV(x,y)}{\sigma_x \sigma_y}$, where $COV(x, y)$ is the covariance between vectors x and y, and σ_x and σ_y are the standard deviations of resp. x and y.

Error comparison with travel-card data

The NRMSE is used during validation, in order to compare the daily estimated rail passenger out-flows to the travel-card out-flows : $NRMSE = \frac{1}{\bar{x_i}} \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x_i})^2}{N}}$ where $\hat{x_i}$ is the estimated number of rail passenger flows for each day, x_i is the travel-card counts over the same day, $\bar{x_i}$ is the daily average of travel-card counts and i is a postcode area. N is the total number of samples being the product of the number of days and the number of postcode areas containing a train station.

5.3 Results

This section presents the main results obtained with our model for urban transport flows. First, the results of the clustering on mobile network sectors are reported in Section 5.3.1. Second, results on transport mode inference, obtained using Bayes probabilities, are described in Section 5.3.2. Third, mobility patterns derived from OD matrices per transport mode are depicted in Section 5.3.3.

5.3.1 Clustering Mobile Network Sectors for Transportation

Mobile networks are dynamic as the signal strength of mobile network equipments is continuously updated for signal optimization and the number of cells and base stations can evolve in time. In this Section, we provide results for the mobile network sectors configuration of April 2018. Several clustering algorithms are tested to separate mobile network sectors from the Greater Paris (See Fig. 5.3). The AHC with ward linkage and the K-means produce several well defined clusters and are

hence the two best candidate clustering algorithms. Meanwhile we observe that the DBSCAN fails to find clusters because of the homogeneous densities. The AHC with average and complete linkage both result in uneven clusters size, with one preponderant cluster which include the majority of the region. These observations still hold while testing higher values of k.

The evaluation metrics are analyzed in function of k for the two algorithms (See Fig.

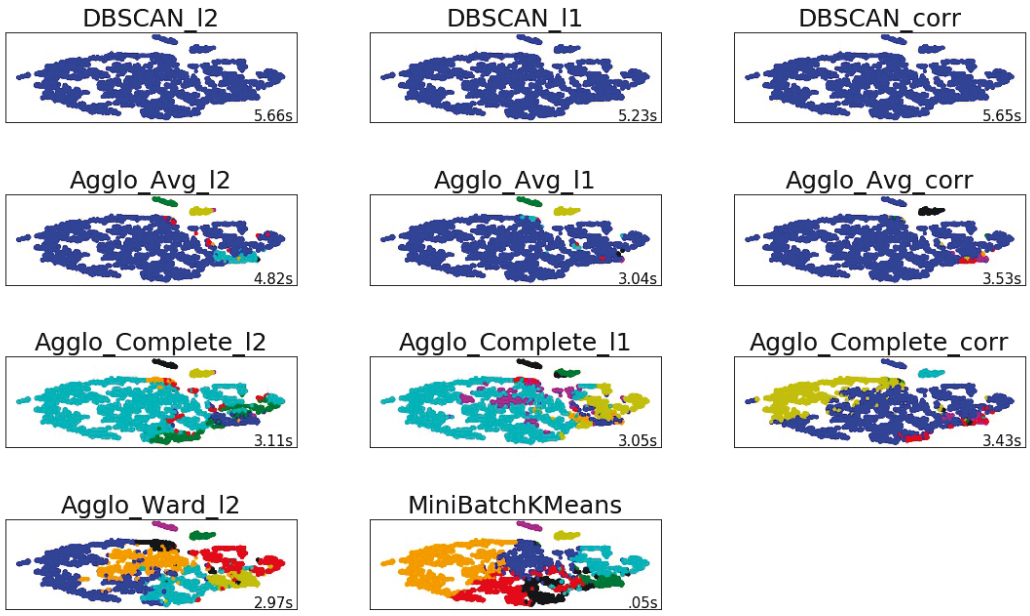

FIGURE 5.3: Benchmark of sectors clustering with t-SNE projections displayed for $K = 8$. t-SNE Parameters : $perplexity = 30$, $learning\_rate = 300$, $n\_iter = 2000$.

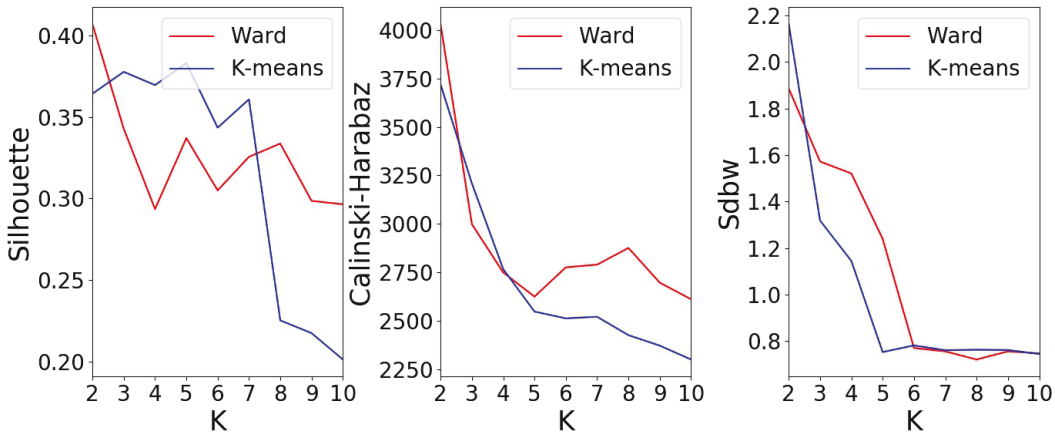

FIGURE 5.4: Clustering evaluation metrics for the standard scaler normalization, correlation distance and hierarchical clustering.

5.4). The performance of the two algorithms is similar regarding the CH and S_{dbw} index. Although the K-means silhouette is higher for small values of k, the silhouette for Ward is relatively more stable with k. K-means grants an acceptable clustering solution for $k = 5$, corresponding to the cut-off value for the three metrics. Still, the Ward AHC performance is slightly higher for higher values of k. For Ward AHC,

it is harder to identify one cut-off value. Therefore, we look for a trade-off among the local maximum and minimum metrics values. The silhouette score S and the CH index both reach a local maximum for $k = 8$ while the minimal S_{dbw} value is attained for $k = 8$. Compared to K-means with $k = 5$, the silhouette and CH index are higher for the Ward AHC with $k = 8$. Therefore we retain this clustering as our final solution. The transport probabilities are calculated for each of the eight clusters and given in Table 5.1. Greater Paris sectors are represented on Fig. 5.5.

Clusters are considered to be dominated by a mode when the probability for this

TABLE 5.1: Transport Mode probabilities and cluster size for $k = 8$

Cluster	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Size (%)	16.3	7.04	13.2	19.9	1.73	2.25	3.38	36.2
P_{Rail}	0.651	0.949	0.639	0.191	0.350	0.896	0.557	0.21
P_{Road}	0.349	0.051	0.361	0.809	0.650	0.104	0.443	0.79
Mode	multi	rail	multi	road	multi	rail	multi	road

mode is significantly high, i.e. above 0.7. When there is no dominant mode, the cluster is considered as multimodal, i.e. having both substantial road and rail mode usage. Clusters C_1, C_3 and C_7 are multimodal clusters with a higher probability for rail while C_5 is multimodal with a higher probability for road. Clusters C_2 and C_6 are rail dominated clusters containing a high proportion of indoor cells, inside the underground and train stations. Most sectors from these two clusters are located in the city center as the underground network is limited to Paris and its closest suburb areas. Eventually, C_4 and C_8 are road dominated clusters. The multimodal and road clusters are equally present in the city center and the suburb. At the time of this study, the mobile network of the Greater Paris region contains nearly 10% rail sectors, 39% road sectors and half of the sectors are multimodal. The rail mode is predominant among sectors from the city center while the road mode dominates sectors from the suburb.

5.3.2 Performance Evaluation of Transport Mode Inference

Confidence Interval

We derive the transport mode probability distribution of trajectories (see Fig. 5.6). The confidence interval for transport probabilities is $z^* = [0, 0.345] \cup [0.645, 1]$. This shows that 95% of all transport probabilities are below 0.345 or above 0.645. The remaining 5% of trips, with probabilities outside the range z^*, are categorized as uncertain mode. The transport mode can be uncertain when devices are detected in multimodal sectors. As an example, the mode is uncertain for a device having the same number of records in sectors being members of clusters C_1 and C_5 or sectors

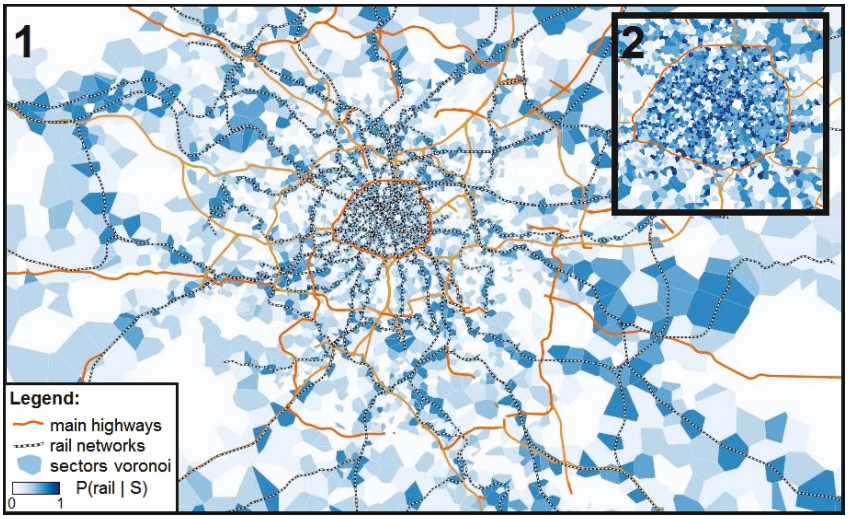

FIGURE 5.5: Sectors projected on the Greater Paris area (1) with a zoom on Paris (2). The color gradient gets a darker blue tone when the rail probability is high. Lighter sectors have higher road probabilities.

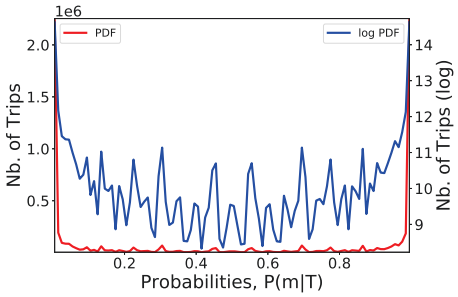

FIGURE 5.6: PDF of transport probabilities

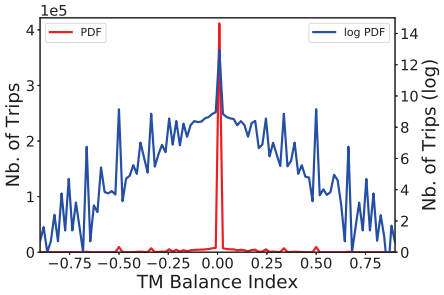

FIGURE 5.7: PDF of Balance Index Δ_{BI}

of clusters C_3 and C_5, as the road and rail probabilities are opposites (see Tab. 5.1). Still the transport probability distribution reveals that the modes are well separated for the majority of trips.

Transport Mode Balance Index

The Balance Index Δ_{BI} is calculated for one month Origin-Destination flows (see Fig. 5.7). Users that change mode for their return trip are assumed to represent a small proportion of the population. Thus, we expect to have a reasonably low amount of round trips with mode switch per OD. After calculation of Δ_{BI} for each OD, we obtain an average and median value both equal to 0 with a confidence interval of ± 0.16. This reveals that, for 95% of OD locations, there is less than 16% round trips where a mobile phone switched modes. As expected, the vast majority of devices used the same mode for both ways of travel. After observing Fig. 5.7, we identify that non-null values correspond to OD with low passenger flows, having less than 1000 trips a month, although there is no correlation between the two variables.

Thus, round trips having a mode-shift can be found in areas with fewest travelers. In real-life scenarios, some travelers have multimodal behavior. As an hypothesis, users that travel on low-frequented itineraries may experience important public transport waiting time or difficult access to public transport and thus decide to switch mode or combine several mode.

5.3.3 Mobility Patterns

Spatial Patterns

Top passenger flows for rail and road modes are displayed in Fig. 5.8 and Fig. 5.9. The top rail passenger flows involve an origin or a destination located in Paris. Top road passenger flows involve at least an origin or a destination in the suburb, or Paris périphérique (the ring road surrounding Paris). In addition, we observe top rail flows between Paris and the suburb in Fig. 5.10. Three long-distance arcs are visible and correspond to the three directions for high-speed trains (Paris-Bordeaux, Paris-Marseille and Paris-Strasbourg). Inter-suburb rail flows are depicted in Fig. 5.11. Two areas attract most suburb flows (La Défense and Saint-Denis).

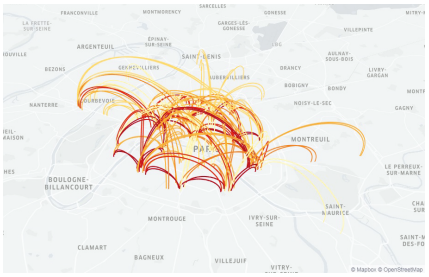

FIGURE 5.8: Top 100 rail passenger flows in the Greater Paris (zoom on Paris and the close suburb)

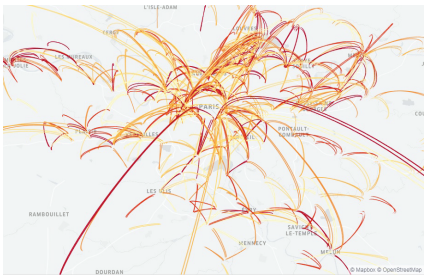

FIGURE 5.9: Top 100 road passenger flows in the Greater Paris, for trips having a distance $d > 5$ km

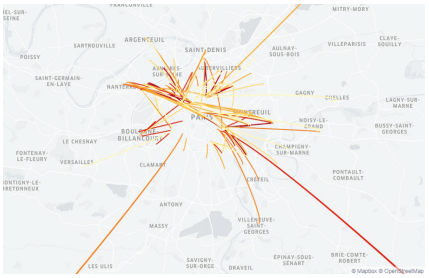

FIGURE 5.10: Top 100 rail passenger flows between Paris and the suburb

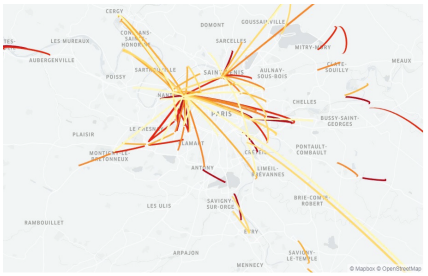

FIGURE 5.11: Top 100 rail passenger flows in the suburb, for trips having a distance $d > 5$ km

Temporal Patterns

Temporal travel patterns for a typical week are observed for each department. Average road and rail trip counts are shown in Fig. 5.12, Fig. 5.13, Fig. 5.14 and Fig. 5.15. Trip counts are calculated over 2 month of data (April and May 2018). Flows are averaged per week day, per start hour and per home department for rail (Fig. 5.12) and road mode (5.14). For business days, peak hours occur in the morning and early evening. A midday peak can also be observed at lunch time. During business days, morning and evening rail peaks are more balanced than road peaks. Rail morning peaks are slightly thiner and higher than in the evening, this phenomenon being more visible in the city center (dep 75). On the contrary the number of road trips is higher in the evening, for any week day. The phenomenon is more pronounced for departments from the second suburb ring (i.e. dep 77, 78, 91 and 95). This suggests that road users travel several times in the end of the day. Unlike for rail mode, the road midday peak height is comparable to the road morning peak. Our findings are consistent with the trend announced in the survey from 2010. The later has indicated that the number of road trips has been decreasing during morning peak (- 8 % between 7-9 AM) while it has increased during off-peak hours (+ 7% between 9AM-4PM) and evening peak time (+ 6 % from 4-7 PM) since 2001. During week-ends, peaks are less visible. Compared to working days, there is a significant drop of mobility, more pronounced for rail transport than for road trips. For rail mode there is a loss of 37% trips on Saturday and 52% on Sunday. For road mode the overall mobility loss is about 12% on Saturday and 24% on Sunday.

5.4 Validation

Our estimates are confronted to external datasets for validation. First, our results are compared to the household travel survey from 2010 to obtain a global validation at macroscopic scales. Second, results are validated against public transport data, constituted of trip counts per train stations during one month. To ensure a correct validation between different data sources, the data is processed to match the same area and time range, as well as the same spatio-temporal scales.

5.4.1 Validation with Survey

Results for one month of data, for the year 2017, are extensively compared with the latest household travel survey of the Greater Paris region, for year 2010. Responses from 43000 of the 12 millions residents were gathered for this survey. The coarser survey scale has three areas : city-center (i.e., Paris, noted CC), first ring (i.e., close

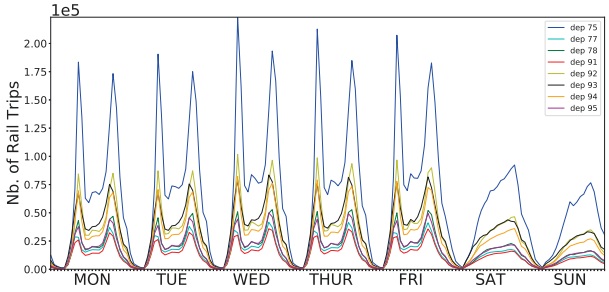

FIGURE 5.12: Weekly pattern for rail passenger flows per home department

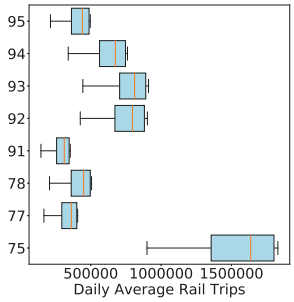

FIGURE 5.13: Boxplot for daily average rail flows per department

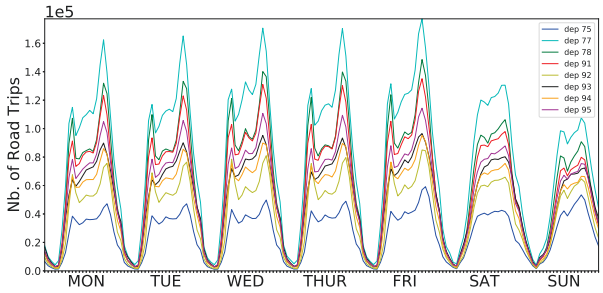

FIGURE 5.14: Weekly pattern for road passenger flows per home department

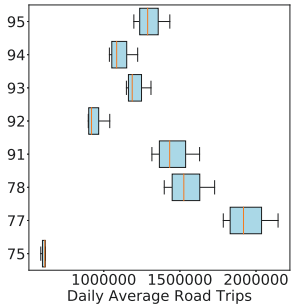

FIGURE 5.15: Boxplot for daily average road flows per department

suburb noted $R1$) and second ring (i.e., furthest suburb $R2$). The intermediate scale is the department. The smallest scale is constituted of 100 areas, called survey blocks, which are smaller than departments and larger than postcode areas. In the survey, transport modes are divided into two main categories. The first category is motorized modes which include public transport (e.g. underground, tramway, bus) and private vehicles (e.g. cars, motorbikes, taxi). The second category is unmotorized modes, namely walk and bike. In order to compare our results on rail and road modes with the survey, we group together survey trips for underground, overground and tramway as rail trips. Similarly, private vehicles and bus trips are aggregated as survey road trips. Transport modes are noted as $m \in \{all, motorized, road, rail\}$.

Average day trips per person

First, the average day trips per individual are calculated from the survey and compared to our results. The average survey trip count per resident is $C_m^{TS} = \frac{\sum_{i=1}^{k} N_{i,m} * w_i}{\sum_{i=1}^{k} w_i}$ where an individual i of weight w_i reported N_i trips for mode m during

one day of survey. The weight w_i is calculated with socio-demographic information to rescale the individual to the entire population living in an area. Similarly the average trip count per day and per mobile phone is : $C_m^{MP} = \sum_{i=1}^{U} \sum_{t=1}^{T} \frac{1}{U} \frac{1}{T} N_{i,t,m}$ where U is the number of phones, T is the number of days and $N_{i,t,m}$ is the number of trips detected for phone i during day t with transport mode m. In addition, the ratio between road and rail trips is calculated as $C_{ratio} = \frac{C_{road}}{C_{rail}}$. Pearson correlation coefficients are calculated between C_m^{TS} and C_m^{MP} and shown in Table 5.2. High positive correlations are obtained between the motorized mode category of the survey and average mobile phone day trips (from 0.47 to 0.99). The smallest correlation is obtained between survey motorized modes and mobile phones trips, from the 100 blocks (0.466). Still, survey blocks from the city center and the first ring grant higher correlations than blocks from the second ring. One possible explanation could be a sampling bias induced by a lack of survey samples in second ring blocks. Yet rail and road modes achieves high correlations for all scales (from 0.87 for blocks, up to 0.99 for rings). The obtained correlations reveal that the evolution of the modal share per individual, across different geographic areas, is consistent with the survey.

TABLE 5.2: Pearson correlation coefficients between survey and mobile phones on average day trips per individual.

Home Scale	$(C_{Motor}^{TS}, C_{All}^{MP})$	$(C_{Road}^{TS}, C_{Road}^{MP})$	$(C_{Rail}^{TS}, C_{Rail}^{MP})$	$(C_{Ratio}^{TS}, C_{Ratio}^{MP})$
Rings (CC, R1-2)	0.993	0.995	0.990	0.999
Deps (CC, D2-8)	0.751	0.960	0.986	0.978
Survey Blocks (S1-100)	0.466	0.931	0.874	0.764
Survey Blocks (CC, R1)	0.669	0.951	0.933	0.901

TABLE 5.3: Average trips per individual for a business day (source : EGT 2010-Île de France Mobilités-OMNIL-DRIEA)

Home Scale	Travel Survey (TS)					Mobile Phone (MP)			
	C_{All}^{TS}	C_{Motor}^{TS}	C_{Rail}^{TS}	C_{Road}^{TS}	C_{Ratio}^{TS}	C_{All}^{MP}	C_{Rail}^{MP}	C_{Road}^{MP}	C_{Ratio}^{MP}
All population	4.16	2.45	0.61	1.85	3.03	2.10	0.80	1.30	1.62
Paris (CC)	4.37	1.93	1.11	0.83	0.75	1.94	1.22	0.72	0.59
1st Ring (R1)	4.03	2.25	0.61	1.64	2.69	2.07	0.80	1.27	1.60
2nd Ring (R2)	4.18	2.86	0.38	2.49	6.55	2.24	0.50	1.74	3.45

Absolute values for average trips per individual are provided for the overall region, the city center and first and second rings, in Table 5.3. According to the survey, Greater Paris residents performed 2.45 motorized trips a day in 2010 while mobile phone users have an average of 2.1 day trips in 2017 (see table 5.3). In Paris, all motorized trips were detected for mobile phones users living in Paris ($C_{Motor}^{TS} \simeq C_{All}^{MP}$). Compared to survey motorized trips, the amount of mobile phones trips decreases for first and second rings (respectively -8% and -22% trips). Meanwhile, the second ring has the highest number of road trips, followed by the first ring. In

parallel, a loss of mobile phone road trips is observed, in comparison with the survey (respectively -22.5% and -30% road trips). This phenomenon is more pronounced in the suburbs. This result reveals a potential bias in phone usage for road travelers, as mobile phone calls are prohibited while driving. Compared to the survey our overall results show a global modal transfer of 13% of road trips in favor of rail transportation in the region, yet this result might be biased by undetected drivers. Paris is the area being the less affected by undetected mobile phones hence has the most reliable results. The city has witnessed a rise of 10% of rail trips per individual (and reciprocally a loss of 10% individual road trips). These findings suggest that since 2010, citizens from the Greater Paris have increased their overall use for rail transportation. The increase in transport usage could be a consequence of the construction of new transport lines in the region (e.g., six new tramways, expansion for rail lines $M4$ and RER E). In addition, the region adopted a unique fare for transport pass in September 2015, hence reducing the cost for long distance trips.

Average total day trips

Daily and hourly passenger flows, for Greater Paris residents, are estimated with our model for two months of data during spring 2018, having holidays and weekends filtered. Flows are averaged per departure hour, for business days. Similarly, flows corresponding to a business day are collected from the survey and compared to our results as depicted on Table 5.4 and Fig. 5.16. The 2010 survey reported 6.0 millions rail flows per business day. After rescaling MP trips, the average rail flows for a business day is 6.4 millions. Compared to year 2010, our results show a raise of $+6.4$% rail transport flows, during spring 2018. Comparatively, the Greater Paris transport authority reported a raise in public transport trips of $+0.8$% for the underground and $+10.1$% for the overground, corresponding to a $+10.9\%$ annual raise for year 2016 compared to 2010 (Source : Île-de-France Mobilités 2017). The 4.7% difference can possibly be caused by seasonal variations or by a certain amount of undetected mobile phones. Meanwhile, only 11 millions daily road trips are detected, against 18.2 millions in the survey, for a typical business day. This finding strengthens the hypothesis that rail passengers may have a more active phone usage while a substantial part of road passengers remains undetected possibly due to a low mobile phone activity. In addition, Pearson correlations between survey and mobile phones trips are resp. equal to 0.95 for rail trips and 0.97 for road trips. Therefore, the hourly patterns of a typical business day remain identical for survey and mobile phone data for both modes, as observed in Fig. 5.16.

In order to correct the road bias, we propose a new rescaling term, based on the survey, to estimate road flows. The number of road flows for day d and hour h can be

TABLE 5.4: Average daily trips per transport mode in the Greater Paris, for mobile phones (MP) and survey from 2010.

Mode	MP raw	MP rescaled	Survey
Rail	1227284	6383103	5999183
Road	2128750	11034581	18215180
Rail / Road	0.55	0.58	0.33

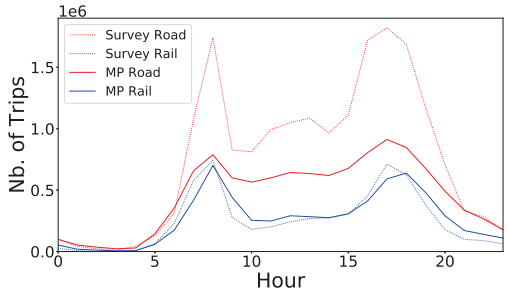

FIGURE 5.16: Daily pattern for a business day for survey and mobile phones (MP) trips volumes

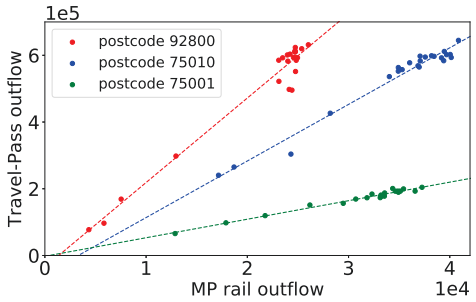

FIGURE 5.17: Regression between daily mobile phones rail trips and travel-card counts, for three postcode areas

calculated as $N_{road}(d,h) = \gamma \cdot f \cdot N^{raw}_{road}(d,h)$, where $N^{raw}_{road}(d,h)$ is the raw number of mobile phones road trips for the same time and f is the expansion factor. The additional rescaling term $\gamma = \frac{N^{S}_{road}(h)}{f \cdot N^{raw}_{road}(h)}$ is the survey scaling factor for road mode for a typical day. For this case study the median and average values for the rescaling term γ are respectively 1.49 and 1.39. For a higher precision, the number of survey road trips should correspond to the latest year available. Traffic counts should also be used for calibration in finer scale areas wherever possible.

5.4.2 Validation with Public Transports Data

Greater Paris commuters swipe their travel-cards when entering public transport. Yet, in most stations, it is not required to swipe a second time when exiting the transport system. The validation dataset is constituted of daily entry counts inside train stations, for one month data in 2017. Our model generates OD matrices containing the daily and hourly number of rail trips between postcode areas, for the same month. Through this validation step, two datasets obtained from different sources, namely mobile phones data and public transport data, are compared. The success of the validation depends on the ability to conciliate the spatial and temporal scales from both datasets. Therefore, train stations are aggregated per postcode in order to up-scale the validation data. In addition, the sum of travel counts is calculated for stations belonging to the same postcode areas. Similarly, mobile phones flows are aggregated per day and per origin location (postcode). For each day and each

postcode, the daily out-flows (i.e., the number of trips starting in the area) are obtained for both mobile phones and travel-card holders. First, Pearson correlation coefficients are calculated between daily rail out-flows estimates and validation data (see Fig. 5.18). The obtained median correlation is 0.98. The minimum correlation value is 0.68, obtained for a major leisure area in the second ring. The later contains Disneyland and the largest shopping center of the region, with two train stations. Disneyland station serves mostly highspeed trains, yet the validation dataset does not account for highspeed train tickets, which explains the lower correlation for this area.

Second, a linear relation is found between mobile phone rail out-flows and travel-card out-flows (see Fig. 5.17). A linear regression model is applied for each postcode area. In median, the NRMSE value is 0.062 (see Fig. 5.19). In comparison, without the linear regression, the state-of-the-art rescaling method with expansion factors obtains a median NRMSE value is 0.346. The calibration with travel-card per post-code area enables to account for bias specific to each train station and divide the error by a factor five. The main bias is caused by the existence of two transport operators in the Greater Paris. Consequently, travelers might need to swipe their travel-cards more than once if they change lines. Thus, validation counts contain both trip starts and transfers. Meanwhile our estimates account for rail flows only for users starting their trip at the station. A second bias is fraud rates (i.e., commuters not swiping any transport tickets), fluctuating among train stations. At last, the travel-cards counts are not a perfect validation data as some technical problems and anomalies can affect the precision of this data. As an example, the outliers in Fig. 5.17 correspond to days for which one train station had anomalies reported by the transport authorities.

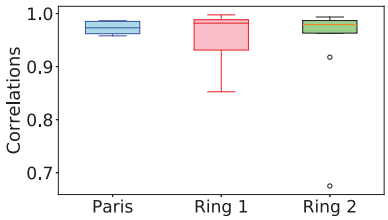

FIGURE 5.18: Daily origin outflows correlations between MP rail trips and travel-card counts

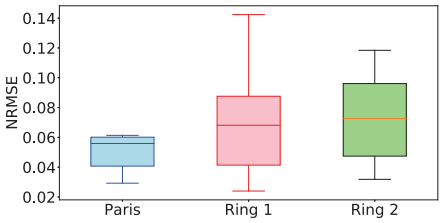

FIGURE 5.19: NRMSE between rescaled MP rail trips and travel-card counts

5.5 Comparative Evaluation

At last, we present a comparative evaluation study of our model against state-of-the-art. Thus, we reproduce the approach proposed by Wang et al (H. WANG et al., 2010) which addresses unsupervised transport mode estimation with CDR and is

our unique competitor. This method consists in calculating the distribution of travel times per OD. Two features, namely the travel time of CDR trajectories and the number of flows with same travel time grouped by same OD, are fed to a k-means algorithm, producing two clusters per OD. For a direct comparison with our results, we apply the aforementioned method to all OD pairs in the Greater Paris, at the Commune scale. In order to apply one clustering phase to the complete dataset, for all OD pairs, we add a recursive normalization procedure. The features are normalized for each OD with a Robust Scaler which is a z-score that considers the mean and standard deviations of features in the range of the first and third quantiles, hence robust to outliers. This step is used to account for very short or very long trips which can be caused by noise. Such normalization enables us to consider all trips, despite abnormal travel time values, instead of simply filtering out anomalies. Thus we can compare these results with our model, considering all trips. In order to identify the mode of each cluster, we calculate the average travel time per cluster and compare it to the survey travel time. This step is similar to Wang et al's method which used Google travel time instead of a survey. Our reason for not using Google travel time is because the exact coordinates of the origin and destination are required. Wang et al worked with triangulated CDR while the location precision of our data is coarse. Therefore, in our case, the survey scale is more convenient to calculate and compare the average travel times. Based on the Greater Paris survey, the average travel time for road and rail modes are respectively 29 and 58 minutes, between Communes.

First we compare the average travel time for our classification to the survey (See Tab. 5.5). There is a perfect match for the road travel time. Meanwhile, we obtain a rail travel time nearly twice smaller. A possible cause can be because the survey reports the travel time of the whole trip between two locations while assigning one majority mode. For instance if a person first walks to a train station, this walking time is accounted, yet the trip has a train mode. If we assume that a person walks approximately 15 minutes to and from a train station, this remains consistent with our finding.

Second we assess the performance of Wang et al's method. After applying the clustering for $k = 2$, we obtain highly unbalanced clusters, with 7% trips for the cluster assumed to be the rail mode. Cluster C_0 exhibits a high average travel time being 2h45 (see. Tab. 5.5). Here it is clear that the clustering fails to separate the mode, based on the travel time. In addition, we test the clustering with $k = 3$ (see. Tab. 5.6). As a result, the sizes of clusters C_0 and C_1 are more balanced. Trips with longest duration are grouped in the smallest cluster C_2. The Adjusted Rand Index is used to compare the clustering outputs with our classification. Let x_1 be the vector representing the transport mode labels assigned with our model. Let x_2 be the vector representing the clusters labels obtained with Wang et al approach. The corresponding ARI is close to zero ($ARI(x_1, x_2) = 0.005$), revealing that the two classifications are independent. The results are compared to the travel

survey, using the correlations of the hourly transport flows for a typical business day (see. Tab. 5.7). Although the baseline obtains high correlations with the survey, our model improves the correlations. Eventually, the absolute errors with the survey are calculated. After rescaling, our model obtains a percentage error of $+6.4\%$ against -43% for the baseline, considering the rail mode.

TABLE 5.5: Travel times for our modal classification, the baseline and the survey

Our Model			Wang et al 2010			Survey
Mode	Size (%)	t_{avg} (min)	Cluster	Size (%)	t_{avg} (min)	t^S_{avg} (min)
Rail	35	25	0	7	165	56
Road	75	28	1	93	25	29

TABLE 5.6: Baseline for k=3.

Cluster	Size (%)	t_{avg} (min)
0	33	37
1	64	16
2	3	112

TABLE 5.7: Correlations with survey flows

Mode	Our Model	Wang et al 2010	
		$k=2$	$k=3$
Rail	0.95	0.72	0.88
Road	0.97	0.95	0.94

5.6 Discussion

The presented work proposes a methodology for the estimation of Origin-Destination flows for rail and road transport modes. Mobile network data and transport networks infrastructures are used to infer trips transport mode. Travel surveys and census data are used to upscale the number of trips. Our results show high correlations with travel survey and similar mobility trends. In addition, we obtain both high correlations and small errors with travel-card data. The extensive comparison of results with these two external datasets proves the validity of our method.

This model brings a significant contribution to the state-of-the-art. First, it outperforms our unique competitor on the task of unsupervised transport mode detection with CDR. Second, the approach is robust to noise and sparse data, making it generalizable to all types of mobile network records (i.e., CDR and passive records with varying frequency). The model can be reproduced by practitioners that have access to CDR and is generalizable to other regions for which transport networks, census and travel survey are available. In order to achieve highest performance in terms of absolute errors, mobile phone data should be used jointly with travel surveys, public transport data and traffic counts, whenever possible, for optimal fine-grained calibration of OD trips.

Although our results stand for a good model performance, several open issues remain. The first limitation is inherent to the coarse spatial scale of the mobile network data, which prevents mobility estimation at a fine resolution. In the present study, our results are computed at the Commune scale.

A second open issue of this model is the assignment of one main transport mode to each trip meanwhile in real life scenarios, multi-modal trips can occur. Detecting when users switch mode during their trip is a delicate task in reason of noisy and coarse geolocation, and delayed times for start and end of a detected trip. To handle multi-modal trips, we believe that one could apply decision rules based on sectors transport probabilities. The following conditions are a non-exhaustive list which could be applied to identify multi-modal trips :

— Several records are located in sectors being member of clusters associated to rail mode (i.e., visited sectors in C_2, C_6) and to road mode (i.e., visited sectors in C_4, C_8).

— The first or last visited sector belongs to a cluster with a transport mode being different from the trajectory mode. For instance, first or last record is road while the rest of the trip is classified as rail mode.

— A trajectory contains at least two sequences of visited sectors, each having at least two sectors, such as one sequence has rail sectors and the other has road sectors.

However the lack of ground truth limits the possibilities for multi-modal trips validation.

The third limitation of the method is that we consider a bi-modal separation into road and rail trips. This work lefts aside the difficult task of separating private vehicles from road public transport users, such as bus passengers. To solve this problem, several challenges have to be addressed in the future. Indeed, private and public vehicle riders often share the same routes. Similarly, high-speed trains, tramway, underground and overground can also share the same railway routes (e.g. serve the same stations). Therefore, a modal classification solely based on spatial features is not sufficient. In order to further identify other transport modes, an extension to the present work could be to construct trajectory features fed into a second clustering model, aimed at clustering the road trajectories into private (car) vs. public (bus) and the rail trajectories into highspeed, underground, overground and tramway. One could derive temporal features coupled with behavioral features such as mobile phone activity or the number of travelers passing by the same sectors etc. The travel survey should be used for validation of the unsupervised model as it is the unique data source reporting travel information from various transport modes.

This work confirms the potential of mobile network data to strengthen travel demand models such as the four step models during trip generation, trip distribution and mode choice. Concerning route assignment, mobile network records are limited by their moderate frequency and their coarse spatial granularity. Mobile network trips are imprecise and incomplete, which makes route determination a complicated task. In case triangulated passive mobile phone data are authorized and available, the higher spatio-temporal precision of such data can help in determining users' routes. Otherwise, other traditional optimization approaches for traffic assignment can be

employed yet the problem of missing ground truth still holds.
Eventually, this work can be used for several applications such as the determination of optimal locations for the construction of new transport infrastructures, and the study of impacts of particular events such as transport strikes, meteorological events, protests or sports games (e.g., world cups, Olympic games). Thus, we believe this work will help the transport community for the development of smart transport applications, and encourage the collaboration between transport authorities and mobile phone operators.

5.7 Summary

Fast urbanization generates increasing amounts of commuters' flows, urging the need for efficient transport planning policies. In parallel, mobile phone data have emerged as the largest mobility data source but is not yet integrated to transport planning models. Currently, transport authorities are lacking a global picture of daily passenger flows on multimodal transport networks. In this work, we propose the first methodology to infer total Origin-Destination flows per transport modes using Call Detail Records, over a complete region having dense transport networks. For this study, we pre-process 360 millions trips for more than 2 millions devices from the Greater Paris as our case study region. The model combines mobile phone data with transport networks geospatial data, travel survey, census and travel-card data. Transport modes are identified through a two-steps algorithm involving clustering of mobile network areas and inference of trips mode with Bayesian probabilities. After attributing the mode with highest probability to each trip, we construct transport mode Origin-Destination matrices. Flows are up-scaled to the total population using expansion factors. For the road mode, expansion factors are up-scaled a second time with the survey to account for undetected cellphones because of inactive drivers and short distance trips. The model outputs the hourly road and rail passenger flows for the complete region, at the zipcode scale. From our results, we observe different mobility patterns for road and rail mode and between Paris and the suburb. Results are extensively validated against survey and travel-card data. Our model brings a major contribution to the state-of-the-art and outperforms our unique competitor on the identification of transport flows per OD with CDR.

6 Conclusion

Mobile network geolocation data provide huge promises toward improving urban planning strategies. Although important pre-processing is required to limit the effect of noise, sparsity and the modest precision of the raw positioning, mobile network data have proven to be extremely valuable inputs for mobility estimation models. In this thesis, two main problematics are solved on the case study of the Greater Paris region. The first work addresses the estimation of population densities and number of visitors in urban areas, while capturing their fluctuations over time. The second work presents a methodology for the estimation of total Origin-Destination flows per transport mode. Both models are solved through an unsupervised learning scheme. Despite having unlabeled data, the estimates are extensively validated against external large-scale mobility datasets. This thesis provides several contributions to the literature on urban mobility mining with mobile network data.

6.1 Contributions

6.1.1 Population Attendances

The dynamic attendances estimation study unveils important sensitivity of model parameters to data frequency and spatial scale. Although the data frequency can modify the range of parameters values without affecting model performance, down-scaling the spatial scale increases the errors. A corrected spatial mapping is proposed, based on a population density weighting factor, reducing the errors at the census block resolution by nearly 20%. This work exhaustively develops the interpretation of the model parameters for the first time, emphasizing the combined influence of the market share, data frequency and penetration rate. In addition, this study unveils the need for a different rescaling term at the smaller census block scale between city center and suburb, depending on the target region. Finally, our model outperforms our competitors on attendances estimation during football games. The novelty of this validation study is that indoor mobile network cells are used to estimate attendances inside stadiums. The high performance on validation (errors below 10%) proves that the dynamic parameters, initially estimated at census block resolution, can be extended to the cellular scale.

6.1.2 Transport Mode OD Matrices

This work is the first unsupervised learning method able to identify the transport mode from all CDR trajectory, without any data filtering. As mobile network geolocation is sparse and noisy, past studies tend to apply several filtering to remove low activity devices and trajectories with few samples or anomalies. The main strength of our model is that only two locations are required to determine the transport mode. The transport probability is updated with each record without the need of the exact complete itinerary. Our work further extends the modal inference to the estimation of total hourly OD flows, traveling between all the 1276 zipcode areas in the region. In this perspective, cellphone flows are upscaled to total population flows using expansion factors. Compared to previous methods on OD matrices construction, we identify that our mobile network records are subject to a bias for the road mode users. Compared to the travel survey, a third of road travelers are undetected, in average, by reason of a lack of mobile phone activity. This is a consequence of both non-calling drivers and short-distance trips remaining in the same location area. Thus we have proposed to correct this bias through a second calibration, using the total flows for a typical day estimated with the survey. Our final model combines several mobility data sources, involved at different stages such as feature construction, calibration and validation. Different mobility patterns can be observed between road and rail users, between the city center and the suburb, and between week-days and week-ends. Our estimates are validated against the survey and travel-cards with correlations higher than 95% and errors below 10%.

6.1.3 General Contributions

All experiments presented in this thesis involve processing Terabytes of mobile network records, which are real datasets, from a large-scale region, wide of 1200km^2. Although two different problematics are addressed, the proposed approaches present several similarities. First, classic normalization techniques perform poorly on the spatio-temporal data at hand. Thus, the proposed data normalization strategies account for heterogeneous urban density bias. Second, the models are capable of handling both sparse and high frequency mobile network records i.e., CDR and passive records. Although high frequency passive records (XDR) are becoming increasingly popular among mobile network operators, such data are more expensive in terms of material resources, computational complexity and might be subject to stricter data regulation laws which are state dependent. In the end, our initial data sparsity constraint has been overcome and turns out being a strength to reduce the running time. Ultimately, the key issue for mobile network providers is to set a record collection frequency ensuring reasonable trade-off between performance and running time.

6.2 Limitations

The present work is constrained by several limitations. The first limitation is inherent to the data in general. On the one hand, the scale of the mobile network is coarse and prevents the estimation of urban mobility at fine-grained resolutions. In other words, mobile network geolocation cannot compete with the precision of GPS tracking outdoor, hence should be used for other purposes such as large scale mobility patterns of population. As a consequence the estimates derived at the zipcode scale obtain highest performance (less than 10% errors) while the smallest census block scale (IRIS) suffers from important error rates, three to six times higher. On the other hand, unlabeled data and missing ground truth are a persistent problem limiting possibilities for model training, calibration, testing and validation. In addition, several open issues persist concerning the transport mode inference model. First, it is not possible to perform an individual validation because of user privacy issues (e.g., risk of de-anonymization, need for individual user consent etc.). Second, the model generates a binary classification into road or rail mode meanwhile there exist additional modes. This choice is motivated by the possibility to extract a subset of bi-modal labels associated to mobile network sectors. For other modes (e.g., bus, tramway, bike etc.), label extraction is not possible. Indeed, as mobile network sectors remain coarse while the transport networks are dense, there are generally several rail modes (e.g., underground and overground) or several road modes (e.g., bus and car), or both, within each sector. Concerning non-motorized modes such as walkers, they are initially undetected because the movements of walkers are too microscopic compared to the coarse scale of the mobile network and cellphones are considered in a stay position. Similarly, the bike mode is not guaranteed to be detected, unless in case of long distance trips involving a change of location area at the network scale (i.e., several kilometers). Third, the model associates one main mode to each cellphone trajectory despite the possible existence of multi-modal trips. Although we propose a heuristic-based solution for multi-modal trips, additional validation tests and ground truth information are required.
Still, despite the aforementioned limitations, the presented works bring several contributions and can be further developed for future work perspectives.

6.3 Perspectives

Eventually, we provide several guidelines for future works. First, concerning the estimation of population densities and attendances, we propose to investigate on training several models on the residential areas partitioned into several administrative areas (e.g., for the Greater Paris, the department and Commune scale can be worth investigating). The objective is to prevent overestimating the population in

low-density areas while underestimating in high-density zones which is the drawback of fitting one model on a large heterogeneous region.
Second, the identification of additional transport modes could be addressed in future studies aiming to understand mobility flows across several modes, namely car, bus, highspeed train, underground, overground, tramway or bike. In our opinion, our binary transport classification is necessary to first cope with sparse and noisy data and classify all trajectories. For additional modes, we propose using our road and rail classification as features for a second clustering on trajectories and suggest investigating the following features :

— Temporal features : start time, end time, duration, speed etc.

— Behavioral features : mobile phone activity, number of cellphones traveling together etc.

— Additional spatial features on railway and road types classified by maximum speed.

Regarding data normalization, we believe one could apply specific normalization on each feature, for each OD pair and possibly for each departure time to account for the impact of rush hours. For validation, travel surveys constitute the most complete data source reporting travel information from various transport modes hence are extremely useful to derive mobility trends.
Third, each mobile provider might collect their data at a different frequency and scale while applying different pre-processing strategies (e.g., noise reduction, segmentation etc.). Therefore it is important to know your data, hence be aware of its initial strength and weaknesses in order to better interpret the results. In other words, one should assess model performance considering both the choice of methodology and the data quality. A poor pre-processing will degrade the performance, regardless of the pertinence of the method. Reciprocally, the flaws of the raw data can be reduced as long as they are well identified, by combining pre-processing to robust models.
Fourth, we encourage sharing up-to-date anonymous mobility datasets on open data platforms and collaborative research works between private stakeholders from transport, telecommunication, construction and retail industries. Eventually, real-time applications of the proposed models are another direction for future works.

To conclude this thesis, we review possible applications of our work for smart city planning.

— Smart transport : OD flows per transport mode can be used to update and strengthen travel demand estimation models. Other applications are the determination of anomalies on transport networks such as accidents and perturbations, the determination of optimal locations for the construction of new transport infrastructures, and the study of impacts of particular events such as transport strikes, meteorological events, protests, recreational gathering etc.

— Smart tourism & retail : the hourly number of visitors can be estimated for indoor environments equipped with indoor base stations such as shopping centers or any closed recreational areas. For outdoor areas, it is possible to estimate the number of additional visitors (i.e. other than the regular visitors) in case of a particular event involving the meeting of a large group of population. This can be done by measuring the variation of the estimates compared to a normal day. Still for outdoor areas, the absolute number of regular visitors is more prone to errors, the error rate being scale dependent.

The new findings demonstrated in this thesis, we hope, will help the research community and encourage future collaborations between multi-disciplinary practitioners, from public and private organizations, toward improving urban planning by mining massive mobility data collected from smartphones.

7 Bibliographie

AGUILÉRA, Vincent, Sylvain ALLIO, Vincent BENEZECH, François COMBES et Chloé MILION (2014). „Using cell phone data to measure quality of service and passenger flows of Paris transit system“. In : *Transportation Research Part C : Emerging Technologies* 43. Special Issue with Selected Papers from Transport Research Arena, p. 198–211 (cf. p. 15, 16, 72).

AHAS, Rein, Anto AASA, Siiri SILM et Margus TIRU (2010). „Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area : Case study with mobile positioning data“. In : *Transportation Research Part C : Emerging Technologies* 18.1. Information/-Communication Technologies and Travel Behaviour Agents in Traffic and Transportation, p. 45–54 (cf. p. 1, 71).

ALEXANDER, Lauren, Shan JIANG, Mikel MURGA et Marta C GONZÁLEZ (2015). „Origin-destination trips by purpose and time of day inferred from mobile phone data“. In : *Transportation Research Part C : Emerging Technologies* 58, p. 240–250 (cf. p. 2, 3, 6, 13, 14, 17, 72, 79).

ANDRIENKO, Gennady, Natalia ANDRIENKO, Salvatore RINZIVILLO et al. (2009). „Interactive visual clustering of large collections of trajectories“. In : *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, p. 3–10 (cf. p. 18).

ARCEP (2014). *ARCEP*. `http://www.arcep.fr`. Online ; accessed October 2018 (cf. p. 55).

ASGARI, Fereshteh, Alexis SULTAN, Haoyi XIONG, Vincent GAUTHIER et Mounim A EL-YACOUBI (2016). „CT-Mapper : Mapping sparse multimodal cellular trajectories using a multilayer transportation network“. In : *Computer Communications* 95, p. 69–81 (cf. p. 1, 15, 71).

BACHIR, Danya, Vincent GAUTHIER, Mounim EL YACOUBI et Ghazaleh KHODABANDELOU (2017). „Using mobile phone data analysis for the estimation of daily urban dynamics“. In : *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*. IEEE, p. 626–632 (cf. p. 1, 16, 40, 71).

BACHIR, Danya, Ghazaleh KHODABANDELOU, Vincent GAUTHIER, Mounim EL YACOUBI et Jakob PUCHINGER (2019). „Inferring Dynamic Origin-Destination Flows by Transport Mode using Mobile Phone Data“. In : *Transportation Research Part C : Emerging Technologies* (cf. p. 72).

BACHIR, Danya, Ghazaleh KHODABANDELOU, Vincent GAUTHIER, Mounim EL YACOUBI et Eric VACHON (2018). „Combining Bayesian Inference and Clustering for Transport Mode Detection from Sparse and Noisy Geolocation Data“. In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, p. 569–584 (cf. p. 1, 71, 72).

BAGROW, James P, Dashun WANG et Albert-Laszlo BARABASI (2011). „Collective response of human populations to large-scale emergencies“. In : *PloS one* 6.3, e17680 (cf. p. 39).

BAYIR, Murat Ali, Murat DEMIRBAS et Nathan EAGLE (2009). „Discovering spatiotemporal mobility profiles of cellphone users“. In : *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a*. IEEE, p. 1–9 (cf. p. 7).

BECKER, Richard A, Ramon CACERES, Karrie HANSON et al. (2011). „Route classification using cellular handoff patterns“. In : *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, p. 123–132 (cf. p. 1, 71).

BERLINGERIO, M., F. CALABRESE, G. DI LORENZO et al. (2013a). „AllAboard : a system for exploring urban mobility and optimizing public transport using cellphone data.“ In : t. pt.III. IBM Research, Dublin, Ireland (cf. p. 13, 39).

– (2013b). „AllAboard : a system for exploring urban mobility and optimizing public transport using cellphone data.“ In : t. pt.III. IBM Research, Dublin, Ireland (cf. p. 72).

BHAT, Chandra R et Frank S KOPPELMAN (1999). „Activity-based modeling of travel demand“. In : *Handbook of transportation Science*. Springer, p. 35–61 (cf. p. 71).

BILJECKI, Filip, Hugo LEDOUX et Peter VAN OOSTEROM (2013). „Transportation mode-based segmentation and classification of movement trajectories“. In : *International Journal of Geographical Information Science* 27.2, p. 385–407 (cf. p. 15).

BLONDEL, Vincent D, Adeline DECUYPER et Gautier KRINGS (2015). „A survey of results on mobile phone datasets analysis“. In : *EPJ Data Science* 4.1, p. 10 (cf. p. 1, 71).

BOTTA, Federico, Helen Susannah MOAT et Tobias PREIS (2015). „Quantifying crowd size with mobile phone and Twitter data“. In : *Royal Society open science* 2.5, p. 150162 (cf. p. 39, 40).

BROCKMANN, Dirk, Lars HUFNAGEL et Theo GEISEL (2006). „The scaling laws of human travel“. In : *Nature* 439.7075, p. 462–465 (cf. p. 26).

CACERES, N, JP WIDEBERG et FG BENITEZ (2007). „Deriving origin destination data from a mobile phone network“. In : *Intelligent Transport Systems, IET* 1.1, p. 15–26 (cf. p. 14).

CADEZ, Igor V, Scott GAFFNEY et Padhraic SMYTH (2000). „A general probabilistic framework for clustering individuals and objects“. In : *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 140–149 (cf. p. 18).

CALABRESE, Francesco, Massimo COLONNA, Piero LOVISOLO, Dario PARATA et Carlo RATTI (2011). „Real-time urban monitoring using cell phones : A case study in Rome“. In : *IEEE Transactions on Intelligent Transportation Systems* 12.1, p. 141–151 (cf. p. 6).

CALABRESE, Francesco, Giusy DI LORENZO, Liang LIU et Carlo RATTI (2011). „Estimating origin-destination flows using mobile phone location data“. In : *IEEE Pervasive Computing* 10.4, p. 0036–44 (cf. p. 13, 72).

CALABRESE, Francesco, Mi DIAO, Giusy DI LORENZO, Joseph FERREIRA JR et Carlo RATTI (2013). „Understanding individual mobility patterns from urban sensing data : A mobile phone trace example“. In : *Transportation research part C : emerging technologies* 26, p. 301–313 (cf. p. 1, 5, 6, 71).

CALIŃSKI, Tadeusz et Jerzy HARABASZ (1974). „A dendrite method for cluster analysis“. In : *Communications in Statistics-theory and Methods* 3.1, p. 1–27 (cf. p. 35).

CHEN, Cynthia, Ling BIAN et Jingtao MA (2014). „From traces to trajectories : How well can we guess activity locations from mobile phone traces ?“ In : *Transportation Research Part C : Emerging Technologies* 46, p. 326–337 (cf. p. 2, 71).

CHEN, Cynthia, Jingtao MA, Yusak SUSILO, Yu LIU et Menglin WANG (2016). „The promises of big data and small data for travel behavior (aka human mobility) analysis“. In : *Transportation research part C : emerging technologies* 68, p. 285–299 (cf. p. 1, 6, 71).

ÇOLAK, Serdar, Lauren P ALEXANDER, Bernardo G ALVIM, Shomik R MEHNDIRATTA et Marta C GONZÁLEZ (2015a). „Analyzing cell phone location data for urban travel : current methods, limitations, and opportunities“. In : *Transportation research record : Journal of the transportation research board* 2526, p. 126–135 (cf. p. 2, 72).

– (2015b). „Analyzing cell phone location data for urban travel : current methods, limitations, and opportunities“. In : *Transportation Research Record : Journal of the Transportation Research Board* 2526, p. 126–135 (cf. p. 7, 14).

CSÁJI, Balázs Cs, Arnaud BROWET, Vincent A TRAAG et al. (2013). „Exploring the mobility of mobile phone users“. In : *Physica A : Statistical Mechanics and its Applications* 392.6, p. 1459–1473 (cf. p. 9, 17, 24).

DEMISSIE, Merkebe Getachew, Gonçalo Homem de ALMEIDA CORREIA et Carlos BENTO (2013). „Intelligent road traffic status detection system through cellular networks handover information : An exploratory study“. In : *Transportation Research Part C : Emerging Technologies* 32, p. 76–88 (cf. p. 1, 13, 71).

DEVILLE, Pierre, Catherine LINARD, Samuel MARTIN et al. (2014). „Dynamic population mapping using mobile phone data“. In : *Proceedings of the National Academy of Sciences* 111.45, p. 15888–15893 (cf. p. 10–12, 39, 40, 43, 56, 57, 61, 65, 66).

DI LORENZO, Giusy, Marco SBODIO, Francesco CALABRESE et al. (2016). „Allaboard : visual exploration of cellphone mobility data to optimise public transport“. In : *IEEE transactions on visualization and computer graphics* 22.2, p. 1036–1050 (cf. p. 72).

DONG, Honghui, Mingchao WU, Xiaoqing DING et al. (2015). „Traffic zone division based on big data from mobile phone base stations“. In : *Transportation Research Part C : Emerging Technologies* 58, p. 278–291 (cf. p. 1, 5, 13, 71).

DOUGLASS, Rex W, David A MEYER, Megha RAM, David RIDEOUT et Dongjin SONG (2015). „High resolution population estimates from telecommunications data“. In : *EPJ Data Science* 4.1, p. 1 (cf. p. 10, 11, 39, 40).

ESTER, Martin, Hans-Peter KRIEGEL, Jörg SANDER, Xiaowei XU et al. (1996). „A density-based algorithm for discovering clusters in large spatial databases with noise.“ In : *Kdd*. T. 96. 34, p. 226–231 (cf. p. 33).

FURNO, Angelo, Marco FIORE, Razvan STANICA, Cezary ZIEMLICKI et Zbigniew SMOREDA (2017). „A tale of ten cities : Characterizing signatures of mobile traffic in urban areas“. In : *IEEE Transactions on Mobile Computing* 16.10, p. 2682–2696 (cf. p. 11, 39, 47, 48).

FURNO, Angelo, Razvan STANICA et Marco FIORE (2015). „A comparative evaluation of urban fabric detection techniques based on mobile traffic data“. In : *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, p. 689–696 (cf. p. 11, 40, 45).

GADZIŃSKI, Jędrzej (2018). „Perspectives of the use of smartphones in travel behaviour studies : Findings from a literature review and a pilot study“. In : *Transportation Research Part C : Emerging Technologies* 88, p. 74–86 (cf. p. 1, 71).

GAFFNEY, Scott et Padhraic SMYTH (1999). „Trajectory clustering with mixtures of regression models“. In : *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 63–72 (cf. p. 18).

GERLAND, Patrick, Adrian E RAFTERY, Hana SEVCÍKOVÁ et al. (2014). „World population stabilization unlikely this century“. In : *Science* 346.6206, p. 234–237 (cf. p. 1).

GIRARDIN, Fabien, Andrea VACCARI, Alexandre GERBER, Assaf BIDERMAN et Carlo RATTI (2009). „Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate“. In : *Intl. Conference on Computers in Urban Planning and Urban Management* (cf. p. 39).

GOH, CY, J DAUWELS, N MITROVIC et al. (2012). „Online map-matching based on hidden markov model for real-time traffic sensing applications“. In : *2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE, p. 776–781 (cf. p. 14).

GONZALEZ, Marta C, Cesar A HIDALGO et Albert-Laszlo BARABASI (2008). „Understanding individual human mobility patterns“. In : *Nature* 453.7196, p. 779–782 (cf. p. 26–28).

GONZALEZ, P, J WEINSTEIN, S BARBEAU et al. (2008). „Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones“. In : *15th World congress on intelligent transportation systems* (cf. p. 15).

HALKIDI, Maria et Michalis VAZIRGIANNIS (2001). „Clustering validity assessment : Finding the optimal partitioning of a data set“. In : *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, p. 187–194 (cf. p. 36).

HORN, Christopher, Stefan KLAMPFL, Michael CIK et Thomas REITER (2014). „Detecting outliers in cell phone data : correcting trajectories to improve traffic modeling“. In : *Transportation Research Record : Journal of the Transportation Research Board* 2405, p. 49–56 (cf. p. 9).

HU, Congwei, Wu CHEN, Yongqi CHEN et Dajie LIU (2009). „Adaptive Kalman filtering for vehicle navigation“. In : *Positioning* 1.04 (cf. p. 14).

HUNTER, Timothy, Teodor MOLDOVAN, Matei ZAHARIA et al. (2011). „Scaling the Mobile Millennium system in the cloud“. In : *Proceedings of the 2nd ACM Symposium on Cloud Computing*. ACM, p. 28 (cf. p. 14).

INSEE (2013). *Type IRIS*. `https://www.insee.fr/fr/information/2438155`. Online ; accessed 2016 (cf. p. 44).

IQBAL, Md Shahadat, Charisma F CHOUDHURY, Pu WANG et Marta C GONZÁLEZ (2014). „Development of origin–destination matrices using mobile phone call data“. In : *Transportation Research Part C : Emerging Technologies* 40, p. 63–74 (cf. p. 14, 39, 72).

JAHANGIRI, Arash et Hesham A RAKHA (2015). „Applying machine learning techniques to transportation mode recognition using mobile phone sensor data“. In : *IEEE transactions on intelligent transportation systems* 16.5, p. 2406–2417 (cf. p. 39).

JÄRV, Olle, Rein AHAS et Frank WITLOX (2014). „Understanding monthly variability in human activity spaces : A twelve-month study using mobile phone call detail records“. In : *Transportation Research Part C : Emerging Technologies* 38, p. 122–135 (cf. p. 5).

JIANG, Shan, Joseph FERREIRA et Marta C GONZALEZ (2017). „Activity-based human mobility patterns inferred from mobile phone data : A case study of Singapore“. In : *IEEE Transactions on Big Data* 3.2, p. 208–219 (cf. p. 2, 6, 14, 17, 71).

JIANG, Shan, Gaston A FIORE, Yingxiang YANG et al. (2013). „A review of urban computing for mobile phone traces : current methods, challenges and opportunities“. In : *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*. ACM, p. 2 (cf. p. 10).

KANG, Hye-Young, Joon-Seok KIM et Ki-Joune LI (2009). „Similarity measures for trajectory of moving objects in cellular space“. In : *Proceedings of the 2009 ACM symposium on Applied Computing*. ACM, p. 1325–1330 (cf. p. 18).

KAUFMAN, Leonard et Peter J ROUSSEEUW (2009). *Finding groups in data : an introduction to cluster analysis*. T. 344. John Wiley & Sons (cf. p. 34).

KHODABANDELOU, Ghazaleh, Vincent GAUTHIER, Marco FIORE et Mounim A EL YACOUBI (2018). „Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata“. In : *IEEE Transactions on Mobile Computing* (cf. p. 11, 12, 55, 56, 59, 65).

KHODABANDELOU, Ghazaleh, Vincent GAUTHIER, Mounim EL-YACOUBI et Marco FIORE (2016a). „Population estimation from mobile network traffic metadata“. In : *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A*. IEEE, p. 1–9 (cf. p. 1, 11, 39, 58, 71).

– (2016b). „Population estimation from mobile network traffic metadata“. In : *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A*. IEEE, p. 1–9 (cf. p. 11, 40, 64).

KRINGS, Gautier, Francesco CALABRESE, Carlo RATTI et Vincent D BLONDEL (2009). „Urban gravity : a model for inter-city telecommunication flows“. In : *Journal of Statistical Mechanics : Theory and Experiment* 2009.07, p. L07003 (cf. p. 10).

LARIJANI, Anahid Nabavi, Ana-Maria OLTEANU-RAIMOND, Julien PERRET, Mathieu BRÉDIF et Cezary ZIEMLICKI (2015). „Investigating the Mobile Phone Data to Estimate the Origin Destination Flow and Analysis ; Case Study : Paris Region“. In : *Transportation Research Procedia* 6, p. 64–78 (cf. p. 15).

LEE, Jae-Gil, Jiawei HAN et Xiaolei LI (2008). „Trajectory outlier detection : A partition-and-detect framework“. In : *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, p. 140–149 (cf. p. 18).

LERNER, Wilhelm et François-Joseph VAN AUDENHOVE (2012). „The future of urban mobility : Towards networked, multimodal cities in 2050“. In : *Public Transport International-English Edition* 61.2, p. 14 (cf. p. 1).

LIU, Fagui et Zhijie ZHANG (2017). „Adaptive density trajectory cluster based on time and space distance“. In : *Physica A : Statistical Mechanics and its Applications* 484, p. 41–56 (cf. p. 18).

LIU, Yanchi, Zhongmou LI, Hui XIONG, Xuedong GAO et Junjie WU (2010). „Understanding of internal clustering validation measures“. In : *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, p. 911–916 (cf. p. 35, 36).

MA, Xiaolei, Yao-Jan WU, Yinhai WANG, Feng CHEN et Jianfeng LIU (2013). „Mining smart card data for transit riders' travel patterns“. In : *Transportation Research Part C : Emerging Technologies* 36, p. 1–12 (cf. p. 71).

McNally, Michael G (2000). „The four step model“. In : (cf. p. 71).

Mennis, Jeremy (2003). „Generating surface models of population using dasymetric mapping“. In : *The Professional Geographer* 55.1, p. 31–42 (cf. p. 11).

Munizaga, Marcela A et Carolina Palma (2012). „Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile“. In : *Transportation Research Part C : Emerging Technologies* 24, p. 9–18 (cf. p. 71).

Ni, Linglin, Xiaokun (Cara) Wang et Xiqun (Michael) Chen (2018). „A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data“. In : *Transportation Research Part C : Emerging Technologies* 86, p. 510–526 (cf. p. 72).

OSM (2018). *OpenStreetMap*. `http://openstreetmap.ord`. Online ; accessed June 2018 (cf. p. 29, 74).

Pang, Linsey Xiaolin, Sanjay Chawla, Wei Liu et Yu Zheng (2013). „On detection of emerging anomalous traffic patterns using GPS data“. In : *Data & Knowledge Engineering* 87, p. 357–373 (cf. p. 2, 18, 71).

Pelletier, Marie-Pier, Martin Trépanier et Catherine Morency (2011). „Smart card data use in public transit : A literature review“. In : *Transportation Research Part C : Emerging Technologies* 19.4, p. 557–568 (cf. p. 71).

Ratti, Carlo, Dennis Frenchman, Riccardo Maria Pulselli et Sarah Williams (2006). „Mobile landscapes : using location data from cell phones for urban analysis“. In : *Environment and Planning B : Planning and Design* 33.5, p. 727–748 (cf. p. 23, 39).

Reddy, Sasank, Min Mun, Jeff Burke et al. (2010). „Using mobile phones to determine transportation modes“. In : *ACM Transactions on Sensor Networks (TOSN)* 6.2, p. 13 (cf. p. 15).

Schneider, Christian M, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda et Marta C González (2013). „Unravelling daily human mobility motifs“. In : *Journal of The Royal Society Interface* 10.84, p. 20130246 (cf. p. 17).

Shang, Jingbo, Yu Zheng, Wenzhu Tong, Eric Chang et Yong Yu (2014). „Inferring gas consumption and pollution emission of vehicles throughout a city“. In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 1027–1036 (cf. p. 39).

Soto, Victor, Vanessa Frias-Martinez, Jesus Virseda et Enrique Frias-Martinez (2011). „Prediction of socioeconomic levels using cell phone records“. In : *User Modeling, Adaption and Personalization*. Springer, p. 377–388 (cf. p. 39).

STIF (2018). *Open Data STIF*. `http://opendata.stif.info`. Online ; accessed June 2018 (cf. p. 29, 74).

Tang, Lu-An, Yu Zheng, Jing Yuan et al. (2012). „On discovery of traveling companions from streaming trajectories“. In : *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, p. 186–197 (cf. p. 18).

Thiagarajan, Arvind, Lenin Ravindranath, Hari Balakrishnan, Samuel Madden, Lewis Girod et al. (2011). „Accurate, Low-Energy Trajectory Mapping for Mobile Devices.“ In : *NSDI* (cf. p. 15).

TIM (2014). *Telecom Italia Big Data Challenge 2014.* (accessed : 01.17) (cf. p. 40, 57, 65).

TOOLE, Jameson L, Serdar COLAK, Bradley STURT et al. (2015). „The path most traveled : Travel demand estimation using big data resources“. In : *Transportation Research Part C : Emerging Technologies* 58, p. 162–177 (cf. p. 1, 2, 10, 13, 15, 71, 72).

WANG, Feilong et Cynthia CHEN (2018). „On data processing required to derive mobility patterns from passively-generated mobile phone data“. In : *Transportation Research Part C : Emerging Technologies* 87, p. 58–74 (cf. p. 6, 8).

WANG, Huayong, Francesco CALABRESE, Giusy DI LORENZO et Carlo RATTI (2010). „Transportation mode inference from anonymized and aggregated mobile phone call detail records“. In : *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, p. 318–323 (cf. p. 1, 3, 9, 15, 71, 72, 91).

WANG, Ming-Heng, Steven D SCHROCK, Nate VANDER BROEK et Thomas MULINAZZI (2013). „Estimating dynamic origin-destination data and travel demand using cell phone network data“. In : *International Journal of Intelligent Transportation Systems Research* 11.2, p. 76–86 (cf. p. 1, 10, 13, 71).

WANG, Pu, Timothy HUNTER, Alexandre M BAYEN, Katja SCHECHTNER et Marta C GONZÁLEZ (2012). „Understanding road usage patterns in urban areas“. In : *Scientific reports* 2, p. 1001 (cf. p. 14).

WANG, Yihong, Gonçalo Homem de ALMEIDA CORREIA, Bart van AREM et HJP Harry TIMMERMANS (2018). „Understanding travellers' preferences for different types of trip destination based on mobile internet usage data“. In : *Transportation Research Part C : Emerging Technologies* 90, p. 247–259 (cf. p. 1, 71).

WANG, Yilun, Yu ZHENG et Yexiang XUE (2014). „Travel time estimation of a path using sparse trajectories“. In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 25–34 (cf. p. 18).

WESOLOWSKI, Amy, Nathan EAGLE, Andrew J TATEM et al. (2012). „Quantifying the impact of human mobility on malaria“. In : *Science* 338.6104, p. 267–270 (cf. p. 39).

WIDHALM, Peter, Yingxiang YANG, Michael ULM, Shounak ATHAVALE et Marta C GONZÁLEZ (2015). „Discovering urban activity patterns in cell phone data“. In : *Transportation* 42.4, p. 597–623 (cf. p. 17).

WU, Wei, Yue WANG, Joao Bartolo GOMES et al. (2014). „Oscillation resolution for mobile phone cellular tower data to enable mobility modelling“. In : *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*. T. 1. IEEE, p. 321–328 (cf. p. 8).

XU, Fengli, Pengyu ZHANG et Yong LI (2016). „Context-aware real-time population estimation for metropolis“. In : *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, p. 1064–1075 (cf. p. 11, 12, 39, 65, 66).

YAN, Xiao-Yong, Xiao-Pu HAN, Tao ZHOU et Bing-Hong WANG (2010). „Exact solution of gyration radius of individual's trajectory for a simplified human mobility model“. In : *arXiv preprint arXiv :1011.5111* (cf. p. 27).

YE, Yang, Yu ZHENG, Yukun CHEN, Jianhua FENG et Xing XIE (2009). „Mining individual life pattern based on location history“. In : *Mobile Data Management : Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*. IEEE, p. 1–10 (cf. p. 18).

YUAN, Jing, Yu ZHENG, Chengyang ZHANG, Xing XIE et Guang-Zhong SUN (2010). „An interactive-voting based map matching algorithm“. In : *Mobile Data Management (MDM), 2010 Eleventh International Conference on*. IEEE, p. 43–52 (cf. p. 15).

YUAN, Yihong et Martin RAUBAL (2014). „Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method“. In : *International Journal of Geographical Information Science* 28.3, p. 496–520 (cf. p. 18).

ZHENG, Yu et Xing XIE (2011). „Learning travel recommendations from user-generated GPS traces“. In : *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.1, p. 2 (cf. p. 18).

ZHONG, Gang, Xia WAN, Jian ZHANG, Tingting YIN et Bin RAN (2017). „Characterizing passenger flow for a transportation hub based on mobile phone data“. In : *IEEE Transactions on Intelligent Transportation Systems* 18.6, p. 1507–1518 (cf. p. 1, 16, 71).

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Estimating Urban Mobility with Mobile Network Geolocation Data Mining

Keywords : mobile network, geolocation, urban mobility, data mining

Abstract : In the upcoming decades, traffic and travel times are expected to skyrocket, following tremendous population growth in urban territories. The increasing congestion on transport networks threatens cities efficiency at several levels such as citizens well-being, health, economy, tourism and pollution. Thus, local and national authorities are urged to promote urban planning innovation by adopting supportive policies leading to effective and radical measures. Prior to decision making processes, it is crucial to estimate, analyze and understand daily urban mobility. Traditionally, the information on population movements has been gathered through national and local reports such as census and surveys. Still, such materials are constrained by their important cost, inducing extremely low-update frequency and lack of temporal variability. On the meantime, information and communications technologies are providing an unprecedented quantity of up-to-date mobility data, across all categories of population. In particular, most individuals carry their mobile phone everywhere through their daily trips and activities.

In this thesis, we estimate urban mobility by mining mobile network data, which are collected in real-time by mobile phone providers at no extra-cost. Processing the raw data is non-trivial as one must deal with temporal sparsity, coarse spatial precision and complex spatial noise. The thesis addresses two problematics through a weakly supervised learning scheme (i.e., using few labeled data) combining several mobility data sources. First, we estimate population densities and number of visitors over time, at fine spatio-temporal resolutions. Second, we derive Origin-Destination matrices representing total travel flows over time, per transport modes. All estimates are exhaustively validated against external mobility data, with high correlations and small errors. Overall, the proposed models are robust to noise and sparse data yet the performance highly depends on the choice of the spatial resolution. In addition, reaching optimal model performance requires extra-calibration specific to the case study region and to the transportation mode. This step is necessary to account for the bias induced by the joined effect of heterogeneous urban density and user behavior. Our work is the first successful attempt to characterize total road and rail passenger flows over time, at the intra-region level, using mobile network data. Although additional in-depth validation is required to strengthen this statement, our findings highlight the huge potential of mobile network data mining for urban planning applications.

Titre : Estimation de la Mobilité Urbaine par l'Exploitation des Données de Géolocalisation de Téléphonie Mobile

Mots clés : réseaux mobile, géolocalisation, mobilité urbaine, exploration de données

Résumé : Dans les prochaines décennies, la circulation et les temps de trajets augmenteront drastiquement en raison du fort taux d'accroissement de la population urbaine. L'augmentation grandissante de la congestion sur les réseaux de transports menace le bon fonctionnement des villes à plusieurs niveaux, tels que le bien-être des citoyens, la santé, l'économie, le tourisme ou la pollution. Ainsi, il est urgent, pour les autorités locales et nationales, de promouvoir l'innovation pour la planification urbaine, à l'aide d'une politique de soutien à l'innovation et de prises de mesures radicales. Pour guider les processus de décisions, il est crucial d'estimer, analyser et comprendre la mobilité urbaine au quotidien. Traditionnellement, les informations sur les déplacements des populations étaient collectées via des rapports nationaux et locaux, tels que les recensements et les enquêtes. Toutefois, ces derniers ont un coût important, induisant une très faible fréquence de mise-à-jour, ainsi qu'une temporalité restreinte des données. En parallèle, les technologies de l'information et de la communication fournissent une quantité de données de mobilité sans précédent, au jour le jour, toutes catégories de population confondues. En particulier, les téléphones portables accompagnent désormais la majorité des citoyens lors de leurs déplacements et activités du quotidien. Dans cette thèse, nous estimons la mobilité urbaine par l'exploration des données du réseau mobile, qui sont collectées en temps réel, sans coût additionnel, par les opérateurs télécom. Le traitement des données brutes est non-trivial en raison de leur nature sporadique et de la faible précision spatiale couplée à un bruit complexe. La thèse adresse deux problématiques via un schéma d'apprentissage faiblement supervisé (i.e., utilisant très peu de données labélisées) combinant plusieurs sources de données de mobilité. Dans un premier temps, nous estimons les densités de population et le nombre de visiteurs au cours du temps, à une échelle spatio-temporelle relativement fine. Dans un second temps, nous construisons les matrices Origine-Destination qui représentent les flux totaux de déplacements au cours du temps, pour différents modes de transports. Ces estimations sont validées par une comparaison avec des données de mobilité externes, avec lesquelles de fortes corrélations et de faibles erreurs sont obtenues. Les modèles proposés sont robustes au bruit et à la faible fréquence des données, bien que la performance des modèles soit fortement dépendante de l'échelle spatiale. Pour atteindre une performance optimale, la calibration des modèles doit également prendre en compte la zone d'étude et le mode de transport. Cette étape est nécessaire pour réduire les biais générés par une densité urbaine hétérogène et les différents comportements utilisateur. Ces travaux sont les premiers à estimer les flux totaux de voyageurs routiers et ferrés dans le temps, à l'échelle intra-régionale, à l'aide des données mobile. Bien qu'une validation plus approfondie des modèles soit requise pour les renforcer, nos résultats mettent en évidence l'énorme potentiel de la science des données de réseaux mobiles appliquées à la planification urbaine.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France