Inferring dynamic origin-destination flows by transport mode using mobile phone data

Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Jakob Puchinger, Mounim El Yacoubi

Inferring Dynamic Origin-Destination Flows by Transport Mode using Mobile Phone Data

Danya Bachir[a,c,*], Ghazaleh Khodabandelou[b,c], Vincent Gauthier[c,**], Mounim El Yacoubi[c], Jakob Puchinger[d,a]

[a] *IRT SystemX, Palaiseau France*
[b] *CNRS LPTMC, Sorbonne Universités UPMC, Paris France*
[c] *CNRS SAMOVAR, Telecom SudParis, Université Paris Saclay, France*
[d] *LGI, CentraleSupélec, Université Paris-Saclay, France*

Abstract

Fast urbanization generates increasing amounts of travel flows, urging the need for efficient transport planning policies. In parallel, mobile phone data have emerged as the largest mobility data source, but are 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 dynamic Origin-Destination flows by transport modes using mobile network data e.g., Call Detail Records. For this study, we pre-process 360 million trajectories for more than 2 million devices from the Greater Paris as our case study region. The model combines mobile network geolocation with transport network geospatial data, travel survey, census and travel card data. The transport modes of mobile network trajectories are identified through a two-steps semi-supervised learning algorithm. The later involves clustering of mobile network areas and Bayesian inference to generate transport probabilities for trajectories. After attributing the mode with highest probability to each trajectory, we construct Origin-Destination matrices by transport mode. Flows are up-scaled to the total population using state-of-the-art expansion factors. The model generates time variant road and rail passenger flows for the complete region. From our results, we observe different mobility patterns for road and rail modes and between Paris and its suburbs. The resulting transport flows are extensively validated against the travel survey and the travel card data for different spatial scales.

Keywords: Mobile phone data, Origin Destination Matrix, Transport mode, Urban mobility, Travel flows, Machine Learning

1. Introduction

In the upcoming decades, travel flows 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 transport planning innovation by adopting supportive policies leading to effective measures. Prior to decision making processes, it is crucial to estimate, analyze and understand daily urban mobility. Traditionally, information on population mobility has been gathered through national and local reports such as census and surveys. Thus, traditional transport planning models, such as four steps and activity based models, extensively rely on travel surveys (McNally, 2000; Bhat and Koppelman, 1999). However, surveys are constrained by their important cost, inducing extremely low-update frequency and lack of temporal variability. In particular, surveys generally report one day of trips per individual, which is not sufficient to capture all the temporal variations in mobility (e.g., seasonality, weekly patterns). In recent years, public transport operators have been

*Corresponding author: danya.bachir@gmail.com
**Corresponding author: vincent.gauthier@telecom-sudparis.eu

collecting daily travel card data (Pelletier et al., 2011; Ma et al., 2013; Munizaga and Palma, 2012). In most urban areas, multiple transport operators are in charge of public transport. Each operator possesses mobility data on its own transport network. Therefore, transport operators usually lack a global picture of real-time travel flows on multimodal transport networks. Such knowledge could be a valuable asset for transport planning, to evaluate the impact of transport policies on urban mobility (e.g., evaluate the evolution of the market share of public transport against private vehicles), predict the effect of perturbations (e.g., congestion, public transport interruption, public transport strikes, road closure, meteorological events etc.) and design new mobility services from the analysis of urban mobility.

On the meantime, mobile phone data has become the largest mobility data source as most individuals carry their mobile phone everywhere through their daily trips and activities. In particular, the pervasiveness and high penetration rates of mobile networks enable mobile phone operators to collect unprecedented quantity of up-to-date geolocation data from Call Detail Records (CDR), across all categories of population, at no additional cost. Several research works have described the potential of mobile network data for mobility analysis (Chen et al., 2016; Gadziński, 2018; Blondel et al., 2015). The most popular research areas are human mobility models (Gonzalez et al., 2008a; Pappalardo et al., 2015; Ni et al., 2018), travel demand modeling (Toole et al., 2015; Wang et al., 2013; Huang et al., 2018), itinerary reconstruction (Asgari et al., 2016; Becker et al., 2011), traveler behavior understanding (Calabrese et al., 2013; Wang et al., 2018; Ahas et al., 2010), population density estimation (Bachir et al., 2017; Khodabandelou et al., 2016, 2018), transport mode detection (Wang et al., 2010; Bachir et al., 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 et al., 2017; Chen et al., 2014). In addition, mobile network data offer the possibility to build day-to-day Origin-Destination (OD) matrices of flows (Çolak et al., 2015a; Iqbal et al., 2014; Alexander et al., 2015; Toole et al., 2015; Wang et al., 2010; Berlingerio et al., 2013a; Di Lorenzo et al., 2016; Ni et al., 2018; Aguiléra et al., 2014; Calabrese 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.

In this work, we present an end-to-end model for the estimation of dynamic Origin-Destination matrices by transport mode using mobile network data. An earlier version of this work briefly presents the transport mode inference model (Bachir et al., 2018). The main contributions of our work are listed below.

- This is the first study 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, multimodal 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 density of Greater Paris transport networks is among the highest worldwide, with an heterogeneous density between Paris and its suburb. Thus our model is generalizable to both high density and low density areas.

- Our transport mode inference model is semi-supervised as we rely on a small subset of labeled data for Base Transceiver Stations. Although mobile network geolocation is sparse and noisy, the mode inference is robust to both low data collection frequency and imprecise geolocation. A trajectory has a minimum of two distinct positions and no data filtering is required. Thus this is the first method identifying road and rail transport modes for all mobile network trajectories.

- The model estimates total flows for road and rail modes over time (e.g., per day, per hour), with Origin Destination aggregated at different spatial scales. From our results, we analyze the recent mobility patterns and modal shares in the region.

- For performance evaluation, extensive validation tests are conducted against two external transport data i.e., survey and travel cards. Estimates are validated with high Pearson correlations, reasonable absolute differences with the survey and small errors with travel cards data.

In Section 2, we review the literature on mobile network data, transport mode detection and OD matrices. The mobile network data and the case study are described in Section 3. Data pre-processing and transport mode inference are detailed in Section 4. The transport mode of trajectories is inferred through a two-steps semi-supervised model which identifies the trip mode among rail or road usage. 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 mode. The second step is the Bayesian inference of 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 operators market share, mobile phone flows are rescaled to the total population with expansion factors using census data. In Section 5 we summarize our main results. Finally, we perform the validation study against household travel survey and travel card data in Section 6.

2. Literature

In the past decade, several works addressed mobility related topics using geolocation data collected from smartphones and mobile networks. Two types of mobile network data have been described so far: communication records such as Call Detail Records (CDR) and mobile network records. The following section describes the traditional mobile network data, followed by a review of previous works on transport mode detection and Origin-Destination matrices estimation. In particular, we highlight limitations in past studies and provide comparison with our methodology.

2.1. Mobile Network Data Types

The first mobile network data type is communication events, or active events, used by mobile phone providers for billing purposes. Such data are traditionally collected in the CDR format and have been used in several mobility studies (Jrv et al., 2014; Calabrese et al., 2013; Dong et al., 2015). The CDR report communications between phones, including calls (i.e., voice communications, unanswered calling attempts), text messages and sometimes Internet usage. Each record contains the anonymized ID (aimsi) of the caller, and optionally of the callee, a timestamp with a duration, the ID of the telecommunication equipment (cell ID) connected to the device and the type of record i.e. incoming or outcoming, voice or text etc. The second data type is network records which are generated from an interaction between a device and the mobile network (Calabrese et al., 2015). Several records can be produced during a call or when the phone is not being used. The phone is said to interact 'passively' with the network while 'active' communications are collected in the CDR. Generally, network records have a higher frequency compared to CDR. Still, the sampling rate can vary depending on operators needs, material resources and legislation. In addition, mobile operators can combine CDR and network records to obtain a greater collection frequency. Such data is named Data Detail Records (XDR) (Graells-Garrido et al., 2018) or sightings (Chen et al., 2016; Wang and Chen, 2018) or signaling data (Huang et al., 2018).

When a device connects to the mobile network, it is located inside a signal area of a Base Transceiver Station (BTS). Such an area is called a network cell. Mobile phone positions are commonly approximated at the cellular scale which is coarse, with radii varying from hundred meters to several kilometers. Mobile network geolocation can have a finer-grained spatial resolution through triangulation (Calabrese et al., 2013; Wang and Chen, 2018; Alexander et al., 2015; Jiang et al., 2017). Triangulation requires signal frequencies of at least three nearby antennas to estimate the coordinates of 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.

In our work, we process mobile network geolocation from CDR and Location Area Updates (LAU) in France, where triangulation is prohibited. A record is retrieved at the beginning and end of a call, when a text is sent or received, when a data session starts and ends and when cellphones change their Location Area. Details on the mobile network data are provided in Section 3. (sightings)

2.2. Transport Mode Detection

Few research has been conducted on transport mode detection with mobile network data. Previous methods have employed map-matching on transport networks (Yuan et al., 2010; Asgari et al., 2016) and supervised learning algorithms (Gonzalez et al., 2008b; Reddy et al., 2010), which became popular initially with GPS data. Contrary to GPS positioning, mobile network geolocation is coarse, noisy and sparse. Two consecutive geolocation can be separated by long distances (from hundred meters to kilometers) and long time periods (ranging from seconds to hours). Consequently, mobile network trajectories are an imprecise and incomplete representation of users' real paths. Thus, traditional methods for transport mode detection implemented for GPS data are difficult to transpose. On the one hand, map-matching requires a substantive number of positions to find users' routes. Hence such technique is hardly generalizable to all mobile network trajectories which may contain a few geolocation points. On the other hand, 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 trajectories are unlabeled regarding transport modes. A transport classification of such data requires unsupervised or semi-supervised approaches. Among the literature on transport mode detection, few studies attempted unsupervised learning. 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. (2015) and Aguiléra et al. (2014) used indoor base stations inside Paris underground to identify underground flows from CDR. No additional modes were addressed in these works. Wang et al. (2010) used triangulated CDR from Boston area U.S., to identify two transport modes, road and public transport. Authors estimated Origin-Destination flows and applied a k-means clustering on travel times, followed by a comparison with Google travel times. Still, CDR frequency induces important incertitude 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 single 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.

In our work, we combine clustering and Bayesian inference to identify two transport modes, road and rail, for all mobile network trajectories in the Greater Paris. Trajectories are unlabeled as the real transport mode of each individual trajectory is unknown. To compensate this issue, we have extracted transport labels (road and rail) for a subset of mobile network areas. Consequently we consider our method as semi-supervised. Due to the absence of individual ground truth, a performance evaluation on individual trajectories is impossible. Instead, we perform a validation by comparing our estimated transport flows with two external datasets i.e., travel survey and travel card flows. Details on our transport mode identification model are provided in Section 4.

2.3. Origin-Destination Matrices

Mobile network data have been used to derive time-variant and pervasive OD matrices for large populations. Common applications are the estimation of travel demand (Wang et al., 2013; Toole et al., 2015; Huang et al., 2018), 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., 2013b), the determination of trips purposes (Alexander et al., 2015) and of weekly travel patterns (Calabrese et al., 2011), studying the effects of urban facilities and transport access on mobility (Ni et al., 2018), to name a few. Past studies on OD matrices constructed with cellular data share a common methodology. The first step is to identify cellphones trips. This is traditionally done by segmenting 'stay' records and 'moving' records. In past studies, segmentation algorithms compare the duration and distance between consecutive points to some thresholds (Wang et al., 2013; Jiang et al., 2013; Toole et al., 2015). The second step is the identification of the origin and destination of each trip, generally first and last visited cells. To form the OD matrix, trips are grouped by origin-destination and departure time. The last step is OD flows rescaling, converting mobile phone flows to total population flows.

Recent studies employed CDR from Dahka, Bengladesh (Iqbal et al., 2014), Boston and San Francisco, U.S. (Wang et al., 2012), Singapore (Jiang et al., 2017), triangulated sightings from Boston, US (Çolak et al., 2015b; Alexander et al., 2015) and XDR from Santiago, Chile (Graells-Garrido et al., 2018). For Dahka, OD flows were

first generated at the tower-to-tower level, a tower being a base station. The matrices were converted to the node-to-node scale on road networks. Each road node obtains an increment vote when a record occurs at a nearby tower. The tower-to-node conversion rule attributes the road node with highest vote over one month for a given phone. Flows were up-scaled using an optimization algorithm minimizing the difference with traffic counts. In our opinion, the tower-to-node conversion is unpractical for regions having high density of multimodal transport networks due to mobile networks coarse granularity. In other studies, OD flows are traditionally aggregated at the same scale as census and survey units. The state-of-the-art rescaling is performed by multiplying flows with expansion factors (Çolak et al., 2015b; Alexander et al., 2015; Jiang et al., 2017). For each user, the home location is identified as the location of longest stay duration during night time. Expansion factors are calculated as the ratio between census population and number of mobile phone subscribers living in the same area.

To the best of our knowledge, only the work of (Graells-Garrido et al., 2018) estimates OD matrices with transport modes using XDR. This study proposes an inference model for rail, car, bus, rail+bus and pedestrian modes at the city level, for commuters in Santiago, Chile. The model inputs a matrix of visited BTS frequencies (columns) for users (rows). The mode inference is based on a non-negative matrix factorization followed by k-means clustering. The study compares transport share from the XDR to the ones from travel survey. In the results, 2% of XDR trips are identified as pedestrian compared to 23% in the survey. Although the model makes strong assumptions on the pedestrian mode i.e., trips within a small distance from home and work, the results suggest that tracking walkers on mobile networks may be infeasible.

For our work, OD matrices are constructed with mobile network data for the Greater Paris (see Section 3). The three state-of-the-art steps, namely segmentation, origin-destination identification and rescaling are successively applied. Details for each step are provided in Section 4. Our approach shares a few similarities with the one of (Graells-Garrido et al., 2018) despite initial differences in data type and pre-processing. Both approaches for transport mode inference are semi-supervised, involving a subset of labeled base stations. Still, Graells-Garrido et al.'s model is different from ours on many aspects. We focus on three main dissimilarities. First, Santiago OD flows are generated for commuting trips, i.e., between users home and work places, which are not representative of the full spectrum of daily activities. Meanwhile our study identifies the mode for all trips involving all activities in order to estimate total OD flows per mode. Second, the Santiago model is static in time. The estimated modal shares correspond to the full observation period i.e., several weeks. On the contrary, our model is dynamic over time and generates OD flows per mode for any time slot resolution and thus can be used for real-time applications. Last, Graells-Garrido et al.'s estimates from 2016 are compared to the travel survey from 2012 using Spearman correlations and absolute differences. The study lacks a performance evaluation with ground truth data over an identical period. Our model performance is assessed through several indicators (confidence intervals of transport probabilities, modal balance index, robustness) and a detailed validation study. Our estimated OD flows are extensively compared against survey flows for different spatial scales. Eventually, we calculate errors with one month travel card flows.

3. Case Study Data

In this work, we process two-months of mobile network geolocation data from all subscribers of a specific mobile providers, living in the case study region. Our mobility study focuses on the Greater Paris which is among the densest areas in transport networks worldwide. In this section, we present the raw mobile network data, including its spatio-temporal characteristics. Then we describe the case study region and the different spatial scales used for the mobility analysis.

3.1. Mobile Network Data

Our main data are mobile network records (see Table 1) representing billions of rows each day (Terabytes). The mobile operator providing the data has a market share of 11.7% in France, at the time of the study. Records are collected for the Greater Paris region over a two months period during spring. 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 data sessions (i.e., IMSI attach/detach). When a mobile phone changes Location Area (LA), a data record is generated from a Location Area Update (LAU), occurring for mobility management of

Table 1: Example rows for our mobile network records

User ID	Timestamp	Sector ID	Type
9221959679262440000	2018-06-01 20:49:01	1500	Start Voice
9221959679262440000	2018-06-01 20:55:00	3452	End Voice
9221959679262440000	2018-06-01 21:13:05	4708	Data
9221959679262440000	2018-06-01 21:34:30	4708	Text

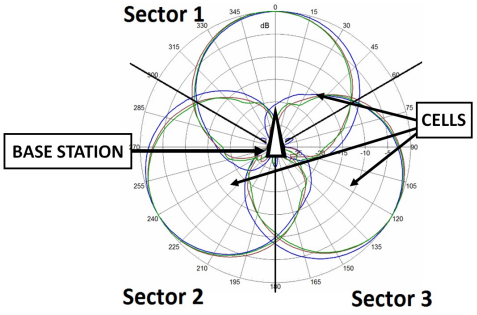

Figure 1: Tri-sector base station with three signal directions and nine overlapping cells

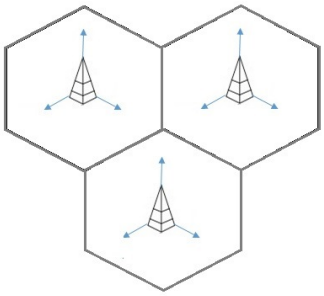

Figure 2: Classic voronoi centered on base stations

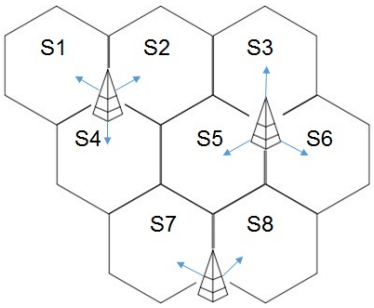

Figure 3: Our voronoi centered on mobile nework sectors

the mobile network. Location areas contain several base stations and are wide from several kilometers. At last, periodic location updates are recorded each 30 minutes. Such location updates occur in order to optimize the speed of signal transmission from the network in case of a new communication. With data records, the time interval between two distinct consecutive records decreases from several hours to a few minutes (4 min in median and 55 min in average). Although the LAU are passively generated, the data frequency remains moderate and is still dependent of the mobile phone usage i.e., when devices are being used (calls, sms).

In addition, mobile phone providers have no access to GPS coordinates of the devices. One way to estimate the coordinates of a mobile phone is to use triangulation. Yet this practice is currently unauthorized in France, except for emergency calls and authorities demands, as it is considered to bypass user consent and privacy. Instead, we use the raw geolocation of mobile phones on the mobile network. Each record is associated to a coarse signal area of the mobile network, traditionally a cell, surrounding a base station (see Fig. 1). Although mobile phones are located near base stations, it is extremely rare to encounter devices positioned exactly at the base station coordinates. Mobile phones can be anywhere inside the cells. Mobile network cells are represented by circular shapes with radii ranging 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. Cells constitute a multitude of overlapping areas. Consequently, we pre-process raw geolocation in order to merge overlaps. In order to obtain distinct non-overlapping areas, we use the direction angle of the antenna to create separate subdivisions around the base stations per signal direction. The resulting partitions are called mobile network sectors. Each record is associated to its corresponding sector position. In the literature, mobile network cells are sometimes represented as voronoï areas centered on base stations (see Fig. 2). For our case study, we rather use sectors which grant a finer spatial scale, as we have in average three sectors per base station. Therefore, we create sectors voronoï (see Fig. 3). Sectors centroids are calculated as the barycenter of cells centroids from the same sector. 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. Although mobile network geolocation is sparse, coarse and noisy, our data have higher spatio-temporal precision than classic CDR and are compliant with data legislation.

3.2. Case Study Region

This work focuses on the case study of the Greater Paris region which has 12 millions inhabitants and spans over 12000km^2. Different spatial scales are provided in the census, travel survey and mobile phone data. The spatial resolutions used for our study are presented in Table 2. The Greater Paris is subdivided into administrative areas of different levels. The three coarser areas are the city center formed by Paris, the first suburb ring and the second suburb ring. The region contains 8 departments. Paris corresponds to one department, the first ring consists of three departments while the second ring groups 4 departments. Rings and departments are represented in Fig. 4. In addition, the Greater Paris has 100 cantons and 1276 postcodes represented in Fig. 5. Postcode areas 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.

Table 2: Spatial scales and characteristics of spatial units in the Greater Paris.

Scale name	Nb. of units	Notations	Area range	Provider
Rings	3, including Paris	CC, R1, R2	$10^2 - 10^4$ km^2	Census & Travel survey
Departments	8, including Paris	CC, D2,...,D8	10^2 km^2	Census & Travel survey
Cantons	100	z_1,z_2,...z_{100}	10^1 km^2	Census & Travel survey
Postcodes	1382	None	$1 - 10^1$ km^2	Census
Mobile Network Sectors	7859	S_1,S_2,...,S_n	$10^1 - 10^2$ m^2	Mobile Operator

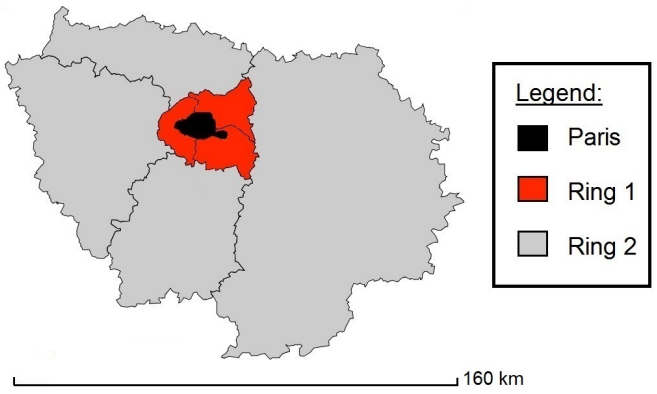

Figure 4: Greater Paris region for rings and departments scales

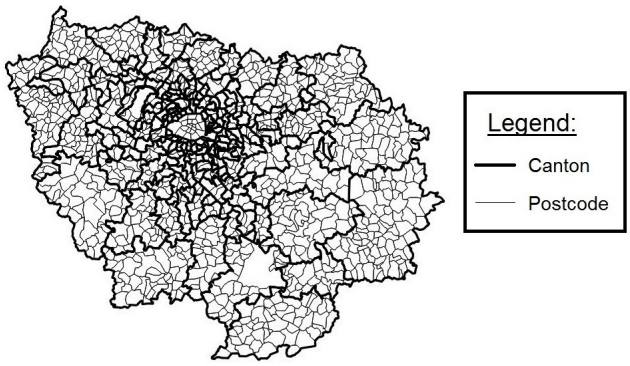

Figure 5: Greater Paris region for cantons and postcodes scales

4. Method

4.1. Overview

In this section, we present our method for OD flows estimation per transport mode using mobile network data. The model workflow is shown in Fig. 6 below. The first step of the model is the collection of the mobile network data presented in Section 3. The geolocation scale on the mobile network corresponds to mobile network sectors, described in Section 3.1. Anonymized mobile network records are pre-processed to generate trajectories as sequences of sectors, see Section 4.2. Then, we successively construct sectors features (see Section 4.4) and extract sectors transport labels (see Section 4.5) using transport networks. In this study, we perform a bi-modal separation between road and rail trips to infer transport flows. The mode inference is two-fold: we first perform clustering on mobile network sectors (see Section 4.6) followed by Bayesian inference used to calculate the transport probabilities for trajectories (see Section 4.7). 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 a mobile phone event is recorded, one knows the location of a device on mobile network sectors. The clustering aims at producing clusters of sectors grouped by transport usage. This step is equivalent to a transport land-use partitioning of the mobile network. Using a small labeled subset of sectors (e.g., base stations inside train stations, near highways etc.) we derive transport mode probabilities per cluster. A transport mode probability is assigned to each sector, depending on its cluster. Then, Bayesian inference is applied to 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 the highest probability is retained. Once transport modes are obtained, we construct modal OD matrices of flows (see Section 4.8). To evaluate the model performance we use several evaluation metrics, presented in Section 4.9. Results are provided in Section 5 followed by validation against travel cards and travel survey in Section 6.

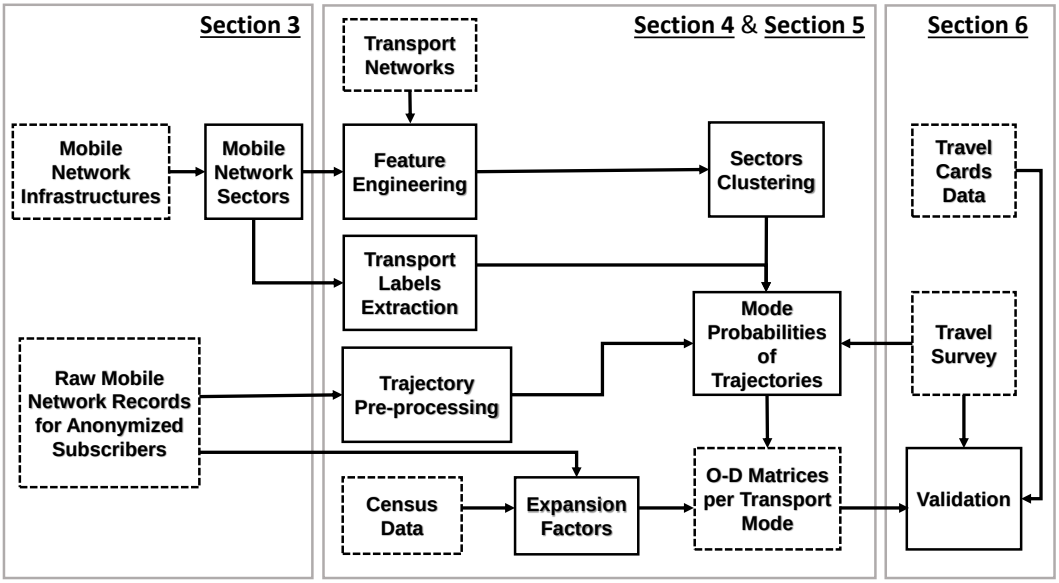

Figure 6: Workflow of the model for construction of OD matrices per transport mode

4.2. Trajectory Pre-Processing

Anonymized raw data are collected and pre-processed by the mobile phone provider. First, the operator filters oscillations identified as 'impossible jumps' (Wu et al., 2014). A mobile phone with three consecutive records detected in cells 'A', 'B' and back in 'A' (i.e., 'ABA') is assigned to position 'A' if the inter-event interval is below a certain time threshold (e.g., $\Delta t \leq 60$ s) and the speed is above some threshold (e.g., $\Delta v \geq 150$ km/h). Second, a smoothing algorithm is applied to strengthen noise reduction. Raw geolocations are smoothed with a weighted moving average, using the technique of (Csáji et al., 2013). Trajectory smoothing is followed by trajectory segmentation, also 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 lasts several minutes, 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 authors work, these steps are not further detailed in this study. After segmentation, moving points are grouped together to form a trajectory corresponding to one trip. 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 two months. Trajectories with at least 2 distinct moving positions are retained, since a single moving point could be noise. When comparing results

with the travel survey, we select users living in the Greater Paris region using their home location. Home locations are identified using the cells where phones are detected the most during night time.

4.3. Feature Construction with Transport Networks

A trajectory is a sequence of visited sectors for which we aim to find transport mode probabilities. In this perspective, we construct sectors features based on related spatial information between the mobile network and transport networks. The road networks are collected from OpenStreetMap (OSM, 2018). In order to reduce the computational cost of feature construction, we filter out residential roads. 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 OpenStreetMap. The following sectors features are constructed:

- $d_{j,road}$: shortest euclidean distance between the centroid of sector j and the road network. The shortest distance is the length of the segment formed by the sector centroid and the closest road point, which is perpendicular to the tangent to the road.
- $d_{j,rail}$: 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.

4.4. Feature Normalization

The range of the sector features is impacted by the densities of transport networks and mobile networks, which are both heterogeneous. Indeed, the city center benefits from a higher concentration of base stations with smaller sectors and denser transport networks. On the contrary suburbs have larger sectors with transport networks of lower densities. Thus, our strategy is to normalize each feature to unit norm, sector by sector, in order to reduce the bias induced by urban density over transport usage. For each sector j, each feature $d_{j,m}$ (distance to transport network for mode m) is divided by the sum of the distance features to all transport networks. Similarly each feature $N_{j,m}$ (intersection with transport network for mode m) is divided by the sum of the intersection to transport networks, for a given sector j. Train stations areas $A_{j,station}$ are divided by sector area A_j.

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

$$\widehat{N}_{j,m} = \frac{N_{j,m}}{\sum_i N_{j,i}} \in [0, 1] \tag{2}$$

$$\widehat{A}_{j,station} = \frac{A_{j,station}}{A_j} \in [0, 1] \tag{3}$$

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 $A_{j,station}$.

4.5. Label Extraction

For our work, we construct labels for a small subset of Base Transceiver Stations (BTS), see algorithm 1. First, we assess whether BTS coordinates are within a small distance to transport networks (e.g., 100 m). The mobile operator has the information whether the BTS are constructed outdoor or indoor. For indoor BTS, it is straightforward to infer that BTS matching rail networks are inside the underground or train stations and BTS matching roads are inside tunnels. Meanwhile outdoor BTS cover most roads and overground rails. For outdoor BTS, a label

is assigned in case there is only one mode in the sector (i.e., only roads or only rails). As a result, we obtain 4% sectors with rail labels and 11% sectors with road labels, hence a total of 15% transport labels for Greater Paris sectors. Initially, we use categorical transport labels i.e., $\{road, rail\}$ on our subset of sectors. Still, categorical transport labels are not appropriate for most sectors, such as outdoor equipment. Indeed, in dense urban areas such as the Greater Paris, the classic scenario is to encounter several transport modes inside an outdoor sector because of mobile networks coarse granularity. BTS constructed near roads or rail are not guaranteed to exclusively detect only one transport mode. When several transport networks are present in a sector, users could have taken any mode 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, using the prior knowledge of the categorical transport labeled subset. The maximal transport probabilities are $P \in \{0, 1\}$ and are restricted to indoor labeled BTS.

Input: Voronoi areas of mobile network sectors ;
Coordinates of Base Transceiver Stations ;
Transport networks for roads, rails and train stations ;
Output: Labels of sectors
foreach *sector j* **do**
  get the coordinates (x,y) of the BTS associated to j ;
  calculate d as the shortest euclidean distance between (x,y) and the closest transport network ;
  if $d < \epsilon$ **then**
    if *the BTS is indoor* **then**
      return the label corresponding to the closest transport mode
    else
      if *j has one transport mode* **then**
        return the label corresponding to this mode
      else
        j is unlabeled
      end
    end
  else
    j is unlabeled
  end
end

Algorithm 1: Label Extraction

4.6. Mobile Network Sectors Clustering

In order to find groups of sectors with similar transport usage we use an agglomerative hierarchical clustering. Clusters are merged according to an euclidean distance-based ward criterion, which minimizes the sum of squared errors. The optimal number of clusters is identified by minimizing the S_{dbw} validity index (Halkidi and Vazirgiannis, 2001). The S_{dbw} performs a trade-off between clusters compactness and separability. A small S_{dbw} grants smallest clusters dispersions and highest density of points around clusters centroids.
For each cluster k we calculate the score $p_{k,m}$ of a given transport mode $m \in \{rail, road\}$ (see Eq. 4). Such score is calculated as the proportion of labeled sectors in cluster k for a mode m, noted $L_{k,m}$, among the total number of labeled sectors for mode m, noted L_m. The score $p_{k,m}$ is normalized by the sum of transport scores for cluster k, for road and rail modes (see Eq. 5). This normalized score is the probability $P(m|S_{i,k}) \in [0, 1]$ of taking transport mode m given a visited a sector $S_{i,k}$ belonging to cluster k. The probabilities satisfy the condition: $\sum_j P(m_j|S_{i,k}) = 1$. Thus, the transport probabilities calculated for unlabeled sectors depend on their cluster membership. In addition, we update the probabilities of outdoor labeled sectors using Eq. 5 while indoor labeled sectors have maximum (or

minimum) transport probabilities in $\{0, 1\}$.

$$p_{k,m} = \frac{L_{k,m}}{L_m} \tag{4}$$

$$P(m|S_{i,k}) = \frac{p_{k,m}}{\sum_j p_{k,j}} \tag{5}$$

4.7. Inference of Trajectory Transport Mode

Bayesian inference is used to determine the main transport mode associated to a mobile phone trajectory. The probability $P(m|T_j^u)$ to take a transport 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, we have $P(T_j^u|m) = P(S_0, ..., S_l|m)$. An independence assumption between the probabilities to visit sectors given the mode is formalized: $P(S_i, S_{i+1}|m) = P(S_i|m)P(S_{i+1}|m)$. Thus, we have $P(T_j^u|m) = \prod_{i=0}^{l} P(S_i|m)$. The Bayes theorem is then recursively applied.

$$P(m|T_j^u) = \frac{P(T_j^u|m) * P(m)}{P(T_j^u)} = \frac{P(m)}{P(T_j^u)} \prod_{i=0}^{l} P(S_i|m) \tag{6}$$

Using Eq.5 we inject $P(m|S_i)$ to Eq. 6:

$$P(m|T_j^u) = \frac{\prod_{i=0}^{l} P(S_i)}{P(T_j^u)} P(m)^{1-l} \prod_{i=0}^{l} P(m|S_i) \tag{7}$$

The prior transport probability $P(m)$ is obtained from the travel survey and depends also 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. Finally we affect the mode obtaining the highest probability to each trajectory. Details of the transport mode inference algorithm are provided in Algorithm 2.

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 2: Transport Mode Inference

4.8. Origin-Destination Matrices

After modal inference, we construct OD matrices of flows which represent the total number of trips per mode. Each user trip corresponds to a cellphone trajectory on the mobile network. The sectors corresponding to first and

last record of the trajectory are considered as the origin and destination of a trip. A matrix is a 3-dimensional array noted $F = (f_{o,d,t})$, such as an element $f_{o,d,t}$ is the number of flows from origin location o to destination location d, 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 follows:

$$F^{tot} = \sum_{\substack{o,d,t \\ o \neq d}} f_{o,d,t} \tag{8}$$

$$F^{out}_{o} = \sum_{\substack{d,t \\ o \neq d}} f_{o,d,t} \tag{9}$$

$$F^{in}_{d} = \sum_{\substack{o,t \\ o \neq j}} f_{o,d,t} \tag{10}$$

The choice of the spatial granularity, for the aggregation of OD flows, 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. In addition, the noise inherent to mobile network geolocation also contributes to inaccurate origin and destination positions. For our matrices, we chose two levels of spatial aggregations: departments and postcodes. These scales are considered coarse enough to reduce the errors on the origin and destination. For each mode and for each day, the department OD matrix has 128 elements and the postcode OD matrix has $3.8 \cdot 10^6$ elements, considering the two ways of travel.

In addition, our mobile phone data corresponds to users from one mobile phone operator only. Therefore, we rescale flows up to the total population, using expansion factors (Alexander et al., 2015). Such 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 subscribers of the operator living in the same area. Population counts are obtained from the most recent census. Subscriber home locations are identified as the area of longest stay duration during night time. 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. In Section 5 and Section 6 we present model results using expansion factors calculated for departments.

4.9. 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.

4.9.1. Confidence Interval

In order to measure the separability between rail and road modes, the confidence interval $z^* \subset [0, 1]$ of the corresponding probability 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 $1 - \alpha = 0.95$.

4.9.2. 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.

$$\Delta_{BI}(A,B) = \frac{N_{A\to B,m}}{max(N_{B\to A,m},1)} - \frac{N_{B\to A,m}}{max(N_{A\to B,m},1)} \in [-1,1] \tag{11}$$

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.

4.9.3. Correlation with external data

The Pearson correlation coefficient r is used to calculate the correlation between mobile phone data results, noted x, and 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.

4.10. Comparison with travel card data

The NRMSE is used during validation, in order to compare the estimated rail flows to travel card data: NRMSE= $\frac{1}{\bar{x_i}}\sqrt{\frac{\sum_{i=1}^{N}(x_i-\hat{x_i})^2}{N}}$ where x_i is the vector of daily rail outflow over N days, $\hat{x_i}$ is the daily travel card outflow over the same period, $\bar{x_i}$ is the daily average travel card outflow and i is a postcode area.

5. Results

This section presents the main results obtained with our model. First, the results of the clustering on mobile network sectors are reported in Section 5.1. Second, results on transport mode estimation obtained with Bayesian inference are described in Section 5.2. Third, from the estimated OD matrices per transport mode, we analyze the Greater Paris mobility patterns in Section 5.3.

5.1. Transport Clustering of Mobile Network Sectors

A clustering algorithm is applied on mobile network sectors from the Greater Paris, as described in Section 4.6. The mobile network is dynamic as the signal strength of mobile network equipments is continuously updated for signal optimization and the number of base stations can also evolve in time. In this Section, we provide results for the mobile network sectors configuration of one given month (April 2018). The S_{dbw} minimization criterion is used to determine the optimal number of clusters k. The minimal S_{dbw} value is 0.305, achieved for $k = 9$. The transport probabilities are calculated for each of the nine clusters and given in Table 3.

Table 3: Transport Mode probabilities and cluster size for $k = 9$

Cluster	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Size	16.3%	7.04%	13.2%	19.9%	1.73%	2.25%	3.38%	17.4%	18.8%
P_{RAIL}	0.651	0.949	0.639	0.191	0.350	0.896	0.557	0.400	0.027
P_{ROAD}	0.349	0.051	0.361	0.809	0.650	0.104	0.443	0.600	0.973
Main mode	multimodal	rail	multimodal	road	multimodal	rail	multimodal	multimodal	road

Greater Paris sectors are represented on Fig. 7. Clusters are considered to be dominated by a mode when the probability for this 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 and C_8 are multimodal clusters with a higher probability for road. Clusters C_2 and C_6 are rail dominated clusters. Eventually, C_4 and C_9 are road dominated clusters. Clusters are equally present in the city center and the suburb, except for C_2 and C_6. 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. 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.

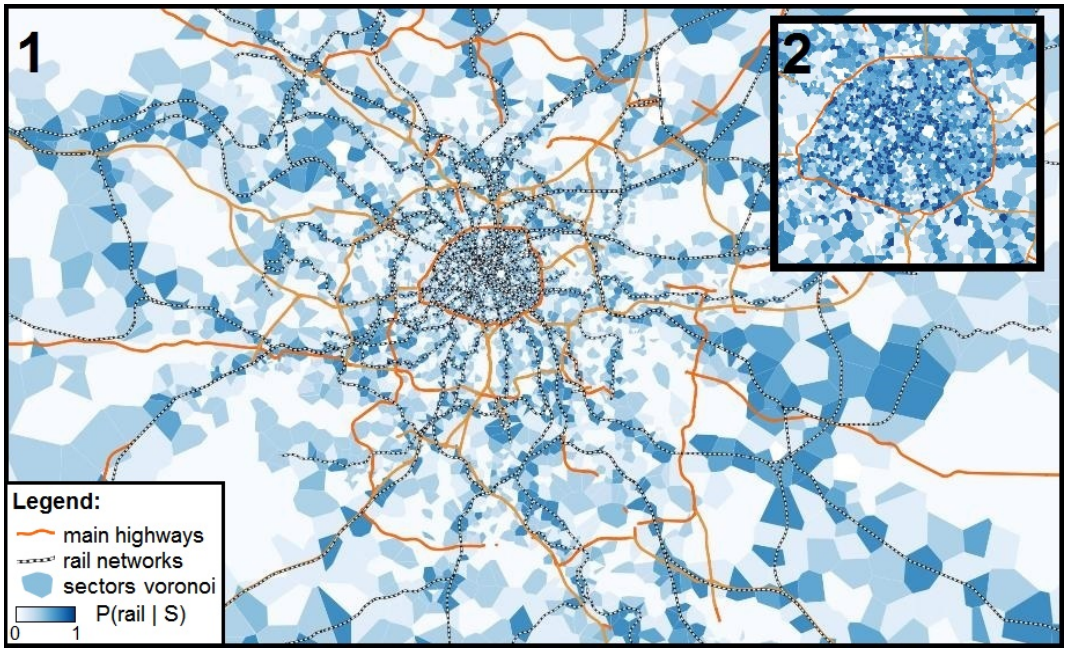

Figure 7: Sectors projected on the Greater Paris area (1) with a zoom on Paris (2). The color gradient gets a darker blue tone when the rail probability is high. Lighter sectors have higher road probabilities.

5.2. Performance Evaluation of Transport Mode Inference

5.2.1. Trajectories Probabilities

Using Bayesian inference, we derive the transport mode probability distribution of trajectories (see Fig. 8). 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 from C_3 and C_5 being multimodal.

Although half of sectors (52%) are considered multimodal, only 5% trajectories are categorized as uncertain. The reason is that multimodal sectors still have a dominant mode in their transport probabilities i.e., either the rail or the road probability is higher and probabilities are never identical. As a result, the transport probabilities of trajectories are well separated given the confidence interval z^*. Here, we observe the strength of the mode inference method through the decomposition of the trajectories into sectors. In case mobile phones are visiting several multimodal sectors, a minimum of one rail sector, or one road sector, can discriminate the mode probability. In addition, we assess the effect of the observed sectors compared to the influence of the prior, on the posterior probabilities $P(m|T)$, see Eq. 7. Such probabilities are determined by two terms. The first term is $P(m)^{1-l}$, where l is the number of sectors in T. The second term is the product of the probability of the mode given the visited sectors $P(m|S_i)$. After testing random values for $P(m)$, the results for $P(m|T)$ remain unchanged by a factor 10^{-3}. Consequently, the term $P(m)$ is not predominant in deciding $P(m|T)$ while the likelihood terms $P(m|S_i)$ determines the posterior $P(m|T)$. Thus, the observed trajectory has a dominant effect on the transport probability compared to any prior assumption, in the Bayesian inference scheme. This finding reveals that the model can be applied in urban areas for which travel survey information might not be available everywhere, for instance at the country level. The model can still be validated for a particular region benefiting from a survey.

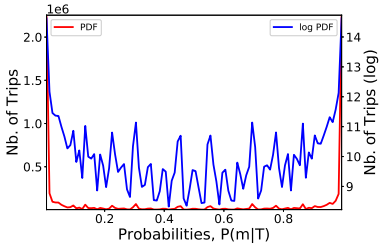

Figure 8: PDF of transport probabilities

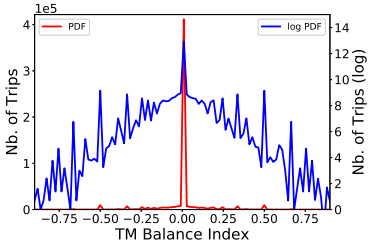

Figure 9: PDF of Balance Index Δ_{BI}

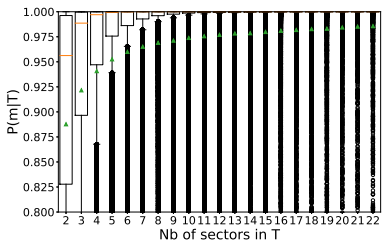

Figure 10: Boxplot for transport probabilities in function of sectors frequency

5.2.2. *Balance Index of OD Flows*

To further assess the performance of the transport mode inference, we calculate the Balance Index Δ_{BI} for OD flows, assessing whether the same mode is taken during round trips (see Fig. 9). 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 standard deviation 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. 9, we identify that non-zero $|\Delta_{BI}|$ values correspond to OD with lowest flows i.e., having small number of trips (less than 1000 trips a month), although there is no correlation between the two variables. This reveals that round trips having a mode-shift are found in areas with fewest travelers.

5.2.3. *Robustness of Mode Inference*

Eventually, we assess the robustness of the mode inference to low frequency in mobile network geolocation. First, we determine whether the actual number of observed sectors influence the transport probabilities. Thus, we visualize $P(m|T)$ in function of the number of visited sectors n_s in Fig. 10. The corresponding Pearson correlation is 0.22 showing there is a small correlation between $P(m|T)$ and n_s. The average and median probabilities gradually increase with the number of sectors. The probabilities converge to 1 for $n_s > 10$. For $n_s \geq 2$ the range of probabilities is above 0.8. showing that high transport probabilities can be obtained with few sectors i.e., few records. Second, we select a subset of trajectories, such as $n_s \geq 10$ and randomly delete records until the median value for n_s i.e., $n'_s = 4$. The same operation is also repeated until having only two sectors left i.e., $n'_s = 2$. The deletion of records at random enables to evaluate the effect of records frequency independently of the distances of real trips, as longer trips generally benefit from more records. The variation between the probabilities is calculated as $\Delta P = |P(m|T)_{n_s \geq 10} - P(m|T)_{n'_s}|$. For $n'_s = 4$, the median and standard deviation for ΔP are respectively 0 and ± 0.15, resulting in a mode changed for 5.1% trajectories, in total, over two months. For $n'_s = 2$, the median and standard deviation for ΔP are respectively 0 and ± 0.25, resulting in a mode changed for 11.3% trajectories. Consequently, the mode inference strategy appears robust to geolocation low frequency.

5.3. *Analysis of the Greater Paris Mobility*

5.3.1. *Visualization of Top Transport Flows*

Top passenger flows for rail and road modes are displayed in Fig. 11 and Fig. 12. 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. 13. 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. 14. Two areas attract most suburb flows (La Défense and Saint-Denis).

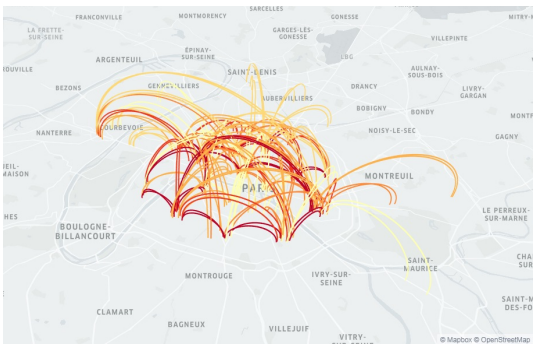

Figure 11: Top 100 rail passenger flows in the Greater Paris (zoom on Paris and the close suburb)

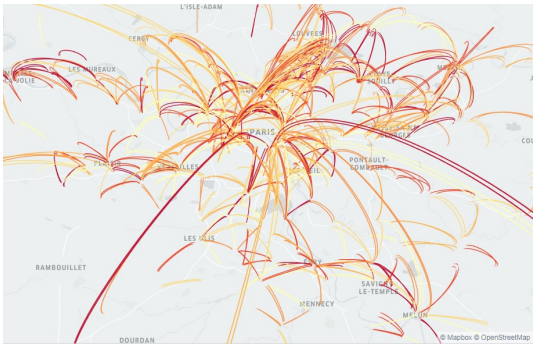

Figure 12: Top 100 road passenger flows in the Greater Paris, for trips having a distance $d > 5$ km

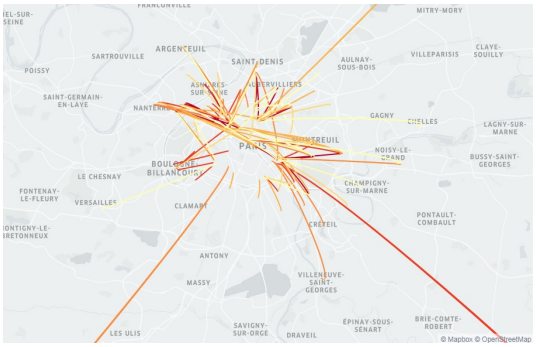

Figure 13: Top 100 rail passenger flows between Paris and the suburb

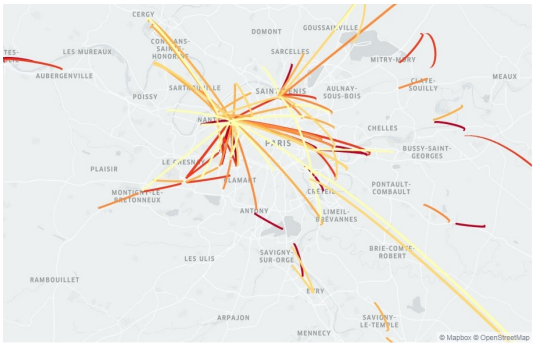

Figure 14: Top 100 rail passenger flows in the suburb, for trips having a distance $d > 5$ km

5.3.2. Temporal Patterns

Temporal patterns for transport flows are represented for a typical week, over a two month period, at the department scale. Daily average rail and road flows are shown in Fig. 16 and Fig. 18. Flows are averaged per week day, per start hour and per home department for rail mode in Fig. 15, and road mode in Fig. 17. For business days, peak hours occur in the morning and early evening. A midday peak can also be observed at lunch time. 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 Paris (department 75). On the contrary the number of road flows is higher in the evening, for any week day. The phenomenon is more pronounced for departments from the second suburb ring (i.e. departments 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. 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 flows. For rail mode there is a loss of 37% flows on Saturday and 52% on Sunday. For road mode the overall mobility loss is about 12% on Saturday and 24% on Sunday.

6. Validation

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

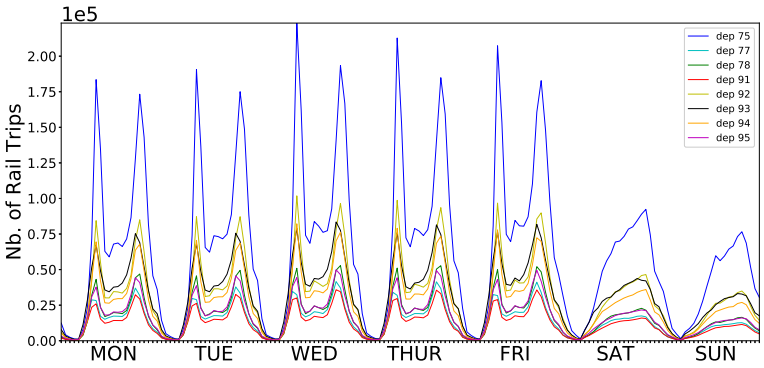

Figure 15: Weekly pattern for rail passenger flows per home department

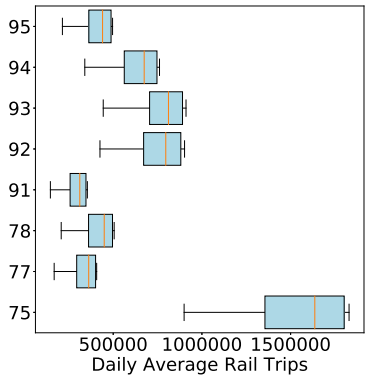

Figure 16: Boxplot for daily average rail flows per home department

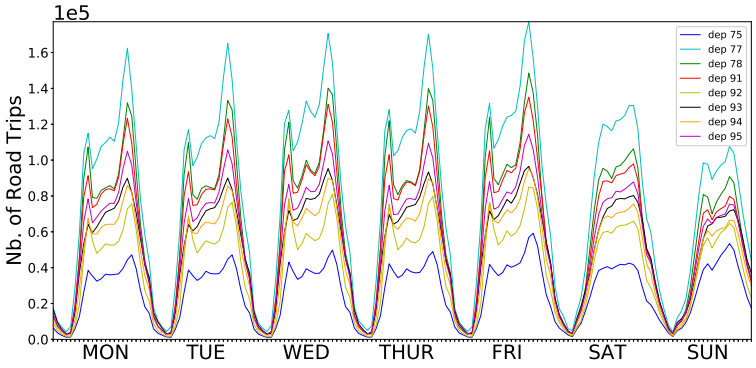

Figure 17: Weekly pattern for road passenger flows per home department

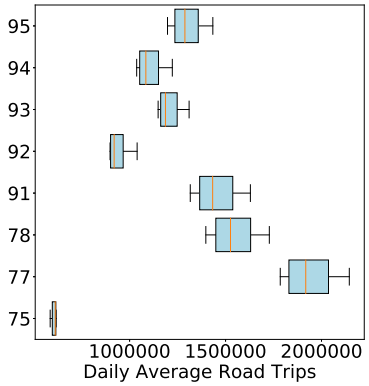

Figure 18: Boxplot for daily average road flows per home department

6.1. Comparison with the Travel Survey

The results with mobile phone data are provided for the period of April-May for the past year (2017) and compared with the latest travel survey (EGT, 2010) of the Greater Paris region, for year 2010. For this comparison study, we exclude week-ends and holidays, from the two datasets. The survey gathers responses from 43000 individuals among the 12 million residents of the Greater Paris. Transport modes are divided into two main categories. The first category is motorized modes, including public transport (e.g. underground, tramway, bus) and private vehicles (e.g. cars, motorbikes, taxi). The second category is unmotorized modes i.e, walk and bike.

In the Greater Paris, 1.5% of trips are realized by bike while 38.8% of trips are made by walkers, according to the survey. In particular, 99% of walk trips are shorter than 2 km. As a result, walkers have short trip distances, in addition to small speeds i.e., below 10 km/h. During CDR pre-processing, devices with low-speed are considered as non-moving. Thus, chances are high for walkers to be either undetected or considered in a non-moving state, hence not accounted in the travel flows. Besides, the double effect of large mobile network areas and geolocation low-frequency (i.e., non-active phone), strengthen the risk of undetected non-motorized trips. Therefore we rather use the CDR to estimate the total OD flows for motorized modes on roads and rails, which have higher speeds and longer trip distances.

In what follows, we compare our results to the survey for rail and road flows. In this perspective, we group underground, overground and tramway flows from the survey into survey rail flows. Similarly, private vehicles and bus flows from the survey are aggregated into survey road flows. Our comparison with the survey is achieved in three steps. First, we confront the total sum of transport flows in the region averaged per day and per hour. Second,

we benchmark our estimated OD flows, between zones, to the survey OD flows. At last, we consider the average number of trips performed each day per individual to compare the modal share per area.

6.1.1. Total transport flows

The total transport OD flows are estimated with our model, per day and per hour, for Greater Paris residents. Flows are grouped by departure hour and averaged over all days for each hour. As a result, we obtain the average transport OD flows per hour for a typical business day. Similarly, we collect the transport flows corresponding to a business day from the survey. First, the Pearson correlations between survey and mobile phone flows per hour are 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. 19. Secondly, we compare the absolute values for the sum of flows during 24 hours of a business day (see Table 4). In the year 2010, the survey reported 6.0 million rail flows for a business day. After rescaling MP rail flows, the average rail flows obtained is

Table 4: Total flows per transport mode in the Greater Paris for a typical business day. Flows are calculated with mobile phones (MP) before and after rescaling. In the column 'Survey' all road and rail trips from the 2010 survey are considered. In the column 'Survey*', we filter short-distance trips i.e., shorter than 1.5 km in suburb ring 1 and shorter than 2.5 km in suburb ring 2 to cope with the heterogeneous density and coarseness of the mobile network.

Mode	MP (raw)	MP (rescaled)	Survey	Survey* (filtered)
Rail flows	1227284	6383103	5999183	5843650
Road flows	2128750	11034581	18215180	11368597
$\frac{\text{Rail flows}}{\text{Road flows}}$	0.55	0.58	0.33	0.51

6.4 millions. Compared to year 2010, our results show a raise of +6.4% rail transport flows, during spring 2017. Comparatively, the Greater Paris transport authority reported a +10.9% annual rise for public transport trips in 2016 compared to 2010 (Source: Île-de-France Mobilités 2017). The 4.7% difference between our results and the transport authority estimates can possibly be caused by seasonal variations or by undetected mobile phones.
Meanwhile, after rescaling MP road flows, we find 11 millions road trips against 18.2 millions in the survey, for a business day. This corresponds to 39% of road flows being undetected from the CDR for a typical business day. Our findings reveal that a consequent part of road flows are undetected while our estimated rail flows are close to the survey by a few percents. One possible interpretation for this phenomenon is that mobile network cells are denser and smaller in Paris while the size of mobile network cells gradually increases in the suburb. Consequently, a subset of CDR trips occurring in the suburb might be too short compared to the mobile network scale and hence remain undetected.

Thus, we decide to filter trips shorter than 1.5 km in suburb ring 1 and shorter than 2.5 km in suburb ring 2 to cope with the coarseness and heterogeneity of the mobile network due to a heterogeneous urban density. As a result, we obtain a total of 11.3 million survey road flows. In comparison, our estimates show a loss of 3% road flows. For rail trips, filtering out the small-distance trips results in less than 1% difference. For the case study of the Greater Paris, we believe this reveals that the inter-distance between train stations is higher than the filtered distance thresholds. When comparing the difference of transport ratio $\frac{\text{Rail flows}}{\text{Road flows}}$, our results reveal a modal shift of 16% from road to rail mode since 2010, considering that small-distance trips are filtered out from the survey.
Compared to the survey, our estimates with CDR (see Fig. 19) reveal that flows during peak hours remained relatively identical for the rail mode (-2% between 7-9 AM and -1% between 4-7 PM) and have decreased for the road mode (-9% between 7-9 AM and -27% between 4-7 PM). Meanwhile off-peak hours flows have increased by 14% for rail mode and 5% for road mode between 9AM-4PM. In addition, we observe a rise of evening flows between 7-11 PM by 48% for rail mode and 28% for road mode. Then, we compare the evolution of the transport trends between the two surveys from 2010 and 2001. The 2010 survey 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. Thus, our findings for off-peak hour flows remain consistent with the trend announced in 2010. Meanwhile our results for 2017 suggest that road

and rail flows span over a longer period during the evening.

6.1.2. Origin-Destination Flows

In addition, we estimate the OD flows for rail and road modes at the ring scale (see Fig. 20). The Pearson correlation for OD flows are respectively 0.94 and 0.97 for rail and road flows. The absolute differences are calculated between our estimates and the survey in Table 5 to supplement Fig. 20. From our results we observe a raise for rail flows for trips inside Paris, trips from ring 1 to ring 1 and trips between Paris and ring 1. The highest raise for the rail mode concerns flows with an origin and a destination in ring 2 which are multiplied by a factor 2.5. Note that trips starting and arriving in suburb rings may pass by Paris. For the road mode, flows have been considerably reduced in Paris and between Paris and ring 1. However, road flows starting from or arriving to ring 2 have increased. Yet, as the road flows contain both private vehicles and public transports for bus mode, it is unclear whether public transports have increased or decreased. For instance multi-modal trips combining rail and bus modes are common in the region e.g., bus-rail-bus, bus-rail etc.

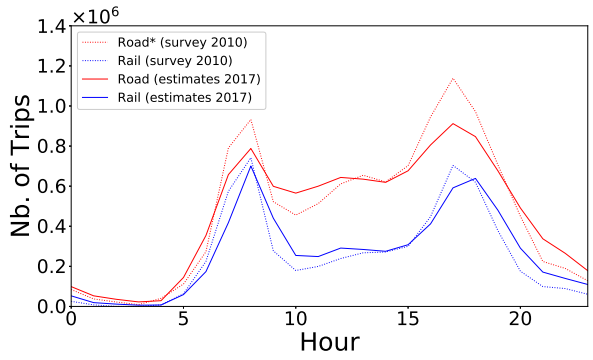

Figure 19: Daily pattern for survey and CDR flows during a business day. We have filtered survey road trips shorter than 1.5 km in ring 1 and shorter than 2.5 km in ring 2 to cope with the heterogeneous density and coarseness of the mobile network.

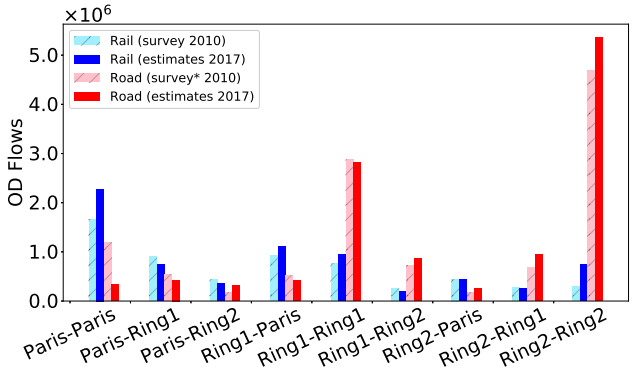

Figure 20: Average daily OD flows estimated with the survey and our model for road and rail modes between Greater Paris rings. We have filtered survey road trips shorter than 1.5 km in ring 1 and shorter than 2.5 km in ring 2.

Table 5: Absolute percentage difference on OD Flows between estimates and survey

OD	Paris-Paris	Paris-R1	Paris-R2	R1-Paris	R1-R1	R1-R2	R2-Paris	R2-R1	R2-R2
Rail	+37	−17	−20	+20	+25	−23	−2	−5	+151
Road	−72	−20	+73	−21	−2	+20	+47	+34	+27

6.1.3. Average day trips per person

The average daily trips per individual are calculated from the survey. The survey data is reported for 3 spatial resolutions. The coarser scale contains the three rings (i.e., Paris, ring 1 (R1) and ring 2 (R2)). The intermediate scale contains the eight departments. Paris forms its own department and the remaining departments are noted D2 to D8. The smaller survey scale corresponds to a partition of the region into 100 canton zones noted z_1 to z_{100}.
In the survey, an individual i has a weight w_i and reported $n_{i,m,t}$ trips for a mode m during a day t. The weights w_i are given in the survey. Such weights are assumed to be calculated with socio-demographic information (e.g., age, sex, job type, home address). The weights are used to rescale the subset of surveyed individuals to the entire population such as the sum of weights equals the total residential population. The function associating a home area j given an individual i is noted H such as $H(i) = j$. For the survey, the average number of trips per resident of an area j, for a mode m, is noted $C_m^{TS}(j)$ (see Eq. 12). In addition, for each mobile phone i, we calculate the number of trips $n_{i,m,t}$ per mode m for each day t during a period of T days, considering a total of U mobile phones. For a

mode m, the average number of trips per day and per mobile phone, with a home address in area j, is noted $C_m^{MP}(j)$ (see Eq. 13). At last, the ratio between the average road trips over the average rail trips per person in an area j is noted $C_{ratio}(j)$.

$$C_m^{TS}(j) = \frac{\sum_{i=1}^{k} w_i \cdot n_{i,m}}{\sum_{i=1}^{k} w_i} \quad \text{where } \forall i \; H(i) = j \tag{12}$$

$$C_m^{MP}(j) = \sum_{i=1}^{U} \sum_{t=1}^{T} \frac{1}{U} \frac{1}{T} n_{i,m,t} \quad \text{where } \forall i \; H(i) = j \tag{13}$$

$$C_{ratio}(j) = \frac{C_{road}(j)}{C_{rail}(j)} \tag{14}$$

The Pearson correlation coefficients are calculated between $C_m^{TS}(j)$ and $C_m^{MP}(j)$ using different spatial scales for the home area j (i.e., rings, departments and cantons). Results are shown in Table 6.

Table 6: Pearson correlation coefficients between the travel survey (TS) and mobile phones (MP) on average day trips per individual. Results are given considering different spatial scales for the home. The scales are the rings, the departments, the cantons, and cantons with suburb R2 filtered out.

Home Scale for j	$r(C_{Motor}^{TS}, C_{All}^{MP})$	$r(C_{Road}^{TS}, C_{Road}^{MP})$	$r(C_{Rail}^{TS}, C_{Rail}^{MP})$	$r(C_{Ratio}^{TS}, C_{Ratio}^{MP})$
Rings (CC, R1, R2)	0.993	0.995	0.990	0.999
Departments (CC, D2-8)	0.751	0.960	0.986	0.978
Cantons (z_1-z_{100})	0.466	0.931	0.874	0.764
Cantons ($\setminus$R2)	0.669	0.951	0.933	0.901

Table 7: Daily average trips per individual calculated for business days (source: EGT 2010-Île de France Mobilités-OMNIL-DRIEA)

Home Area j	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

When comparing survey motorized trips to all CDR trips, the highest correlation (0.99) is obtained for the ring scale, which is the coarsest. Meanwhile the smallest correlation (0.466) is obtained for the canton scale. When filtering out the cantons from the suburb ring R2, a higher correlation is obtained (0.669). The most probable cause of the lower correlation is a higher number of undetected trips in R2. A second possible explanation could be a sampling bias in the survey induced by a lack of surveyed individuals in cantons from R2. The rail and road modes achieve high correlations for all scales, ranging from 0.87 for cantons, up to 0.99 for rings. The obtained correlations reveal that the average numbers of trips per individual, calculated across different geographic areas, are consistent with the survey for road and rail modes.

The daily average trips per individual are provided at the region level and for the ring scale in Table 7. According to the survey, Greater Paris residents performed a total of 4.16 daily trips in 2010, among which 2.45 trips are realized using motorized modes. Meanwhile mobile phone users have an average of 2.1 day trips during spring 2017 (see table 7). For Paris residents, the daily average motorized trips for mobile phones users is identical to the survey: $C_{Motor}^{TS} \simeq C_{All}^{MP}$. Compared to survey motorized trips (C_{Motor}^{TS}), the value for mobile phones trips (C_{All}^{MP}) decreases for first and second suburb rings, respectively by -8% and -22%. Similarly, the daily average road trips witness a loss of respectively -13.2% for Paris residents, -22.5% for R1 residents and -30% for R2 residents.

When comparing the average daily trips, we observe an increase of rail trips and less road trips compared to the survey, as $C_{Rail}^{TS} < C_{Rail}^{MP}$ and $C_{Road}^{TS} > C_{Road}^{MP}$. Although we are aware of undetected road trips, the results suggest

an increase for rail transportation usage in the Greater Paris since 2010. In particular, we assume our results are the most reliable for Paris residents for which $C^{TS}_{Motor} \simeq C^{MP}_{All}$. Paris results show a global modal transfer of 13% of road trips in favor of rail transportation. Concerning the suburb, the results are biased by undetected trips, therefore the modal share transfer rate remains uncertain. Still, it is possible to estimate a modal share transfer considering

6.2. Validation with Public Transport Data

Greater Paris travelers swipe their travel cards when entering public transport, yet it is not required to swipe a second time when exiting the transport system. The validation dataset consists of daily entry counts of travel cards, swiped when entering train stations, for one month data (May 2017). Our model generates OD matrices containing daily and hourly rail flows between postcode areas, for the same month. Through this validation step, two datasets obtained from different sources, namely mobile phone 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. The rail outflows, calculated as the sum of travel card entry counts, is calculated for stations grouped by same postcode zone. Similarly, mobile phone flows are aggregated per day and per origin location at the postcode scale. For each day and each postcode, the daily outflows (i.e., the number of trips starting in the area) are obtained for both mobile phones and travel card holders.

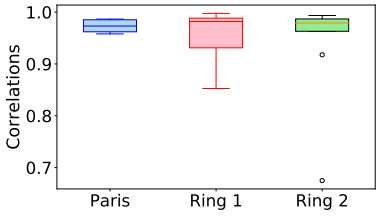

Figure 21: Correlations between daily MP rail outflows and travel card outflows

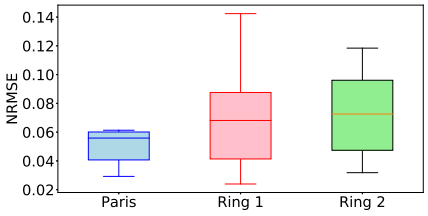

Figure 22: NRMSE between daily MP rail outflows and travel card outflows

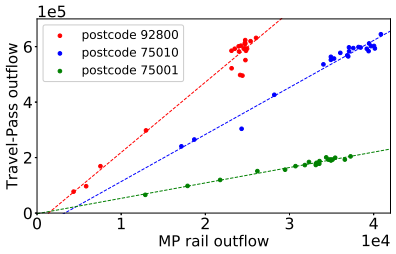

Figure 23: Regression between daily MP rail outflows and travel card outflows, for three examples postcode zones

First, Pearson correlation coefficients for daily rail outflows are calculated between our estimates and the validation data (see Fig. 21). The median correlation obtained is 0.98. The minimum correlation value is 0.68, appearing as an outlier for ring 2. This value corresponds to a major leisure area in the second ring, containing 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 are different from travel cards. Highspeed train passenger flows are accounted in our estimates while not accounted in the validation data, which explains the lower correlation for this postcode zone. The second outlier in the second ring corresponds to the zone with the biggest airport (roissy charles de gaulle) which also has a highspeed train station. Still, the corresponding correlation of 0.93 remains high.

Second, a linear relation is found between mobile phone rail outflows and travel card outflows (see Fig. 23). Consequently we apply linear regression models between the two datasets. The lowest NRMSE values are obtained after applying several linear regression models, one model being applied to each postcode area. The median NRMSE value is 0.062 (see Fig. 22). In comparison, with the state of the art rescaling method based on expansion factors, the median NRMSE value is 0.346. Here, the calibration with travel cards, aggregated per postcode zone, enables to account for bias specific to each train station. 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 departures and transfers. Meanwhile our OD estimates account for rail flows starting at the station. A second bias is fraud rates (i.e., travelers not swiping any travel card), fluctuating among train stations. At last, the travel card counts are not a perfect validation data as some technical problems and anomalies can affect the precision of this data. Still, the validation results regarding correlations and NRMSE show the performance of our model at the postcode scale for daily rail flows.

7. Discussion

In this paper, we propose the first methodology to estimate daily Origin-Destination flows of the total population, for rail and road transport modes, at the intra-region level. Mobile network geolocation, transport networks and travel survey information are jointly used in our transport mode inference model associating a mode to mobile network trajectories. The modal OD flows are rescaled using expansion factors calculated with both mobile network data and census data. The resulting OD matrices of transport flows present high correlations with travel survey flows and modal share, with reasonable absolute differences, despite minor dissimilarities since 2010. Although similar mobility trends are observed compared to the survey, such as a rise of flows during off-peak hours, a different pattern has also been revealed by our study which is the spreading of the evening-peak over a longer period, for both modes. Yet it remains unclear whether this observation is due to the difference of year between the datasets or to a seasonal effect, as our mobile phone data corresponds to two months in spring. In addition, our results unveil a raise of the rail transport mode usage for inter-ring trips (i.e., Paris-Paris, R1-R1, R2-R2). Meanwhile, we report an increase of road trips starting from or arriving in the farthest suburb (R2), given trip distances greater than 2.5 km. Still, the proportion of private vehicles and bus transportation involved in this increase remains unknown. To justify the growth in rail transport usage in the region we formulate two hypothesis. One is the construction of new transport lines in the region e.g., six new tramways and two rail lines expansions (M4 and RER E). A second hypothesis is the adoption of a unique fare for travel cards since year 2015, which may have encouraged public transport usage in the region.

For the validation with travel card data, we obtain both high correlations and small NRMSE. Compared to the state-of-the-art rescaling method, the validation error is reduced after calibrating the estimated rail outflows with the travel card outflows through regression models. This step is necessary to account for several bias in the travel card data i.e., pass-by flows, fraud rates and highspeed train tickets not accounted. The extensive comparison of our results with these two external transport datasets verifies the validity of our method at different spatial scales. The survey enables to assess the model for coarser scales such as rings, departments and cantons while the travel cards are used for postcode zones validation. The model can be reproduced by practitioners that have access to mobile network data and is generalizable to other areas for which transport networks, census and travel survey are available. In order to obtain the best results, mobile network data should be used jointly with travel surveys, public transport data and traffic counts, whenever possible, for effective calibration of OD flows at finer granularities.

Although our results stand for a good model performance, our work has several open issues. The first limitation of the method is that we consider a bi-modal separation into road and rail trips. This paper lefts aside the difficult task of separating private vehicles from road public transport users, such as bus passengers. This requires to have access to bus networks which are often shared with car routes. In addition, non-motorized modes by bike and walk are often preferred for short distance trips, thus those modes are hard to detect with coarse mobile network geolocation. A second limitation of this work is that we associate one mode to each trip while in real life scenarios, multimodal 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. Besides, these aforementioned issues remain open challenges due to the lack of up-to-date validation datasets for bus, multimodal flows etc.

Despite open issues, the presented work has several applications such as the evaluation of the impact of a transport policy on urban mobility, the determination of optimal locations for the construction of new transport infrastructures or studying the effects of particular events such as meteorological events, transport strikes, protests, sports events e.g., world cups, Olympic games etc. In particular, the modal OD matrices estimated with mobile network data could be used to strengthen traditional transport planing models such as the four step model during trip generation, trip distribution and mode choice. Therefore, we believe this work will help the transport community in planing travel demand, analyzing daily large-scale urban mobility, developing smart transport solutions and encourage the collaboration between transport authorities and mobile phone operators.

Acknowledgments

This research work has been carried out in the framework of IRT SystemX, Paris-Saclay, France, and therefore granted with public funds within the scope of the French Program Investissements dAvenir. Collection and pre-

processing of anonymized mobile network data were conducted at Bouygues Telecom, Meudon, France. The authors thank data scientists and data engineers from Bouygues Telecom Big Data Lab for their collaboration on data access.

8. References

Aguiléra, V., Allio, S., Benezech, V., Combes, F., Milion, C., 2014. Using cell phone data to measure quality of service and passenger flows of paris transit system. Transportation Research Part C: Emerging Technologies 43, 198 – 211. URL: http://www.sciencedirect.com/science/article/pii/S0968090X13002349, doi:https://doi.org/10.1016/j.trc.2013.11.007. special Issue with Selected Papers from Transport Research Arena.

Ahas, R., Aasa, A., Silm, S., Tiru, M., 2010. Daily rhythms of suburban commuters movements in the tallinn metropolitan area: Case study with mobile positioning data. Transportation Research Part C: Emerging Technologies 18, 45 – 54. URL: http://www.sciencedirect.com/science/article/pii/S0968090X09000400, doi:https://doi.org/10.1016/j.trc.2009.04.011. information/Communication Technologies and Travel Behaviour Agents in Traffic and Transportation.

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies 58, 240–250.

Asgari, F., Sultan, A., Xiong, H., Gauthier, V., El-Yacoubi, M.A., 2016. Ct-mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network. Computer Communications 95, 69–81.

Bachir, D., Gauthier, V., El Yacoubi, M., Khodabandelou, G., 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. pp. 626–632.

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, Mounim, V.a.E., 2018. Combining bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data (accepted), in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer.

Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C., 2011. Route classification using cellular handoff patterns, in: Proceedings of the 13th international conference on Ubiquitous computing, ACM. pp. 123–132.

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M., Blockeel, H., Kersting, K., Nijssen, S., Zelezny, F., 2013a. Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data., IBM Research, Dublin, Ireland. URL: http://search.ebscohost.com.gate6.inist.fr/login.aspx?direct=true&db=inh&AN=13840057&lang=fr&site=eds-live.

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M., Blockeel, H., Kersting, K., Nijssen, S., Zelezny, F., 2013b. Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data., IBM Research, Dublin, Ireland. URL: http://search.ebscohost.com.gate6.inist.fr/login.aspx?direct=true&db=inh&AN=13840057&lang=fr&site=eds-live.

Bhat, C.R., Koppelman, F.S., 1999. Activity-based modeling of travel demand, in: Handbook of transportation Science. Springer, pp. 35–61.

Biljecki, F., Ledoux, H., Van Oosterom, P., 2013. Transportation mode-based segmentation and classification of movement trajectories. International Journal of Geographical Information Science 27, 385–407.

Blondel, V.D., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. EPJ Data Science 4, 10.

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Computing 10, 0036–44.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies 26, 301–313.

Calabrese, F., Ferrari, L., Blondel, V.D., 2015. Urban sensing using mobile phone network data: a survey of research. Acm computing surveys (csur) 47, 25.

Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transportation Research Part C: Emerging Technologies 46, 326 – 337. URL: http://www.sciencedirect.com/science/article/pii/S0968090X14002022, doi:https://doi.org/10.1016/j.trc.2014.07.001.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies 68, 285–299.

Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015a. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. Transportation research record: Journal of the transportation research board , 126–135.

Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015b. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. Transportation Research Record: Journal of the Transportation Research Board , 126–135.

Csáji, B.C., Browet, A., Traag, V.A., Delvenne, J.C., Huens, E., Van Dooren, P., Smoreda, Z., Blondel, V.D., 2013. Exploring the mobility of mobile phone users. Physica A: Statistical Mechanics and its Applications 392, 1459–1473.

Demissie, M.G., de Almeida Correia, G.H., Bento, C., 2013. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transportation Research Part C: Emerging Technologies 32, 76 – 88. URL: http://www.sciencedirect.com/science/article/pii/S0968090X13000739, doi:https://doi.org/10.1016/j.trc.2013.03.010.

Di Lorenzo, G., Sbodio, M., Calabrese, F., Berlingerio, M., Pinelli, F., Nair, R., 2016. Allaboard: visual exploration of cellphone mobility data to optimise public transport. IEEE transactions on visualization and computer graphics 22, 1036–1050.

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., Zhou, X., 2015. Traffic zone division based on big data from mobile phone base stations. Transportation Research Part C: Emerging Technologies 58, 278–291.

EGT, 2010. Enquête Global Transport (EGT). http://www.omnil.fr/spip.php?article81. Online; accessed February 2018.

Gadziński, J., 2018. Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. Transportation Research Part C: Emerging Technologies 88, 74 – 86. URL: http://www.sciencedirect.com/science/article/pii/S0968090X18300366, doi:https://doi.org/10.1016/j.trc.2018.01.011.

Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L., 2008a. Understanding individual human mobility patterns. nature 453, 779.

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N.L., Perez, R., 2008b. Automating mode detection using neural networks and assisted gps data collected using gps-enabled mobile phones, in: 15th World congress on intelligent transportation systems.

Graells-Garrido, E., Caro, D., Parra, D., 2018. Inferring modes of transportation using mobile phone data. EPJ Data Science 7, 49.

Halkidi, M., Vazirgiannis, M., 2001. Clustering validity assessment: Finding the optimal partitioning of a data set, in: Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, IEEE. pp. 187–194.

Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., Wang, F.Y., 2018. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transportation Research Part C: Emerging Technologies 96, 251–269.

Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies 40, 63–74.

Jiang, S., Ferreira, J., Gonzalez, M.C., 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. IEEE Transactions on Big Data 3, 208–219.

Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr, J., Frazzoli, E., González, M.C., 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.

Jrv, O., Ahas, R., Witlox, F., 2014. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transportation Research Part C: Emerging Technologies 38, 122 – 135. URL: http://www.sciencedirect.com/science/article/pii/S0968090X13002301, doi:https://doi.org/10.1016/j.trc.2013.11.003.

Khodabandelou, G., Gauthier, V., El-Yacoubi, M., Fiore, M., 2016. Population estimation from mobile network traffic metadata, in: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A, IEEE. pp. 1–9.

Khodabandelou, G., Gauthier, V., Fiore, M., El Yacoubi, M.A., 2018. Estimation of static and dynamic urban populations with mobile network metadata. IEEE Transactions on Mobile Computing .

Larijani, A.N., Olteanu-Raimond, A.M., Perret, J., Brédif, M., Ziemlicki, C., 2015. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. Transportation Research Procedia 6, 64–78.

Ma, X., Wu, Y.J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders travel patterns. Transportation Research Part C: Emerging Technologies 36, 1–12.

McNally, M.G., 2000. The four step model .

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. Transportation Research Part C: Emerging Technologies 24, 9–18.

Ni, L., Wang, X.C., Chen, X.M., 2018. A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. Transportation Research Part C: Emerging Technologies 86, 510 – 526. URL: http://www.sciencedirect.com/science/article/pii/S0968090X17303601, doi:https://doi.org/10.1016/j.trc.2017.12.002.

OSM, 2018. OpenStreetMap. http://openstreetmap.ord. Online; accessed June 2018.

Pang, L.X., Chawla, S., Liu, W., Zheng, Y., 2013. On detection of emerging anomalous traffic patterns using gps data. Data & Knowledge Engineering 87, 357–373.

Pappalardo, L., Pedreschi, D., Smoreda, Z., Giannotti, F., 2015. Using big data to study the link between human mobility and socio-economic development, in: Big Data (Big Data), 2015 IEEE International Conference on, IEEE. pp. 871–878.

Pelletier, M.P., Trpanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, 557 – 568. URL: http://www.sciencedirect.com/science/article/pii/S0968090X1000166X, doi:https://doi.org/10.1016/j.trc.2010.12.003.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN) 6, 13.

STIF, 2018. Open Data STIF. http://opendata.stif.info. Online; accessed June 2018.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C: Emerging Technologies 58, 162–177.

Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies 87, 58 – 74. URL: http://www.sciencedirect.com/science/article/pii/S0968090X17303637, doi:https://doi.org/10.1016/j.trc.2017.12.003.

Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C., 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. pp. 318–323.

Wang, M.H., Schrock, S.D., Vander Broek, N., Mulinazzi, T., 2013. Estimating dynamic origin-destination data and travel demand using cell phone network data. International Journal of Intelligent Transportation Systems Research 11, 76–86.

Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. Scientific reports 2, 1001.

Wang, Y., de Almeida Correia, G.H., van Arem, B., Timmermans, H.H., 2018. Understanding travellers preferences for different types of trip destination based on mobile internet usage data. Transportation Research Part C: Emerging Technologies 90, 247–259.

Wu, W., Wang, Y., Gomes, J.B., Anh, D.T., Antonatos, S., Xue, M., Yang, P., Yap, G.E., Li, X., Krishnaswamy, S., 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, IEEE. pp. 321–328.

Yuan, J., Zheng, Y., Zhang, C., Xie, X., Sun, G.Z., 2010. An interactive-voting based map matching algorithm, in: Mobile Data Management (MDM), 2010 Eleventh International Conference on, IEEE. pp. 43–52.

Zhong, G., Wan, X., Zhang, J., Yin, T., Ran, B., 2017. Characterizing passenger flow for a transportation hub based on mobile phone data. IEEE Transactions on Intelligent Transportation Systems 18, 1507–1518.