TRANSIT: Fine-Grained Human Mobility Trajectory Inference at Scale with Mobile Network Signaling Data

Loïc Bonnetain[a,*], Angelo Furno[a], Nour-Eddin El Faouzi[a], Marco Fiore[b], Razvan Stanica[c], Zbigniew Smoreda[d], Cezary Ziemlicki[d]

[a] *University of Lyon, ENTPE, University Gustave Eiffel, LICIT, 25 avenue François Mitterrand, Lyon, France*
[b] *IMDEA Networks Institute, Avda del Mar Mediterraneo 22, Madrid, Spain*
[c] *Univ Lyon, INSA Lyon, Inria, CITI, 20 Avenue Albert Einstein, Villeurbanne, France*
[d] *Orange Labs, 44 avenue de la République, Chatillon, France*

Abstract

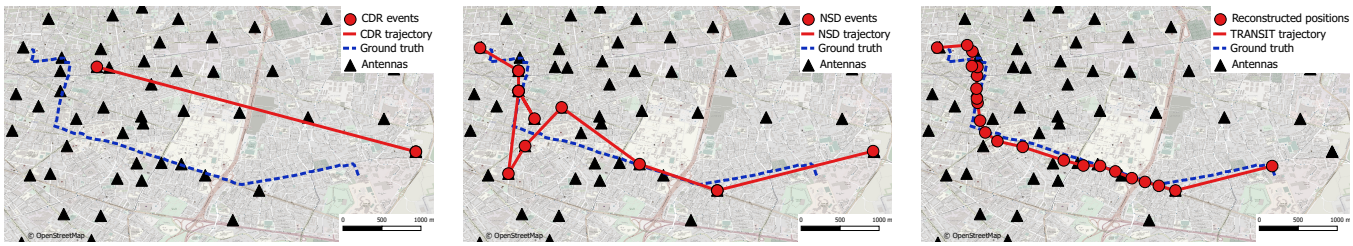
Call detail records (CDR) collected by mobile phone network providers have been largely used to model and analyze human-centric mobility. Despite their potential, they are limited in terms of both spatial and temporal accuracy thus being unable to capture detailed human mobility information. Network Signaling Data (NSD) represent a much richer source of spatio-temporal information currently collected by network providers, but mostly unexploited for fine-grained reconstruction of human-centric trajectories. In this paper, we present TRANSIT, TRAjectory inference from Network SIgnaling daTa, a novel framework capable of processing NSD to accurately distinguish mobility phases from stationary activities for individual mobile devices, and reconstruct, at scale, fine-grained human mobility trajectories, by exploiting, with a DBSCAN-based clustering approach, the inherent recurrence of human mobility and the higher sampling rate of NSD. The validation on a ground-truth dataset of GPS trajectories showcases the superior performance of TRANSIT (80% precision and 96% recall) with respect to state-of-the-art solutions in the identification of movement periods, as well as an average 190 m spatial accuracy in the estimation of the trajectories. We also leverage TRANSIT to process a unique large-scale NSD dataset of more than 10 millions of individuals and perform an exploratory analysis of city-wide transport mode shares, recurrent commuting paths, urban attractivity and analysis of mobility flows.

Keywords: `Mobile Phone Data, Human-Centric Mobility, Individual Trajectory, Big Data, Urban Computing`

1. Introduction

For decades, household surveys have been the only source of data to analyze and understand human-centric mobility, yet they are expensive to run, get quickly outdated, and are unavoidably based on relatively small samples of the population. Triggered by new technologies, this situation has now changed with the emergence of new data sources such as smart cards, the Global Positioning System (GPS), location-based social media, or mobile phone records, which all offer new possibilities to study individual and mass movement patterns. The research community has largely demonstrated the potential of these data in the context of mobility and transportation research [1, 2, 3], [4], where they allow analyses at unprecedented scales compared to traditional surveys [5].

*Corresponding author



(a) Trajectory from Call Detail Record (b) Trajectory from Network Signaling Data (c) Trajectory inferred with TRANSIT

Figure 1: Examples of inference of one trajectory of a volunteer from (a) CDR, (b) NSD, and (c) our NSD-based TRANSIT approach.

Among new mobility data sources, Call Detail Records (CDR) issued from mobile network operators are a privileged option for research, and have been employed to derive and validate general laws that govern human movements [6], reconstructing Origin-Destination (OD) matrices [7], understanding urban land use dynamics [8, 9], or inferring population density shifts in time [10]. Indeed, CDR present a unique combination of desirable properties: (*i*) they offer unprecedented penetration, as they are available for the whole subscriber base of a network provider, which typically covers tens or hundreds of millions of users; (*ii*) they are recorded continuously over long time periods, allowing fine-grained longitudinal studies over months or years; and, (*iii*) they are passively collected and maintained in curated databases for billing purposes, which makes them a very cost-efficient source of data for secondary use and analysis.

However, and despite their significant advantages for human-centric mobility studies, CDR have fundamental limitations in terms of positioning accuracy in both space and time. In space, the mobile device locations can only be mapped to the coverage area or position of the base stations to which it is associated [11]; in time, the sampling process is driven by the occurrence of voice call establishments or text message transmission, which are both sparse and irregularly distributed [12]. Ultimately, these problems limit the utility of CDR for studies that require a high level of spatiotemporal detail [13].

A prominent example of application with stringent needs in terms of spatial and temporal accuracy is the *inference of fine-grained human-centric trajectories in urban settings*. Here, the objective is reconstructing the separate trips of each mobile phone device with high accuracy to infer information about the exact paths traveled (*e.g.*, as the sequence of road segments, or the combination of transport modes). City environments exacerbate the problem, as they feature difficult-to-track short trips over entangled dense road layouts with multiple transportation modes. Traditional CDR are not suitable to address the task, due to their limited spatial resolution and sampling frequency. For instance, Figure 1a shows the localization samples recorded by CDR for an exemplary urban displacement; a linear interpolation of the CDR samples (solid red) is superposed to the actual user trajectory recorded via GPS (dotted blue). The figure makes it clear that inferring the actual movement from CDR is an arduous mission.

In this paper, we tackle the problem of fine-grained trajectory inference in urban areas using an emerging type of data from mobile networks, *i.e.*, Network Signaling Data (NSD), which, as discussed in Section 2, have been recently sparking research in the mobilitly modelling and transportation domains. These data capture control-plane events in the network, which are generated by all interactions with mobile devices that are needed for the operation

and management of the telecommunication system. NSD occur at much higher frequency than the sole call- and text-related events present in CDR: as an example, Figure 1b illustrates the more numerous NSD samples and the resulting, improved interpolated trajectory in the same case of Figure 1a. We provide complete details on NSD and how they compare to CDR are provided in Section 3. By using NSD, we provide the following main contributions.

- We present TRANSIT *(TRAjectory inference from Network SIgnaling daTa)*, a new framework that processes NSD to (*i*) tell apart movement intervals from stationary activity periods for each mobile device, and (*ii*) infer fine-grained human mobility trajectories during the associated movement intervals. The framework exploits the repetitive nature of human mobility *i.e.,* the same individual typically performs many trips between two same given locations over time, generally following very similar paths. This creates redundancy in the mobility information that TRANSIT uses to increase the spatiotemporal accuracy of the trajectories. TRANSIT hinges on the inherent high sampling rate of NSD to achieve an accurate and scalable reconstruction of the path followed by a device during each trip. Figure 1c shows the trajectory inferred with TRANSIT for the same exemplary trip of Figure 1a. TRANSIT is presented in Section 4.

- We validate TRANSIT with ground-truth GPS trajectories collected by a small set of volunteers, showing that it achieves 80% precision and 96% recall in the identification of movement periods, as well as an average 190 m spatial accuracy in the estimation of the trajectories. Comparisons with previous tools for the reconstruction of movements from mobile phone data also show gains in the order of 50%-70%. Details are in Section 5.

- We apply TRANSIT at scale, to the whole subscriber base of a major network operator in two major cities in France, Paris and Lyon. This lets us identify 480 million trajectories of over 10 millions of individuals during a period of three months in 2019 – and improve substantially the accuracy of 100 millions of those. We leverage such a unique information to carry out preliminary explorations of: (*i*) the fraction of trips using public transport versus other modes, (*ii*) the metropolitan-scale commuting paths, (*iii*) the attractivity of specific urban areas hosting special events, and (*iv*) the mobility patterns of trips passing through different sectors of the ring-way that surrounds the metropolitan area of Paris. To the best of our knowledge, we are the first to employ NSD to conduct mobility analysis at such a large scale. Details are in Section 6.

Conclusions from our work, a discussion of its limitations, and directions for future research are finally outlined in Section 7.

2. Related work

In the last two decades, CDR have been at the core of a large corpus of research related to reconstructing human mobility from large-scale passively collected data. These works have traditionally targeted the estimation of travel demand [14, 15, 16], [17], [18], the construction of signatures for automated identification of land use and urban fabrics [8, 19], the analysis of urban dynamics [20], the estimation of population density [21] and patterns discovery in human activities [6, 22]. However, despite their potential, CDR present inherent spatio-temporal biases and sparsity that have impeded their universal adoption for operational purposes related *e.g.*, to city planning and transportation.

Conversely, research has flourished around the challenges aimed at improving the quality of CDR-based approaches [23] for human mobility reconstruction and modelling.

Concerning the temporal dimension, several approaches have been proposed to exploit the repetitive nature of human activities, which can be captured via a sufficiently long observation of the same user over time. The general idea is to recover information from multiple observations of the user's communication activity and thus increase the generally low frequency at which mobile phone traces are normally available. Such methods are traditionally based on machine learning techniques [24] and rely on custom spatio-temporal distances to detect trajectory similarity [25]. The repetitive nature of human mobility has also been exploited with other sources of data. For instance, Choi *et al.* [26] use a deep learning approach to perform trajecotry prediction at individual level with bluetooth data, whereas Kim *et al.* [27] leverage trajectory clustering on GPS traces to infer spatial and temporal patterns at aggregate scale.

Regarding the spatial dimension, the geographical information associated to CDR usually comes only in the form of the coordinates of the base station to which the user is associated to when a mobile phone event is issued and logged. Traditionally, the geographical area assigned to each base station is roughly determined via Voronoi or other regular (*e.g.*, grid-based) tessellations of the mobile network topology and elected as the user's position whenever an event is logged at that base station. As a result, in the most traditional case of a Voronoi tessellation, the spatial resolution of CDR only depends on the density of base stations, ranging from hundreds of meters at best in dense urban areas to several kilometers in rural ones. Another important issue affecting both the spatial and the temporal dimension of CDR is represented by the oscillation phenomenon that traditionally characterizes cellular communications [28]. Since user association in mobile networks follows operator-specific schemes based on dynamic metrics such as received signal power or base station load, oscillations can easily take place between two or more antennas, even in absence of an actual mobility of the user. These characteristics add noise to the localization information that can be inferred from CDR data and make extremely hard the task of reliably discriminating between static and mobile sessions with CDR [29].

In the following, we focus on two approaches recently proposed in the literature to overcome location-related limitations, that represent the most related proposals to our solution, TRANSIT. Wu *et al.* [28] propose a framework, called DECRE, to remove oscillations from CDR and reduce spatial uncertainty for enhanced human mobility modeling. To that purpose, the authors adopt a heuristic-based approach composed of three major steps, namely *detect*, *expand* and *remove*. The first step identifies *suspicious sequences* of events that could be responsible for an oscillation, by detecting high-speed transitions between pairs of consecutive events recorded in the user CDR trace. The second step expands the previously identified *suspicious sequences* by also exploring the mobile phone activity of the user in a fixed time interval before and after the suspicious sequence. The last step consists in removing from the CDR trace those events that belong to the suspicious sequence identified as responsible for the oscillation. To this purpose, each antenna of a suspicious sequence receives a score that depends on its frequency of occurrence in the sequence and its average distance to the other antennas of the sequence. The events corresponding to the antenna with the highest final score are kept, while the others are filtered out. Although such filtering procedure could improve spatial accuracy by removing the noise deriving from the oscillations, the resulting trajectory is still bounded to the original location information from the cellular network (*i.e.*, antennas coordinates), thus exhibiting large spatial uncertainty.

A different strategy has been proposed in [30] and later improved by Bachir *et al.* [14], and applied to both a simple CDR dataset and a second one containing CDR enriched with location update events (a type of control traffic generated on the mobile network when a user moves over medium to long distances). Instead of filtering out the oscillations directly, the authors argue that these oscillations can be used to infer with increased accuracy user locations, by assuming that, if oscillations occur, the user should be, by triangulation, in the barycenter of these oscillation antennas. The approach, named *Cumulative Weighted Moving Average* (CWMA), consists on smoothing each mobile phone position by computing a weighted barycenter of all the consecutive antennas the user connects to within a given time-window. In particular, Bachir *et al.* [14] exploit the CWMA technique to segment the sequence of mobile phone events generated by a given user into a set of mobile and static sessions. To that purpose, the user speed is computed between consecutive smoothed positions obtained via CWMA; if the speed is below a certain threshold, the event is labeled as static. A static session is then defined if there is a sequence of events labeled as static and the duration of the session is superior to a given threshold. TRANSIT is built on the same assumption, *i.e.*, that averaging multiple locations from mobile phone events observed over related static or mobile sessions allows improving the spatial accuracy of the resulting trajectory. However, with TRANSIT we address two main limitations of CWMA: firstly, the moving average smoothing tends to excessively distort the reconstructed trajectories; secondly, both of the approaches proposed in [14, 30] do not take into account the existence of high regularity in human movements, and consequently in mobile phone events, that leads individuals to perform the same trips over time.

Other approaches, tested on small samples of CDR, aim at reducing the spatial inaccuracy by relying on map-matching methods [31, 32]. These methods match sequences of mobile phone events from the operator network to the nodes and edges of the transportation one, by relying on hidden Markov modeling. Despite promising results in terms of spatial accuracy, the computation time for processing a single mobile phone trace in urban environments with a dense transportation network is extremely high, thus making these approaches hard to scale to city-wide populations of mobile phone users. Some solutions, *e.g.*, [15], manage to assign trips extracted from CDR to the transportation network at scale, but require external information and assumptions, such as a route choice model.

Although CDR represent the type of passive mobile phone data most widely used in the literature, other species of mobile phone data have been used for mobility related purposes. Some studies [33, 34] address the low spatial granularity of CDR (or other kind of mobile phone data such as sighting data), by including precise user positioning data, obtained from the user radio signal information at multiple cellular base stations. This allows signal triangulation combined with spatial clustering [34] or the application of some probabilistic radio wave propagation models [33] for precise user position estimation. However, radio signal level information is difficult to obtain, as it is not regularly logged by mobile network operators. The same is true for approaches [35] based on the timing advance computed by the base stations for each user. While this information improves the user localization, it is not commonly logged by the network.

Some types of network signaling data have also been used in the literature. Ahas *et al.* [36] use Positium, an active data collection tool, which allows them to control the temporal granularity in their dataset. However, such tools are not commonly deployed by network operators and they are more intrusive from a privacy point of view than passive approaches that simply log the user activity. Janecek *et al.* [37] use handover and location

area update information for travel time estimation and map matching on the highway network. However, all these approaches tend to exhibit low performance in urban environments with a dense road network, due to the large set of similar alternative paths and low-resolution of the spatial information that is directly derived from the location of the antennas in the cellular network. Leontiadis *et al.* [38] achieve better results, but on a small mobile network signaling data, recorded by a smartphone application on a few tens of users. Recent studies included information regarding the user data connections [39] or even information regarding the increasingly popular machine-type communications [40], but without focusing explicitly on human-centric mobility. Zhao *et al.* [41] use large scale Internet access data and propose a machine learning approach to detect public/private transpotation mode.

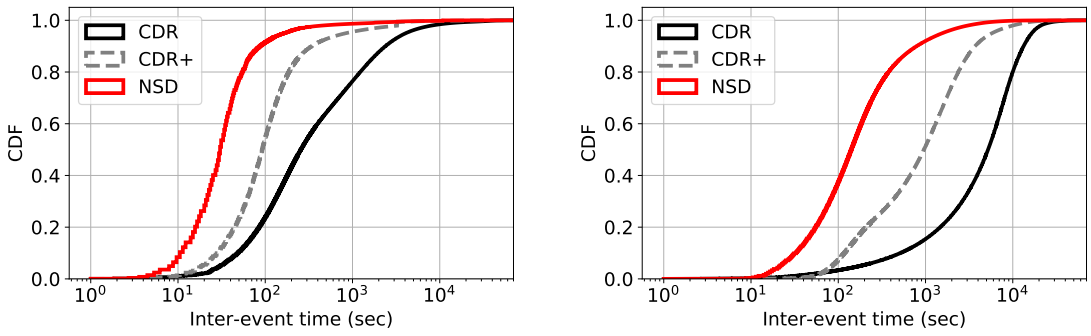
Very recently, some authors have started harnessing the potential of large-scale Network Signaling Data for different purposes. Qin *et al.* use NSD for sensing traffic conditions in urban networks [42] and making individual cellular usage prediction [43]. In [44], Zhao *et al.* compared different mathematical-based human mobility models from the literature by using NSD as ground truth. Such a study allows to improve the understanding of human mobility as well as providing tools for the simulation of mobility at both individual and population levels. The very recent work by Song *et al.* [45] is that closest to ours. The authors propose MIFF, a tool that leverages similar mobility patterns of individuals as a preliminary step before performing a map-matching of NSD to derive personal trajectories. With respect to this work, TRANSIT yields two key differences. First, while MIFF requires a database of users with similar trajectories, our framework overcomes that limitations by operating on the data of each user independently. This potentially also limits privacy concerns, as it enables inferring the accurate mobility of one target individual by accessing exclusively her data, and without any need to disclose or mix other people's sensitive mobility information. Second, TRANSIT does not involve any map-matching, which is known to be a computationally consuming task and still investigated by the research community [46]. This becomes a substantial advantage when scaling the method to very large populations as we do, and indeed we are the first to demonstrate NSD-based trajectory inference on millions of users in large metropolis.

3. Network Signaling Data

The Network Signaling Data (NSD) used in our study were collected in the production infrastructure of Orange, a leading mobile operator internationally and the largest telecommunications provider in France. We next present the data content and collection process, in Section 3.1, and then investigate their statistical properties, in Section 3.2. We provide a comparison of NSD against other mobile network data sources in order to contextualize our framework, and broaden the understanding of NSD, whose adoption is still at early stages.

3.1. Large-Scale Data Collection and Ethical Considerations

NSD include the network data-plane events generated by all devices associated with the Orange radio access network across 2G, 3G and 4G cellular technologies. NSD events are triggered by a variety of interactions: **(i)** voice and texting communications (*i.e.*, call establishments and SMS transmissions, which are fully equivalent to those logged by CDR), **(ii)** handovers (*i.e.*, device cell changes during communication), **(iii)** Location Area (LA) and Tracking Area (TA) updates (*i.e.*, cell changes that cross boundaries among larger regions named LA in 2G/3G and TA in 4G, affecting also idle devices), **(iv)** active paging (*i.e.*, periodic requests to update the location of the device



(a) Median (b) Average

Figure 2: CDF of inter-event times recorded in NSD, CDR, and CDR+. The plots refer to (a) median, and (b) average times per user.

started from the network side), **(v)** network attaches and detaches (*i.e.*, devices joining or leaving the network as they are turned on/off), and **(vi)** data connections (*i.e.*, requests to assign resources for traffic generated by mobile applications running on the device).

The NSD used in our work cover all Orange subscribers observed in two major metropolitan areas of France, *i.e.*, Paris and Lyon; in the following, the NSD datasets in the two cities are denoted by $\mathcal{D}_P$ and $\mathcal{D}_L$, respectively. The resulting total user base tallies to over 10 millions of individual mobile subscribers identifiers (IMSI) and over 3 millions of estimated residents in the two considered cities. The data were gathered during three consecutive months in 2019, from March 15 to June 15, including more than 150 billions of logged events overall, observed on a mobile phone network including more than 4,600 antennas. More details on the $\mathcal{D}_P$ and $\mathcal{D}_L$ NSD datasets are reported in Table 1.

The data from the Orange network probes used in this work were collected as part of the CANCAN - *Content and Context based Adaptation in Mobile Networks* collaborative research project founded by the French National Research Agency (ANR). The collection of this personal data has been authorized by the Data Protection Officer (DPO) of Orange according to article 89 of the General Data Protection Regulation (GDPR)[1], which provides an exemption for research, in particular for scientific and research purposes. The data were collected and processed exclusively on the Orange Labs secure Big Data platform. The data were pseudonymized and stored in a private directory in a server located in the operator premises, and accessible only to authorized researchers. All source data were deleted 12 months after the collection.

3.2. Comparison With Other Mobile Network Data Sources

The assortment of situations **(i)**–**(vi)** captured by NSD is much wider than the sole call- and text-related events in **(i)**; this naturally leads to a much higher sampling frequency of the locations of devices (hence, users) over time in NSD with respect to traditional CDR. Below, we investigate the added accuracy of NSD along the temporal and spatial dimensions.

[1] https://gdpr.eu/tag/gdpr/

3.2.1. Temporal accuracy

A quantitative inspection of the increased temporal accuracy of NSD is provided in Figure 2. The two plots present Cumulative Distribution Functions (CDF) of the time between subsequent NSD events; specifically, the distributions are computed over the (a) median and (b) mean inter-event time recorded for each device, hence they provide a fair view of the statistics across the observed population. We also report equivalent CDF obtained using other kinds of mobile network data: (*i*) CDR, which, as already mentioned, only capture voice and texting communication events in **(i)**, and (*ii*) CDR augmented with LA and TA update events in **(iii)**, which we term CDR+. The rationale is that CDR are the most widely adopted source of data from mobile networks, whereas CDR+ have been previously used for human mobility trajectory inference in the literature [47]. We directly extrapolated CDR and CDR+ from the available NSD database, by simply retaining only the spatiotemporal samples generated by the events that are captured by such data sources (*i.e.*, types **(i)**, and **(i)**+**(iii)**, respectively), while filtering out the information associated to all other network event types.

The distributions in Figure 2 yield a number of interesting observations. NSD grants a median inter-event time below 1 minute for 90% of the users, while that figure grows to 5 minutes for CDR+ and over 30 minutes for CDR. Per-user averages that are biased by long inactivity periods highlight even more the difference between the data sources: NSD keeps averages below 15 minutes for 90% of the users, whereas CDR+ and CDR record mean inter-arrivals of up to 1 hour and 3.5 hours for the same user fraction. Similar considerations hold for users with very heterogeneous levels of network activity, as the CDF remain neatly separated across the whole domain in abscissa. The conclusion is that NSD ensure a sampling rate increase of more than one order of magnitude with respect to CDR and of a factor 5 over CDR+. Importantly, these results are fairly uniform over the considered population.

3.2.2. Spatial accuracy

NSD do not bring any advantage over other classes of mobile network positioning data in terms of the absolute spatial accuracy *of each location sample*. As a matter of fact, NSD, CDR, CDR+, and any other network data types, are collected on the same radio access network infrastructure: therefore, the locations used to geo-reference the events are those of a matching set of base stations to which mobile devices associate over time. To prove our point, we run experiments with ground truth GPS data collected by a small set of volunteers, described in details later in Section 5. For each volunteer, we compute the distance between the location of the antenna associated to all generated network events and the corresponding GPS position at the time. Repeating the process for all CDR, CDR+ and NSD events yields very similar average distances, between 0.26 and 0.28 km, in the three cases.

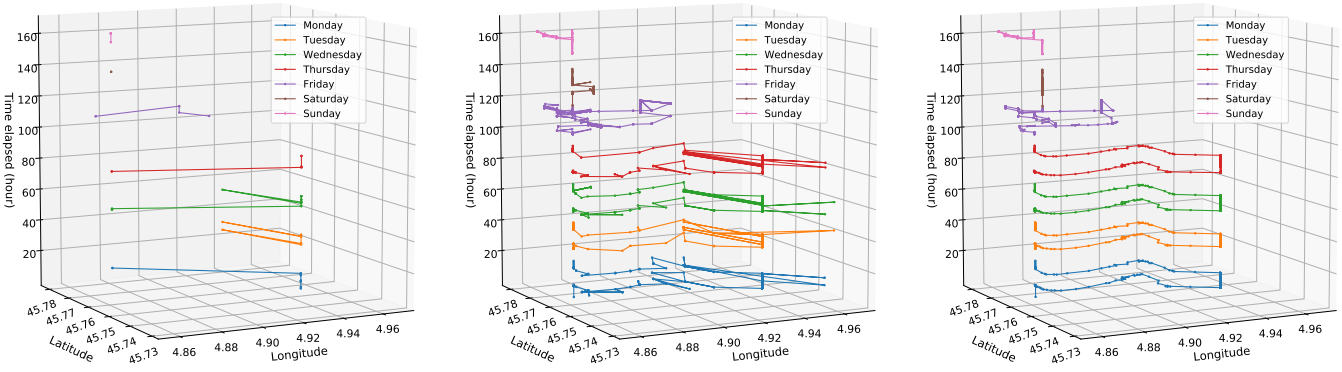
However, NSD provide a much more accurate spatial representation *of the trajectory as a whole*, as a direct consequence of the increased sampling rate. This is clearly shown in plots (a) and (b) of Figure 1 for a single trajectory, as well as in plots (a) and (b) of Figure 3 for multiple trips of a same user. These figures highlight the capability of NSD to capture individual mobility patterns in a much more exhaustive way compared to CDR. The unprecedented spatiotemporal resolution of NSD is at the basis of TRANSIT.

3.3. Impact of the radio technology

An important aspect of the data employed for our study is that it covers three generations of cellular network technologies. This lets us investigate the relevance of events generated by 2G, 3G, and 4G events on the accuracy

Table 1: Statistics on the large-scale network signaling data

Dataset	City	Area (km^2)	Nb antennas	2G events ($\cdot 10^6$)		3G events ($\cdot 10^6$)		4G events ($\cdot 10^6$)	
				Nb IMSI	Nb events	Nb IMSI	Nb events	Nb IMSI	Nb events
$\mathcal{D}_L$	Lyon	1,506	646	1.7	83	2.8	1,470	2.9	20,994
$\mathcal{D}_P$	Paris	5,784	3,972	5.9	850	6.5	10,166	6.1	116,461



(a) CDR (b) NSD (c) TRANSIT

Figure 3: Sample weekly trajectories of one voluntary user inferred from CDR: (a), NSD: (b) and TRANSIT: (c).

of the positioning data. Table 1 breaks down the number of unique devices observed under each technology, as well as the number of events recorded, separately reported for the two large-scale datasets related to Paris and Lyon, $\mathcal{D}_P$ and $\mathcal{D}_L$, respectively. The figures evidence how the number of users that can be monitored by the three radio access technologies is comparable, and partially overlapping. However, the sets of geo-referenced NSD collected for the monitored devices is completely different: the number of events grows by more than one order of magnitude when moving from one cellular generation to the next.

While this is a clear result of the increased consumption of mobile services and associated growth of mobile data traffic that newer network technologies support, it further distinguishes our study from the many previous works that date back to the 2005-2015 period, and that could only rely on limited 2G and 3G data.

4. The TRANSIT Framework

The rationale behind TRANSIT is to leverage the inherent *regularity* of individual mobility, in combination with the high temporal resolution of NSD, to reconstruct the fine-grained mobility of individuals in urban areas. Previous works have already identified the high regularity that characterizes human movements [48, 49], and possibly used it to help coarse mobility inference at, *e.g.*, hourly resolution [12]. Indeed, regularity is already visible with the CDR employed in such earlier studies, as exemplified by Figure 3a. Yet, NSD provide a much more accurate perception of individual movement regularity, as illustrated in Figure 3b, which TRANSIT takes advantage of.

Our framework receives as input the set of NSD events of a mobile device i denoted by $\mathcal{T}^i = \{e_1^i, \ldots, e_n^i, \ldots, e_{N_i}^i\}$, where e_n^i is the n^{th} NSD event recorded for device i. Each NSD event is the result of a communication activity between

Symbol	Description
i	Generic mobile device, also referred as user.
$\mathcal{T}^i$	Temporally sorted set of NSD events of mobile device i.
N_i	Number of NSD events in $\mathcal{T}^i$.
e_n^i	The n^{th} NSD event recorded for device i.
c_n^i	Generic antenna of the mobile network where mobile device i is attached when e_n^i is recorded.
t_n^i	Timestamp of the instant when e_n^i is recorded.
l_n^i	Location of the antenna that handled e_n^i, expressed in terms of (latitude, longitude) coordinates.
T_w	Minimum cumulated time for an antenna to be labeled as static (*tunable parameter*: a default value of 20 minutes is used).
a_k^i	Generic static activity session of device i, *i.e.*, maximal set of consecutive events only associated to static antennas.
$\mathcal{A}^i$	Set of all static activity sessions across the whole observation period $[t_0^i,\, t_{N-1}^i]$ of device i.
N_o	Maximum number of unique antennas, associated to events recorded after the end of a_k^i and before the beginning a_{k+1}^i, required for merging a_k^i and a_{k+1}^i in one single static activity session (*tunable parameter*: a default value of 2 is used).
T_s	Minimum duration of a static session (*tunable parameter*: a default value of 20 minutes is used).
m_h^i	Generic mobile session (*i.e.*, trajectory) of device i defined as the maximal set of consecutive events not belonging to any static activity session, after their merging process. It includes the last static event of the preceding static session, if any, and the first static event of the following static session, if any.
$\mathcal{M}^i$	Set of all mobile sessions across the whole observation period $[t_0^i,\, t_{N-1}^i]$ of device i.
$\mathcal{M}_R^i$	Set of trajectories from $\mathcal{M}^i$ that are classified in a cluster by DBSCAN and identified as recurrent by TRANSIT.
$\widehat{\mathcal{M}_R^i}$	Set of recurrent trajectories from $\mathcal{M}^i$ that are spatially augmented by TRANSIT.
$\mathcal{M}_O^i$	Set of unique trajectories from $\mathcal{M}^i$ that are classified as outliers by DBSCAN and left spatially unmodified by TRANSIT.
$\widehat{\mathcal{M}^i}$	Final set of trajectories retrieved by TRANSIT corresponding to $\widehat{\mathcal{M}_R^i} \cup \mathcal{M}_O^i$.
$d_H(.,.)$	Hausdorff distance.
$d_(.,.)$	Geodesic distance.
D_s	Maximum distance allowed between pairs of static positions in the DBSCAN clustering ensuring consistency in the location of events belonging to a cluster of static activity sessions (*tunable parameter*: a default value of 0.15 km is used).
D_m	Maximum distance allowed between pairs of mobile trajectories in the DBSCAN clustering process aimed at grouping trajectories with similar spatial geometries (*tunable parameter*: a default value of 2.5 km is used).

Figure 4: Main Notation

a mobile device and a base station antenna of the telecommunication network, across all 2G, 3G and 4G technologies; it is defined as a tuple $e_n^i = (c_n^i, t_n^i)$, where c_n^i is the antenna at location l_n^i that handled the network event, and t_n^i is the timestamp of the instant at which the event was recorded. The NSD events in a mobile phone trace $\mathcal{T}^i$ are ordered by their timestamps t_n^i, and N_i denotes the number of events for device i. Then, TRANSIT processes $\mathcal{T}^i$ to produce two outputs in succession, as follows.

- *Trajectory identification.* The framework labels each NSD event $e_n^i \in \mathcal{T}^i$ as either static, if the user i is deemed to be engaged in an activity at a same location at the event time t_n^i, or mobile, if i is performing a movement at t_n^i. The labeling factually allows telling apart the continuous time intervals during which an individual is moving or not, and building a set $\mathcal{A}^i$ of *static activity sessions* and a set $\mathcal{M}^i$ of *mobile sessions*. As a result, the set $\mathcal{M}^i$ also identifies all the *trajectories*, *i.e.*, continued sequences of movement in time, of user i. This step is described in details in Section 4.1.

- *Trajectory augmentation.* The framework enhances the trajectories associated to mobile sessions in $\mathcal{M}^i$, by exploiting the fact that the same individual typically performs many trips between two given locations over time, generally following very similar paths. This creates redundancy in the mobility information that can be used to increase the spatiotemporal accuracy of the trajectories, as shown in Figure 3c. The resulting set of mobile sessions possibly augmented trajectories is denoted as $\widehat{\mathcal{M}}^i$. Details are in Section 4.2.

Ultimately, the output of TRANSIT are the set $\mathcal{A}^i$ of static activity sessions of user i, and the set $\widehat{\mathcal{M}}^i$ of mobile sessions with augmented trajectories. Figure 4 summarizes our notation, and Figure 5 presents a flowchart of the stages of TRANSIT.

4.1. Trajectory identification

As anticipated, the trajectory segmentation step is applied to the individual set of NSD events $\mathcal{T}^i$ recorded for device i, and returns a subset of $\mathcal{T}^i$ where each event is labeled as static or mobile and detected oscillations are removed.

Figure 6 illustrates the process of trajectory identification using TRANSIT. The interpolation of NSD events is portrayed as the black solid line. Figure 6a refers to static antennas with daily accumulated association time above T_w. Figure 6b identifies static activity sessions as obtained from consecutive sequences of static antennas only, and detected oscillations. Figure 6c exhibits the final static activity sessions upon removal of oscillations, as well as the consequent detected mobile sessions.

We start by assuming that the time spent by user i at the antenna c_n^i associated to event e_n^i is $t_{n+1}^i - t_n^i$, *i.e.*, the temporal span to the subsequent event e_{n+1}^i. Given the high temporal resolution of NSD, this simple approach already provides a very good estimation of the time the user is associated to a given antenna, at a low computational cost. Then, a preliminary labeling is performed to trim down candidate static events. To this end, we calculate the cumulated time spent by user i at each antenna c_n^i, on a daily basis. As devices stay connected to a limited set of antennas while still, we expect such antennas to yield a non-negligible cumulated time during the target day. We thus tag as *static antennas* for user i those antennas with a daily cumulated time above a threshold T_w. In our experiments, we set T_w to 20 minutes, which falls within the range of commonly accepted values for the typical

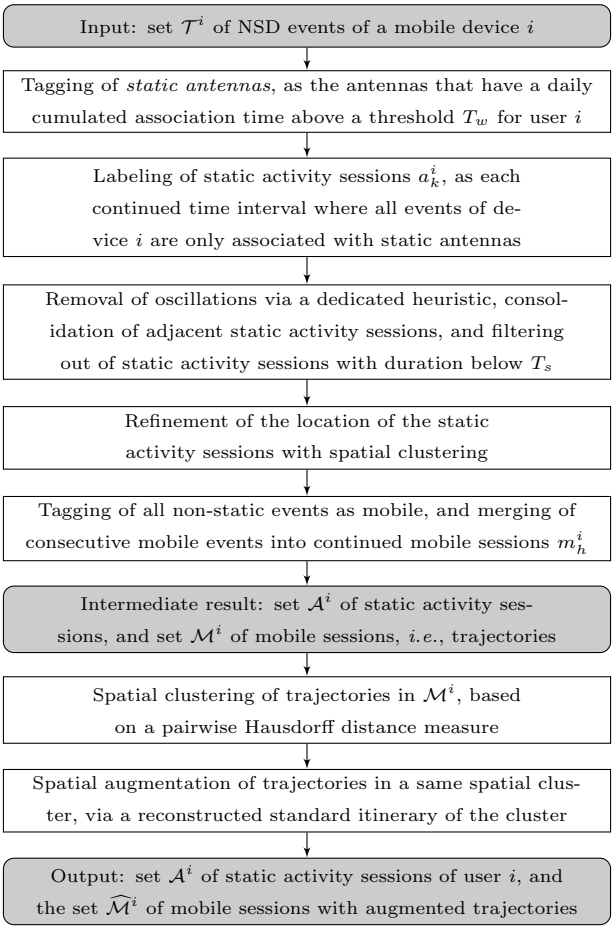
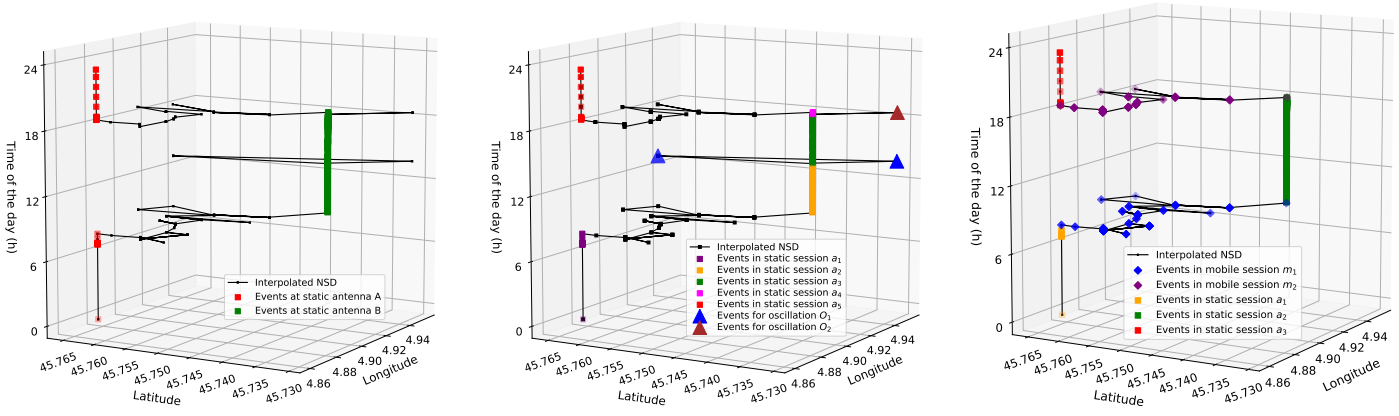


Figure 5: Flowchart of TRANSIT



(a) (b) (c)

Figure 6: Main steps of the trajectory identification via TRANSIT.

minimum duration of a significant activity carried out by an individual at a same location [50, 22], and is employed also with high-frequency longitudinal (*e.g.*, GPS) data [29]. An example is provided in Figure 6a.

A continued time interval where all events of device i are only associated with static antennas is then denoted as a static activity session a_k^i. The set of all such sessions across the whole observation period is $\mathcal{A}^i = \{a_1^i, \ldots, a_{K_i}^i\}$. Typically, during one day, a user can have several static sessions, and each can be composed of one or multiple antennas.

After the stage above, only part of the antennas are labeled. Unlabeled antennas are either encountered during movements, or the result of oscillations that are known to characterize mobile device association to the radio access infrastructure [29]. Oscillations can in fact affect both static and mobile users. In the former case, they can cause the separation of continuous static activities into different static sessions in $\mathcal{A}^i$ interleaved by non-static antennas. In order to address the issue, and remove oscillations from $\mathcal{A}^i$, TRANSIT adopts the following heuristic. If (i) two consecutive static sessions a_k^i and a_{k+1}^i present at least one common (static) antenna, and (ii) the number of unique antennas associated to events observed after a_k^i and before a_{k+1}^i is below a threshold N_o, we merge all the events in a_k^i and a_{k+1}^i into a new, single static session. The new sessions replaces the former pair in $\mathcal{A}^i$. An example of oscillation detection and static sessions before the merging process is shown in Fig 6b.

The single events identified as oscillations in the previous stage are in fact removed from $\mathcal{T}^i$ entirely, so as to limit uninformative noise in the data. The revised static sessions in $\mathcal{A}^i$ are further filtered based on their total duration, and only those with time span higher than a threshold T_s are retained. The value of T_s corresponds to the assumed minimum duration of a static activity, so that we do not include, *e.g.*, waiting periods at red traffic lights for pedestrian or vehicular trips, or dwell times at stops for bus trips. For the same reasons explained above in relation to threshold T_w, used to identify static antennas, the value of 20 minutes has been adopted for T_s as well.

TRANSIT also enforces consistency in the locations of events associated to static activity sessions, as follows. First, we compute the centroid of the locations l_n^i of all events in each session a_k^i; then, the well-known DBSCAN clustering algorithm[2] is run on the centroids of all $a_k^i \in \mathcal{A}^i$. This lets us group together all static sessions related to a same activity, and compute a consolidated location for the activity as the barycenter of all centroids in a same cluster. The locations l_n^i of all events in each session a_k^i are then replaced with the barycenter of the corresponding cluster. Note that the position of the static activity sessions that are labeled as outliers by the DBSCAN algorithm are left unchanged. An example of the resulting $\mathcal{A}^i$ is in Fig 6c.

Finally, all events that have not been labeled as static are labeled as mobile. This directly identifies the mobile sessions m_h^i of user i, as the time-continuous sequences of mobile events; an important remark is that the two static events immediately preceding and following the mobile session are also integrated into m_h^i. As a result, the set of mobile sessions is $\mathcal{M}^i = \{m_1^i, \ldots, m_{H_i}^i\}$. Each m_h^i corresponds to one trajectory of user i identified by TRANSIT. An example is also in Fig 6c.

4.2. Trajectory Augmentation

The sequences of NSD events in $\mathcal{T}^i$ that correspond to the single trajectories m_h^i of user i are still affected by the limited spatial accuracy of mobile network data, which affects NSD as explained in Section 3.2. In its second

[2]The parametrization of DBSCAN for static session clustering leverages is discussed later in Section 5.4.

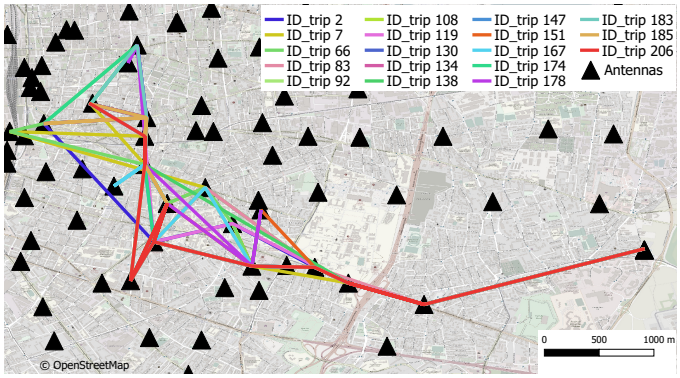


Figure 7: Set of trajectories of a same voluntary user clustered by DBSCAN, for the Origin-Destination path in Figure 1.

phase, TRANSIT thus aims at improving the geographical correctness of the movement information. As anticipated, the framework relies on the regularity of human mobility; more precisely, we use the the information from multiple similar trajectories identified for a same user to mutually improve their accuracy.

As a first step, a similarity measure is computed for all pairs of mobile session $m_h^i \in \mathcal{M}^i$. We employ the Hausdorff distance [51], which is defined as:

$$d_H(m_{h_1}^i, m_{h_2}^i) = \max\{D(m_{h_1}^i, m_{h_2}^i), D(m_{h_2}^i, m_{h_1}^i)\}, \quad \text{where} \quad D(m_{h_1}^i, m_{h_2}^i) = \sup_{l_{n_1}^i \in m_{h_1}^i} \inf_{l_{n_2}^i \in m_{h_2}^i} d(l_{n_1}^i, l_{n_2}^i), \qquad (1)$$

where $m_{h_1}^i$ and $m_{h_2}^i$ are the two mobile sessions to be compared and $d(\cdot,\cdot)$ is the geodesic distance between the two argument locations. This results in a matrix of pairwise distances between all mobile sessions of a same user i.

Then, DBSCAN is applied[3] to the distance matrix, in order to group trajectories that have similar spatial geometries, and correspond to diverse trips of the user between the same two static activity locations. Figure 7 shows an example of a set of mobile sessions, *i.e.*, trajectories, grouped together in the same cluster by DBSCAN, for the origin-destination activity locations in Figure 1. Based on the result of DBSCAN, we can tell apart the mobile sessions in $\mathcal{M}^i$ into two subsets: (i) trajectories that fall into a cluster, *i.e.*, which refer to a path that is recurrent in the mobility of user i, and which we denote as the set $\mathcal{M}_R^i$; and, (ii) outlier trajectories that represent unique movements of i, which are grouped in set $\mathcal{M}_O^i = \mathcal{M}^i \setminus \mathcal{M}_R^i$.

For trajectories in $\mathcal{M}_R^i$, TRANSIT operates a spatial augmentation, as follows. First, the average duration is computed for all trajectories assigned to a same spatial cluster by DBSCAN above; this corresponds to the expected time that user i takes to travel between the same origin-destination activity locations. The time information is used to filter out trajectories whose duration deviates from the median by 50% or more: these mobile sessions are considered not representative of the routine mobility patterns along the target path. The retained trajectories in a same cluster are then temporally scaled (*i.e.*, stretched or compressed) in time so as to match the average travel duration for the cluster. Finally, the scaled trajectories are temporarily binned according to a fixed time period of one minute, and the spatial coordinates of all different events that fall in a same time bin are averaged.

The previous steps lead to a set of positions, one per minute, which represent the reconstructed itinerary. If there is no event within a particular time slot, the resulting enhanced trajectory will have missing positions. All trajectories

[3]The parametrization of DBSCAN for mobile session clustering is discussed later in Section 5.4.

in the cluster are then matched to the reconstructed one, and become thus identical in the space dimension. However, they are re-conducted to their original duration (*i.e.*, via compression or stretching) so as to keep them faithful to their recorded travel time in the NSD.

As a result, each original mobile sessions in $\mathcal{M}_R^i$ is replaced by a set of reconstructed positions without any temporal deformation, and is enriched with information derived from multiple similar trajectories traveled by the same user. This set of enhanced mobile sessions is referred as $\widehat{\mathcal{M}_R^i}$. We recall that Figure 1c shows the final spatial trajectory inferred from the cluster in Figure 7. Trajectories in $\mathcal{M}_O^i$ stay instead unchanged, corresponding to those obtained from the simple interpolation of NSD data. The final set of mobile sessions is $\widehat{\mathcal{M}^i} = \widehat{\mathcal{M}_R^i} \cup \mathcal{M}_O^i$.

5. Validation and sensitivity analysis

We validate TRANSIT by using ground-truth information on the trajectories of a small set of volunteers. More precisely, we collected high-resolution trajectories of each volunteer using a GPS logger running on their smartphones; also, we recovered the NSD data generated by the volunteers' devices (which used Orange as their network provider) during the same observation period. Although reduced in size, this is one of the first dataset allowing a direct comparison of NSD and GPS data.

In the following, we present the validation datasets (Section 5.1), which we then employ to demonstrate the quality of the trajectories identified (Section 5.2) and augmented (Section 5.3) via TRANSIT, in absolute terms as well as with respect to solutions in the literature. Also, we use the validation datasets for a thorough analysis of the sensitivity of TRANSIT to the model parameters (Section 5.4).

5.1. Small-Scale Data Collection

The trajectory data used in our validation was collected by four Orange subscribers who voluntarily agreed to be monitored by a GPS tracking app installed on their smartphones, and who provided informed consent for their NSD to be extracted from the network operator database before pseudonymization and employed for the purpose of this research. Once gathered, all data were in any case pseudonymized, and accessed by authorized personnel of the research team only. The combined GPS and NSD data of the four users, denoted as A, B, C and D in the following, were collected during a continued period of three months, March 15 and June 15 2019, in the city of Lyon, France. We stress that size of the volunteer set, although limited, is aligned with that of state-of-the-art studies [45], with respect to which we collect a much larger number of human trajectory samples.GPS and NSD traces generated by the users A, B, C and D are represented in Fig 8c and in Fig 8d for one day, as well as for the whole observation period in Fig 8e and Fig 8f.

The dataset of GPS locations, named $\mathcal{E}_{GPS}$ in the following, contains GPS data collected via a custom Android application installed on the volunteers' personal mobile phone, so as to track their movements with high resolution and in a continued manner during the observation period. For battery saving purposes, GPS data have been collected with a sampling rate of 5 seconds. Due to the higher spatial accuracy (in the order of meters) and temporal granularity (order of seconds), we employ $\mathcal{E}_{GPS}$ as ground truth information about the mobility of the users.

The NSD dataset, named $\mathcal{E}_{NSD}$ in the following, contains all network signaling events associated to the mobile devices of the four voluntaries, across 2G, 3G and 4G technologies. We highlight that (i) all volunteers were Orange

Table 2: Performance evaluation results for the trajectory identification task. The second and third column report the temporal span of the combined GPS and NSD data, and the number of ground-truth trajectories, respectively. Best values are highlighted in bold.

User	Hours	Trajectories	TRANSIT				CWMA			
			Precision	Recall	F1	Trajectories	Precision	Recall	F1	Trajectories
A	64	17	0.44	1	**0.61**	8	0.65	0.35	0.45	**25**
B	202	78	0.83	0.96	**0.89**	**72**	0.89	0.76	0.82	164
C	426	138	0.77	0.97	**0.86**	**125**	0.79	0.87	0.83	217
D	208	77	0.83	0.94	0.88	**60**	0.88	0.91	**0.90**	98
All	900	310	0.80	0.96	**0.87**	**265**	0.85	0.82	0.83	504

subscribers at the time of the data collection campaign, and (*ii*) their were explicitly invited to maintain their regular mobile communication and service consumption habits during the measurement period. This limits biases, and we indeed observe that A, B, C and D have fairly heterogeneous profiles in the way they use mobile network services: Figure 8 shows that the median and average inter-event times in their NSD fall between the 60^{th} and 93^{th} percentiles of the distributions for all users in the $\mathcal{D}_P$ and $\mathcal{D}_L$ datasets that capture all subscribers in Paris and Lyon.

Overall, the validation datasets $\mathcal{E}_{GPS}$ and $\mathcal{E}_{NSD}$ provide corresponding GPS and NSD data for over 900 hours, and encompass over 300 ground-truth trajectories of the four volunteer users. Such ground-truth trajectories were identified by first applying a recent segmentation approach for spatiotemporal GPS data [29] to $\mathcal{E}_{GPS}$, and then having the volunteers verify the resulting movement patterns via visual inspection.

5.2. Validation of TRANSIT *Trajectory Identification*

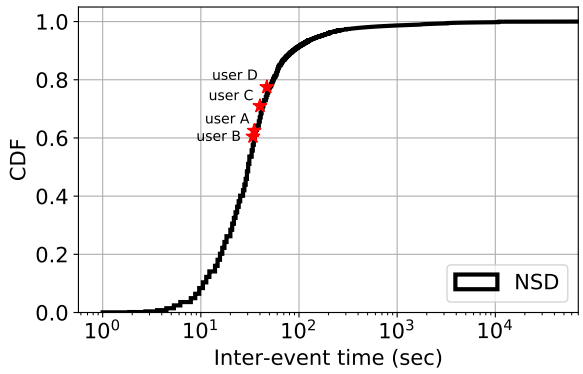
We first assess the performance of TRANSIT in identifying trajectories, by separating the static activity sessions and mobile sessions of a user. To this end, we compare the sessions identified by our approach applied on $\mathcal{E}_{NSD}$ against the ground truth extracted from $\mathcal{E}_{GPS}$. We also include in our analysis one recent benchmark from the literature, *i.e.*, the CWMA approach [30, 14] presented in Section 2. We use classical *precision*, *recall* and *F1* metrics to evaluate the performance of the trajectory segmentation approaches. Formally

$$\text{Precision} = \frac{TP}{(TP+FP)}\,, \quad \text{Recall} = \frac{TP}{(TP+FN)} \quad \text{and} \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{2}$$

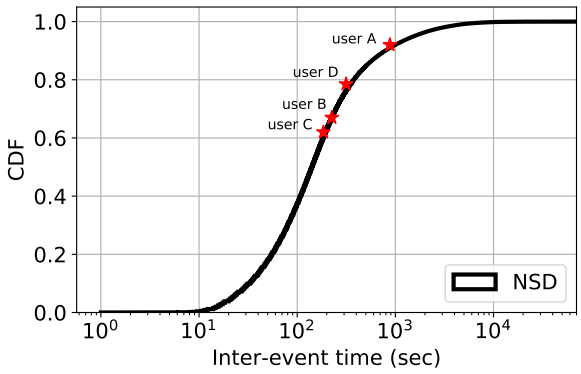
where: *(i)* the number of true positives TP is the number of NSD events labeled as static when the user is also considered as static in GPS data; *(ii)* the number of false positives FP represents the number of NSD events labeled as static while the user is in fact mobile according to the ground truth; *(iii)* the number of false negatives FN maps to the number of NSD events labeled as mobile while the user is static in the GPS data.

Overall and per-user results are summarized in Table 2. Both TRANSIT and CWMA attain rather high values of precision and recall, typically in the 75–100% range. For users with enough trajectories, *i.e.*, B, C and D, this leads to F1 scores between 0.8 and 0.9. However, the session classification approach of TRANSIT performs consistently better, yielding a 5% relative improvement in the total F1 score with resect to CWMA.

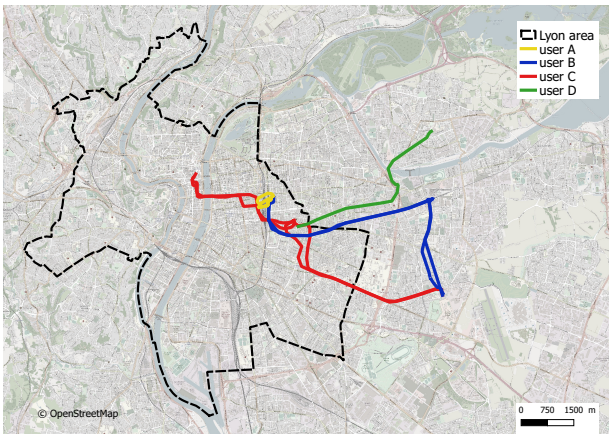
A closer inspection reveals how CWMA tends to yield higher precision than recall, *i.e.*, to incorrectly label static events as mobile. We ascribe the problem to the oscillation phenomenon discussed in Section 4.1: CWMA lacks a tool



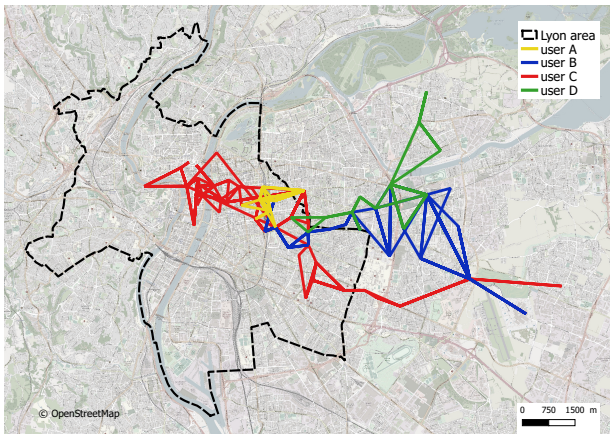
(a) Median



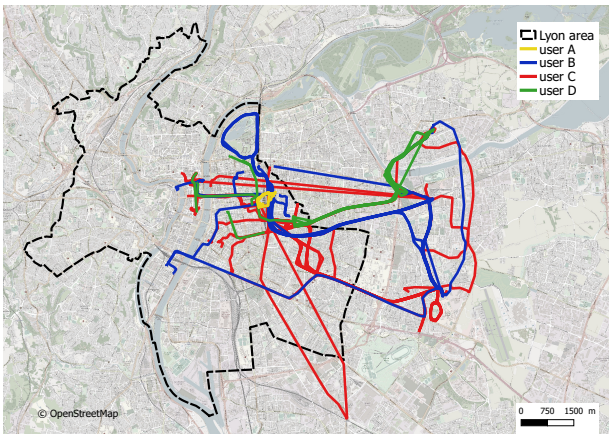
(b) Average



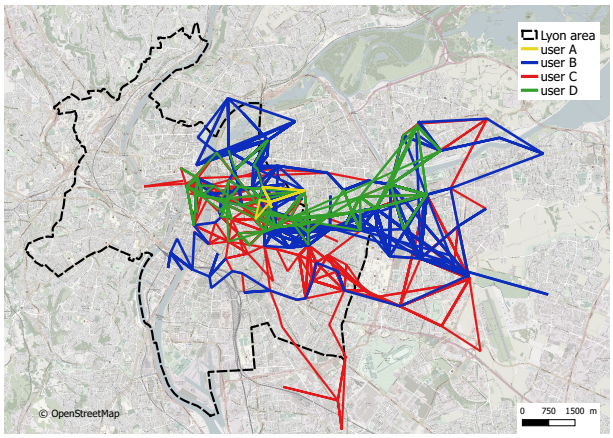
(c) GPS - one day



(d) NSD - one day



(e) GPS - all days



(f) NSD - all days

Figure 8: CDF of inter-event times recorded in NSD for the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$ (solid curve), and corresponding values for the voluntary users in the validation dataset $\mathcal{E}_{NSD}$. The plots refer to (a) median, and (b) average times per user. Daily traces of all users for GPS (c) and NSD (d) and all traces of all users for GPS (e) and NSD (f).

to remove oscillations occurring during static activity phases, hence tags such events as movements and overestimates the incidence of mobile events. As a by-product, CWMA detects a large number of non-existent trajectories (504 against 310 in the ground truth), which are in fact network-driven changes of the antenna serving the static user.

TRANSIT is designed to cope with these situations: it labels as oscillations and removes around 3% of the events. As a result, the number of identified trajectories (265) is much closer to the real one, and the result is fairly consistent across individual users. TRANSIT thus achieves near-perfect recall, while it slightly penalizes precision, by wrongly labeled static events where the user is in fact mobile. In-depth investigations revealed that this can appear in two situations. First, when a user performs very short displacements between the locations of two consecutive static activities, there is a risk that the two nearby static sessions will be merged into a single one, due to the limited spatial accuracy of NSD. Second, in round trips where the origin and the destination of the trajectory are the same, if the user connects to two or less different antennas, the mobility will be ignored altogether. These issues are caused by the finite spatial and temporal resolution of NSD, which our framework can mitigate only to a point.

5.3. Validation of TRANSIT *Trajectory Augmentation*

We now explore the capability of TRANSIT to improve the spatial representation of the individual trajectories identified above. We thus compare the augmented trajectories returned by our framework in $\widehat{\mathcal{M}}^i$ against the ground truth inferred from the GPS data in $\mathcal{E}_{GPS}$. We also consider a comprehensive set of benchmarks to contextualize the performance of our framework, as follows: (*i*) *DECRE/CDR* is the trajectory reconstruction method implemented by DECRE [28] as presented in Section 2 – in this case, we apply DECRE on CDR data extrapolated from NSD as explained in Section 3.2, as the method was originally conceived for this type of data; (*ii*) *CWMA/CDR+* is the trajectory reconstruction approach adopted by CWMA [30, 14] – here, it is applied to CDR+ data, also extracted from NSD as explained in Section 3.2, since these are the kind of data the approach was tested with by its authors; (*iii*) *Raw NSD* are the trajectories interpolated from the NSD directly, which is an important baseline for comparison; (*iv*) *DECRE* is the trajectory reconstruction method implemented by DECRE, run on NSD; (*v*) *CWMA* is the the trajectory reconstruction approach adopted by CWMA, run on NSD. Note that we are interested in comparing the different techniques in the specific task of trajectory augmentation: therefore, for the sake of fairness, we run TRANSIT and all benchmarks on the same set of trajectories $\mathcal{M}^i$, *i.e.*, those identified by our approach, as it provided the most accurate result in Section 5.2 above.

In all cases, two distance measures are used to evaluate the trajectory enhancement. On the one hand, D_{GPS} denotes the distance from the GPS ground-truth trajectory to that inferred from mobile network data: it is calculated by averaging the geodesic distance between each GPS point and the closest network data position in space. On the other hand, D_{NSD} is the distance from the mobile network trajectory to the GPS-based one: it is computed as the average geodesic distance between each point in the inferred trajectory from network data and its closest GPS point in space. Formally:

$$D_{GPS} = \frac{1}{|m_{GPS}|} \sum_{e_{n'} \in m_{GPS}} \min_{e_n \in m_{NSD}} d(l_{n'}, l_n) \quad \text{and,} \quad D_{NSD} = \frac{1}{|m_{NSD}|} \sum_{e_n \in m_{NSD}} \min_{e_{n'} \in m_{GPS}} d(l_n, l_{n'}) \tag{3}$$

where m_{GPS} and m_{NSD} are, respectively, two trajectories inferred from GPS and mobile network data. The operator $|\cdot|$ denotes the cardinality of the argument set, *i.e.*, the number of samples in the case of a trajectory, and $d(\cdot,\cdot)$

Table 3: Performance evaluation results for the trajectory augmentation task. Numbers represent the mean plus/minus the standard deviation, expressed in kilometers. Best values are highlighted in bold.

User	Measure	TRANSIT	CWMA	DECRE	Raw NSD	CWMA/CDR+	DECRE/CDR
A	D_{NSD}	**0.15 ± 0.03**	0.15 ± 0.12	0.20 ± 0.14	0.35 ± 0.15	0.12 ± 0.09	0.34 ± 0.14
	D_{GPS}	**0.15 ± 0.05**	0.26 ± 0.07	0.23 ± 0.06	0.23 ± 0.09	0.25 ± 0.07	0.23 ± 0.06
B	D_{NSD}	**0.14 ± 0.03**	0.18 ± 0.06	0.18 ± 0.05	0.22 ± 0.05	0.26 ± 0.10	0.30 ± 0.19
	D_{GPS}	**0.30 ± 0.08**	0.47 ± 0.27	0.50 ± 0.33	0.41 ± 0.08	0.75 ± 0.29	1.20 ± 0.32
C	D_{NSD}	0.19 ± 0.20	**0.16 ± 0.18**	0.30 ± 0.26	0.30 ± 0.27	0.24 ± 0.22	0.41 ± 0.44
	D_{GPS}	**0.20 ± 0.07**	0.30 ± 0.15	0.34 ± 0.14	0.34 ± 0.19	0.51 ± 0.28	0.92 ± 0.33
D	D_{NSD}	**0.13 ± 0.02**	0.14 ± 0.05	0.19 ± 0.08	0.20 ± 0.08	0.26 ± 0.10	0.33 ± 0.14
	D_{GPS}	**0.18 ± 0.03**	0.45 ± 0.27	0.49 ± 0.25	0.49 ± 0.15	0.95 ± 0.40	1.09 ± 0.28
All	D_{NSD}	**0.16 ± 0.12**	0.16 ± 0.13	0.23 ± 0.18	0.26 ± 0.19	0.25 ± 0.17	0.36 ± 0.32
	D_{GPS}	**0.22 ± 0.08**	0.38 ± 0.15	0.41 ± 0.25	0.38 ± 0.18	0.68 ± 0.36	1.01 ± 0.37

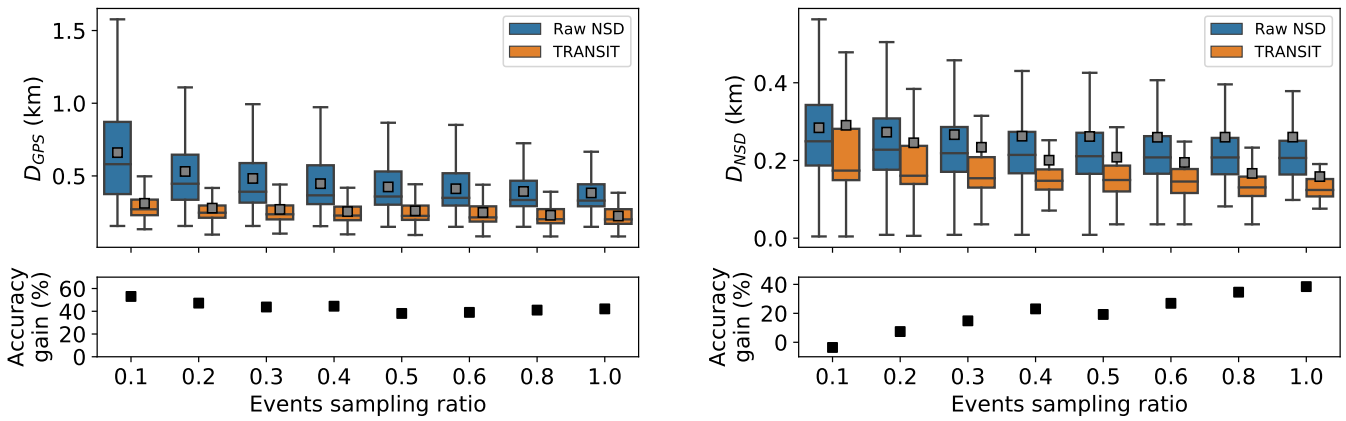
the geodesic distance. We use both metrics as they are complementary: while D_{GPS} is representative of the error observed for continuously tracked user, D_{NSD} measures the error specific to events recorded by the mobile phone network.

The results are reported in Table 3, for each user and in total. Trends are clear and consistent across users: there is a neat increase of accuracy in the inferred trajectories when moving from the right to the left in the table. Clearly, using CDR and CDR+ data penalizes DECRE and CWMA in the two rightmost columns, where the average error in the trajectory locations is 680–1,000 meters for D_{GPS}, and 250–360 meters for D_{NSD}. A simple interpolation of the Raw NSD already improves the result substantially, with average errors at 380 and 260 meters, for D_{GPS} and D_{NSD}, respectively. Interestingly, DECRE cannot improve that performance, mainly because its oscillation removal process has alternating effects, and can also eliminate events that are in fact useful to reconstruct the correct itinerary. CWMA improves the average D_{NSD}, bringing it down to 160 meters, however does not affect D_{GPS}. TRANSIT achieves the best performance in nearly all situations, and attains average errors that are as low as 220 meters for D_{GPS} and 160 meters for D_{NSD}.

Overall, the relative performance in Table 3 prove that TRANSIT does not simply rely on the added temporal resolution of NSD to advance the current state of the art; instead, it also introduces original processing that can take full advantage of NSD. From an absolute performance viewpoint, TRANSIT sets a new bar for the quality of individual trajectories inferred from mobile network data: with errors in the order of 150 meters, it demonstrates that a tailored processing of NSD can result in positioning information that is sufficiently accurate to support mobility monitoring applications at scale. We will provide multiple examples later, in Section 6.

5.3.1. Impact of Network Data Sampling Rate

We investigate further the settings that help TRANSIT achieve such a remarkable result in terms of accuracy of the inferred trajectories. As a first step, we consider the impact of the spatiotemporal sparsity of the NSD that is



(a) D_{GPS} (b) D_{NSD}

Figure 9: Analysis of the performance of TRANSIT versus the ratio of NSD events retained by subsampling, for the (a) D_{GPS} and (b) D_{NSD} distance metrics.

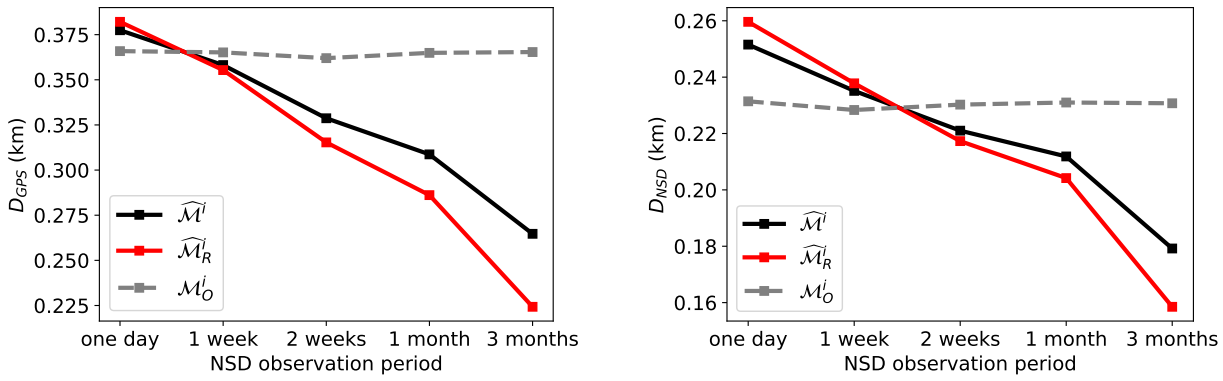
fed to TRANSIT. We do this by randomly subsampling the NSD of each user $i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ down to a fraction of original mobile events in every sessions in $\widehat{\mathcal{M}}^i$; we then run the trajectory augmentation method of TRANSIT on the sparser trajectories. Due to the stochastic nature of the subsampling, we averaged the metrics D_{GPS} and D_{NSD} over 10 trials for each distinct sampling ratio.

Figure 9 shows the results. When looking at D_{GPS}, the impact of the sampling ratio on TRANSIT performance is marginal, even when retaining as little as 10% of the NSD events. The relative gain in term of spatial accuracy of TRANSIT compared to trajectories obtained from a naive interpolation of the Raw NSD grows from 40% to 60%, as the latter are obviously negatively impacted by a reduced NSD sampling frequency. Concerning D_{NSD}, the trend is different. Indeed, the average value of D_{NSD} remains constant for raw NSD, regardless the sampling frequency, whereas it decreases for TRANSIT in the case of higher sampling ratio. Indeed, there is no reason that an increased number of NSD events would improve the intrinsic spatial uncertainty of Raw NSD: as this error is linked to the geographical sparsity of the antennas, the distance between NSD and the closest GPS position stays constant at around 300 meters. However, TRANSIT decouples trajectory samples from base station locations, and can better approximate the actual position of the user by averaging over a higher number of NSD samples collected at different antennas. This lets TRANSIT increase its gain up to 40% as the sampling ratio grows.

5.3.2. Impact of Data History.

As a second test, we study the effect of NSD temporal coverage on the performance of TRANSIT. To this end, we divide the 3-month NSD datasets $\mathcal{E}_{NSD}$ into non-overlapping shorter chunks; we consider chunks of one day in a first experiment, then of 1 week, 2 weeks, and 1 month in subsequent trials. We run TRANSIT's trajectory augmentation method on each chunk separately, and then compute the usual metrics D_{GPS} and D_{NSD} between the inferred trajectories and the ground truth.

The results are in Figure 10. The average accuracy of all trajectories identified by TRANSIT in $\widehat{\mathcal{M}}^i$ substantially improves for longer observation periods. The errors decreases by 40% for both D_{GPS} and D_{NSD} when NSD are



(a) D_{GPS} (b) D_{NSD}

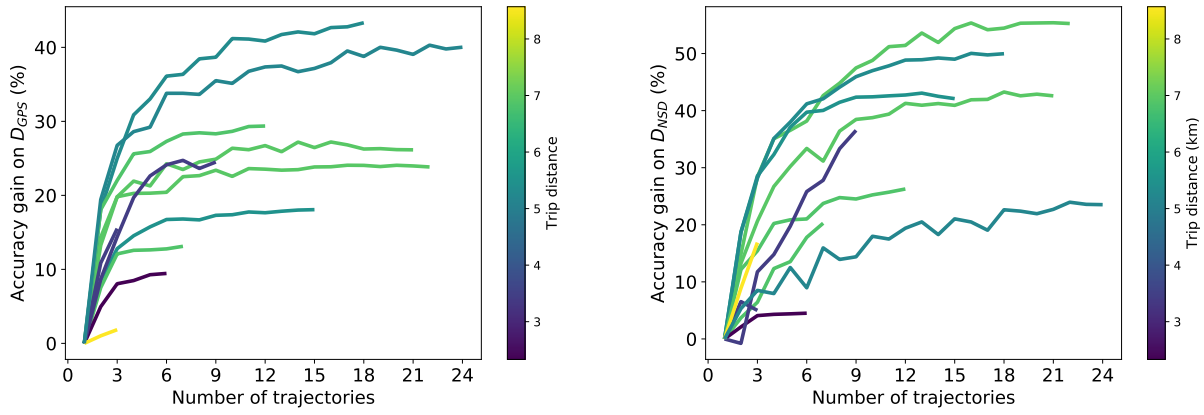
Figure 10: Analysis of the performance of TRANSIT versus the time span of the NSD data, for the (a) D_{GPS} and (b) D_{NSD} distance metrics. Different curves report the results for trajectories in $\widehat{\mathcal{M}}^i$, $\widehat{\mathcal{M}}^i_R$, and $\mathcal{M}^i_O$.

collected during three months rather than in a single day. Also, recall that $\widehat{\mathcal{M}}^i$ is in fact composed of trajectories that are actually augmented by TRANSIT, in the set $\widehat{\mathcal{M}}^i_R$, and trajectories that the framework could not improve due to the lack of similar movements in the user data, in the set $\mathcal{M}^i_O$. Thus, Figure 10 also breaks down the results for these two categories. As expected, different NSD time spans do not affect the accuracy in $\mathcal{M}^i_O$. Instead, in the case of the recurrent trajectories in $\widehat{\mathcal{M}}^i_R$, a longer history of mobility helps clustering and averaging a larger number of similar mobility patterns of the user, hence reducing the natural spatial bias of the original NSD.

5.3.3. Impact of the number of clustered trips.

The results in Figure 10 highlight that at least a few weeks of NSD data are needed in order for TRANSIT to be able to enrich recurrent trajectories. However, this is an artifact of the availability of additional comparable trajectories as we observe the user mobility in time. We thus decouple this phenomenon from the time dimension, and investigate how the number of clustered trajectories used to improve the spatial accuracy of NSD affects D_{GPS} and D_{NSD} directly. We conduct the following experiment: we select clusters of at least 3 trajectories, ending up with 11 clusters across all voluntary users. For these clusters, we test how the number N of trips within each cluster affects the spatial accuracy of the reconstructed itinerary. For instance, for one cluster, we select randomly N trips among all the trips within the cluster, we reconstruct the itinerary using these N trips and then compute the distance metrics. For each cluster, we are able to test N ranging from 1 to the number of trips within the cluster.

The results are shown in Figure 11. On the ordinate, we represent the spatial accuracy gain compared to the scenario using 1 trajectory for doing the reconstruction. For D_{GPS} in Figure 11a, we can observe that most of the curves have similar shapes, with a gain for a relatively low number of trajectories (between 2 and 6) and the emergence of a clear diminishing return effect afterwards. A similar phenomenon can be observed for D_{NSD} in Figure 11b, albeit with less neat transition. This behavior is consistent across trajectories covering different spatial distances (colors), and achieving diverse accuracy gains (final value in the ordinate). We conclude that, at least in the set of trajectories we could study, a fairly small number of less than 10 instances of the same route is typically sufficient to achieve the maximum error reduction that TRANSIT can grant.



(a) D_{GPS} (b) D_{NSD}

Figure 11: Analysis of the performance of TRANSIT versus the number of averaged trajectories per cluster in $\widehat{\mathcal{M}}_R^i$, for the (a) D_{GPS} and (b) D_{NSD} distance metrics. Different colors map to diverse geographical lengths.

5.4. Parameter Setup and Implementation Settings

5.4.1. Parameter Setup

TRANSIT requires the setting of five tunable parameters (*i.e.*, T_w, T_s, N_o, D_s and D_m), reported in the notation table of Figure 4. As mentioned in Section 4, the two thresholds T_w and T_s are representative of the minimum duration allowed for a human activity to be considered static (*i.e.*, cumulated time at a same location larger than the threshold). They have been both set to 20 minutes based on typical reference values from the related literature on activity detection via mobile phone and GPS data [50, 22, 29] that suggests to consider this order of values to avoid including short stops (*e.g.*, bus stops, red-light waits, etc.) in the classification.

The N_o parameter refers instead to oscillation removal and corresponds to the maximum number of unique antennas where a user can be observed between two consecutive static activity sessions in order that the two static sessions can be merged into a single one. To select the default value of No, we have performed trajectory identification with TRANSIT over a large range of values, *i.e.*, $[1, 10]$, using the segmentation information from dataset $\mathcal{E}_{GPS}$ as ground-truth. Accuracy is consistent and close to 100% for $N_o = 1$ and $N_o = 2$, while the number of retrieved trajectories rapidly decreases to 20 trips out of 310 when $N_o = 10$. These results lead to a final choice of 2 as the default value of N_o.

Finally, we report in Figure 12 the results of the sensitivity analysis performed to determine the two distance thresholds D_s and D_m, used as the maximum distance allowed in DBSCAN cluster for both location enhancement of static sessions and for the identification of similar mobile sessions. As performance criterion of the analysis, we have used the average of the two distance metrics D_{GPS} and D_{NSD} described in Equation 3. For each configuration of the parameters D_s and D_m in the ranges reported in Figure 12, we obtained a different set of $\widehat{\mathcal{M}}_R^i$ for all users in $\mathcal{E}_{NSD}$ and computed the corresponding value of our performance metric. The figure highlights that the selected performance criterion attains its minimum value (*i.e.*, better reconstruction of the real trace) when $D_s = 0.15Km$ and $D_m = 2.5Km$. These values have been thus selected as the default values for the two parameters.The sensitivity analysis on the thresholds T_s and T_w is reported in Appendix 8.1.

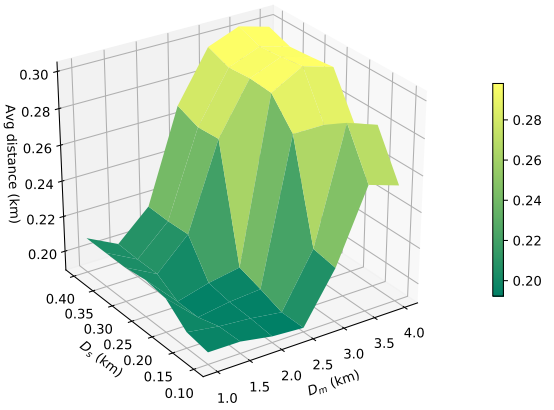


Figure 12: Parameter D_m and D_s sensitivity on trajectory enhancement performance.

5.4.2. Implementation settings

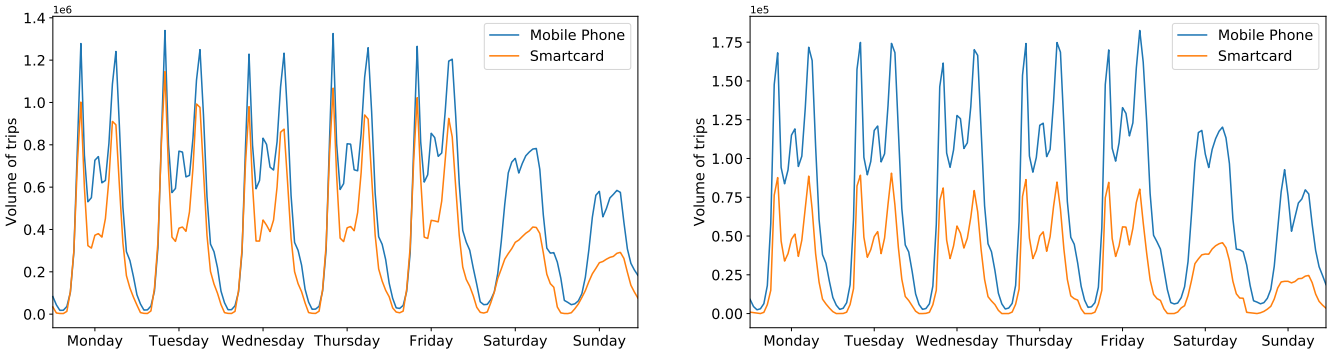
TRANSIT has been implemented in PySpark and run on a Spark cluster deployed at the mobile network provider's facilities. The Spark execution environment consists in 50 executors, each configured with 4 cores and 28 Gigabytes of memory. All the main algorithmic components of TRANSIT from Figure 4 have been implemented via PySpark User-Defined Functions (UDF) and applied in a distributed manner to the whole sets of subscribers' network signaling traces considered in our analyses. Specific optimizations have been required in order to process the three months of NSD from the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$. Among the different optimizations, a special attention was dedicated to the computation of the pair-wise Hausdorff distance matrix, which represents the most time-consuming step of our approach (taking approximately 70% of the total computation time). Specifically, we avoid computing the Hausdorff distance for all pairs of trajectories having different origin and/or destination. In such case, we set their distance to a value larger than the D_m parameter, thus making it impossible for DBSCAN to cluster them together. Similarly, the Hausdorff distance is immediately limited to D_m when a value larger than D_m is found during the iterative computation of the inner distances $D(.\ ,.)$ from Eq. 1. It is worth to note that this simple optimization allows us saving significant computation time, as well as keeping the result of the clustering unchanged.

6. Large-Scale Applications

While the validation results are related to a reduced number of users, the interest of TRANSIT reveals at city-wide scales, where it can enable a number of mobility-related applications. This section analyzes four case studies related to urban mobility that leverage the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$ described in Section 3.

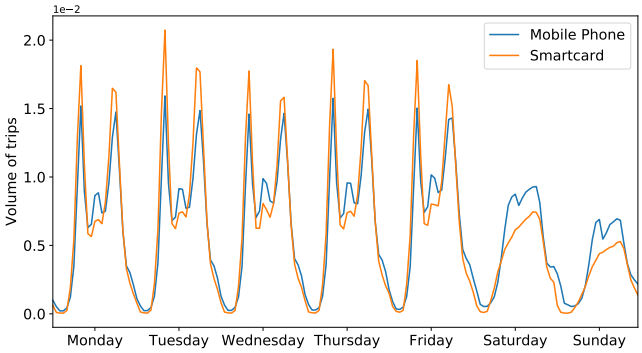
6.1. Urban Mobility and Public Transport

By counting the number of concurrent active trips inferred via TRANSIT over time, we are able to reconstruct accurate temporal profiles of the travel demand in urban regions. For such profiles to be dimensionally correct, a rescaling is needed to account for the penetration rate of the technology (close to 100% in developed countries like France) and the market share of Orange (at 37% over the French territory). The resulting average weekly demand

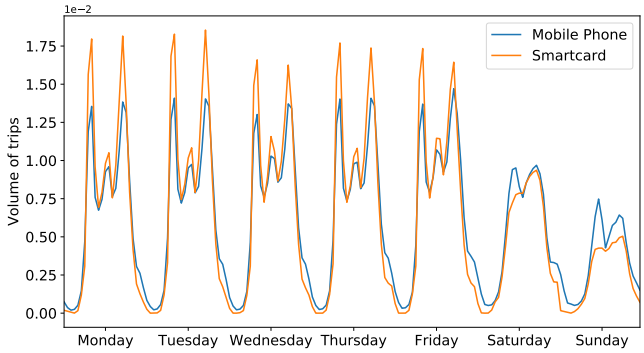


(a) Paris

(b) Lyon



(c) Paris, normalized



(d) Lyon, normalized

Figure 13: Average weekly profiles of the number of concurrent trips in (a) Paris and (b) Lyon, as inferred from TRANSIT and smart card data. Normalized versions with integral one of the same profiles are in (c) and (d).

profiles computed in Paris and Lyon are depicted in blue in Figure 13a and Figure 13b, respectively. Our estimates are that around 1,300,000 individual trips occur at the same time in Paris during commuting peaks, while the figure is at 180,000 for Lyon.

We compare the profiles obtained with TRANSIT with equivalent ones from smart card data, which capture mobility via public transportation systems. For Paris, data were provided by the transportation company Ile-de-France-Mobilité. Concerning Lyon, data were shared by the transportation company Keolis-Lyon. For both cities, public transport data were provided in the same period of the year of NSD, and all smart-card transactions were anonymized in the form of aggregate measures at the scale of the whole agglomeration.

Also in this case, a rescaling is required: while the TRANSIT trajectories refer to the resident population, the smart card data include both residents and non-residents. In order to make the numbers comparable, we apply a scaling factor of 0.81 to the smart card temporal profile; the factor has been calculated from the raw network signaling data, by computing the average instantaneous fraction of resident subscribers present in the target cities, over the total number of observed users. Thus, after the scaling (*i.e.*, multiplication by 0.81) applied to smartcard, the smartcard and NSD profiles are directly comparable. The weekly profiles from smart cards are superposed to the TRANSIT-inferred ones, as the orange curves in Figure 13a and Figure 13b.

The comparison of the profiles reveals interesting facets of mobility in Paris and Lyon. Clearly, the volume of trips identified by TRANSIT is higher than that reconstructed with smart card data: NSD allows monitoring virtually all transport modes, including those beyond public means, *e.g.*, private vehicles, biking, or walking. This lets us quantify which proportion of trips is performed with underground, buses or tramways, and which using personal means. We find that a significant fraction of trips is performed using public transports in both cities: we estimate the percentages of movements captured by smart card data to 66% and 39% of the total, in Paris and Lyon, respectively. The difference between these values is explained by the more developed multimodal transit network available in Paris, as required to support mass mobility in such a large metropolis.

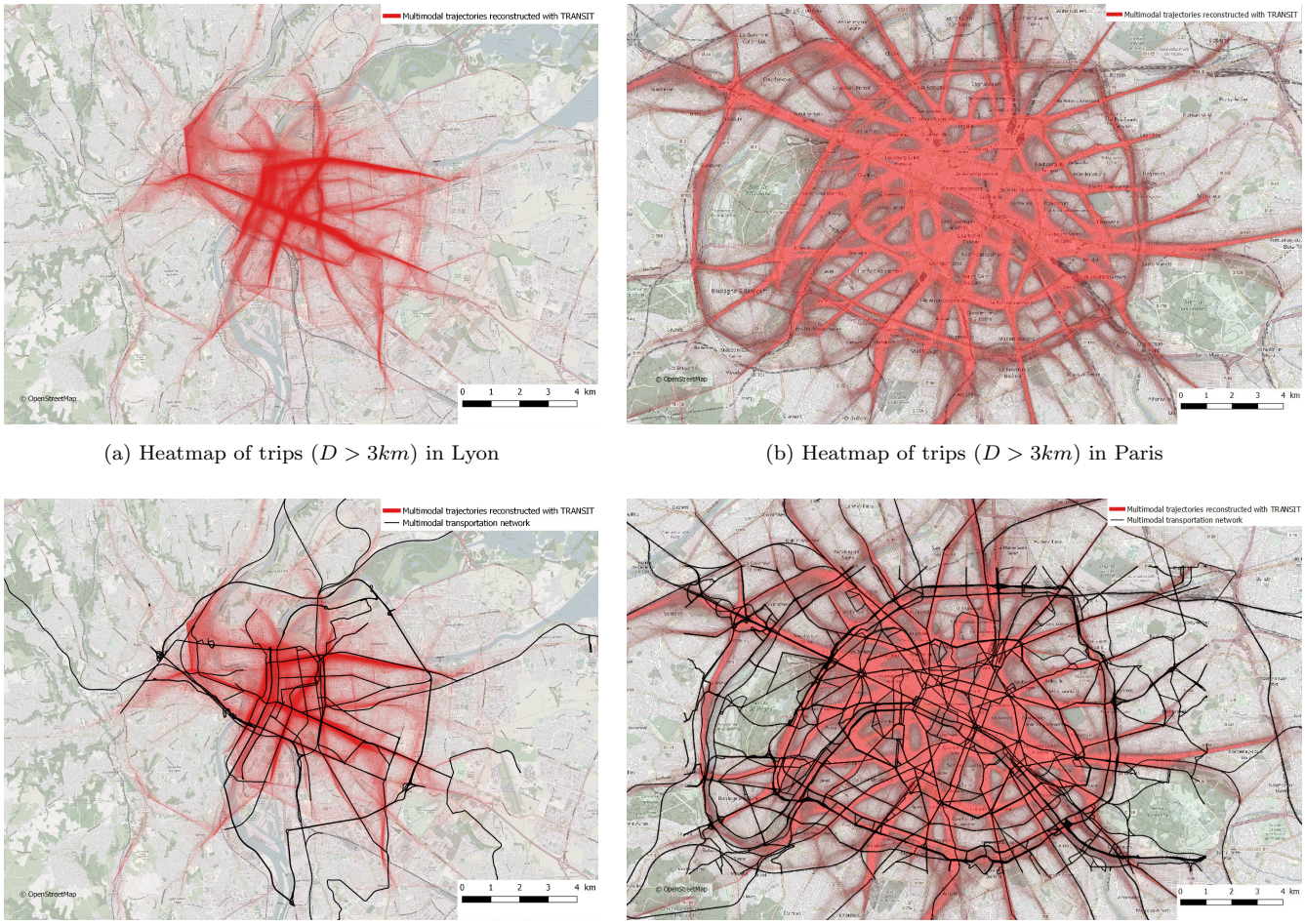
In addition, we can investigate the temporal incidence of public transports by looking at versions of the same profiles that are normalized so that the integral of all curves is one. Figure 13c and Figure 13d show the result. This perspective lets us appreciate how in Paris public transports are especially important during commuting hours, but relatively less used during the lunch break or weekends. A slightly different pattern emerges in Lyon, where public transports are also very much used around midday, but have a lower incidence on total mobility during evenings and weekend mornings. We highlight that obtaining this type of insights is hardly achievable by solely relying on surveying, which demonstrates the value of NSD and a method like TRANSIT that can exploit them.

As a final remark, we highlight that the results in Figure 13 can also be considered as a partial validation of the trajectories inferred by TRANSIT in large-scale settings. Indeed, the near-perfect match of the timing of commuting peaks or overnight low mobility among curves proves the capability of our trajectory segmentation approach to identify trips that are very consistent with data collected in the field over time.

6.2. Popular Paths of Commuting Trips

The previous section indicates that TRANSIT accurately extracts urban mobility patterns. Therefore, as a second application, we focus on inferring popular commuting trips within a city. The knowledge of such trips is

an extremely precious source of information for transport authorities and city planners as: *(i)* they represent the largest share of the the daily urban traffic demand of a city; *(ii)* they identify the typical commuting behaviors of travelers which regularly stress the transport network infrastructure, especially during peak hours; *(iii)* they are hard-to-quantify and characterize at city-scale because of the absence of dedicated sensors or probes that can precisely capture the multi-modal, diverse and time-varying nature of such trips.



(a) Heatmap of trips ($D > 3km$) in Lyon

(b) Heatmap of trips ($D > 3km$) in Paris

(c) Heatmap of trips ($D > 3km$) in Lyon with the multimodal transportation network of Lyon

(d) Heatmap of trips ($D > 3km$) in Paris with the multimodal transportation network of Paris

Figure 14: Heatmap of commuting trips in Lyon and Paris.

By applying our framework to the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$, we associate to each user i of the two analyzed cities a set of trips $\widehat{\mathcal{M}^i}$ for the whole period of 3 months. As explained above, $\widehat{\mathcal{M}^i}$ can be divided in two subsets: $\widehat{\mathcal{M}^i_r}$, a subset of recurrent trips enhanced by TRANSIT, and $\mathcal{M}^i_o$, a set of non-recurrent trips of user u. Considering that commuting trips are, by definition, recurrent, in the remainder of this section we focus our analysis only on subset $\widehat{\mathcal{M}^i_r}$. Furthermore, to extract commuting trips from $\widehat{\mathcal{M}^i_r}$, we filter only those trips associated to the two most popular locations of each user, under the constraint that at least 10 trips are present between these two locations. The underlying assumption is that the remaining set should mostly contain the two most popular trips performed by users in their daily routine, *i.e.*, home-to-work and work-to-home trips (commuting trips).

The spatial density (heatmap) of the reconstructed trips is represented in Figure 14a for Lyon and Figure 14b for Paris. As a first consideration, the recurrent trips appear to have overall a good match with the multi-modal urban transportation network, graphically overlapped to the heatmap in Figure 14c for Lyon and in Figure 14d for Paris. A more in-depth inspection of the figures highlights that, for both cities, the subway network, the tramway lines and most important urban roads clear show up among the commuting trips reconstructed via TRANSIT. In the case of Paris, NSD trips appear to have a near perfect match to the underlying multi-modal transport network. The less evident match for the case of Lyon, especially characterizing some peripheral roads (however present in the heatmap), can be explained by the lower number of available trips and the lower density of the cellular network of Lyon in these areas, compared to those from the capital city.

Of course, the fact that the majority of commuting trips maps to the public transportation network is not unexpected. However, TRANSIT opens the door to a detailed analysis of these trips, which we leave as future work: the obtained trips can be easily map matched to the different transportation lines and modes, showing their share of trips, in different days of the week and at different times of the day. Such information would be highly valuable for any public transportation company or municipality.

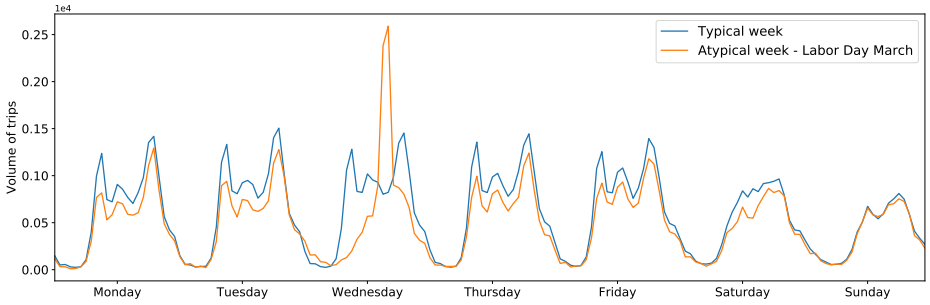
6.3. Human Mobility Analysis during Abnormal Events

As a third application, we use TRANSIT to detect abnormal mobility situations that can occur in the city. For this, we segment the city of Paris into a set of squares of dimension 800m $\times$ 800m, with a temporal bin size of one hour. This spatio-temporal granularity makes it possible to analyze human mobility at a fine-grained scale. For each zone, we compute the *attraction demand profile*, which corresponds to the number of trips having as destination the studied zone at any given hourly time slot. These profiles have been obtained by retaining such trips from the whole set of trajectories $\widehat{\mathcal{M}}^i$ computed via TRANSIT on $\mathcal{D}_P$ for each user i. This allows us to build a typical weekly attraction profile for each zone and, at the same time, to distinguish abnormal patterns during certain events. We use three such abnormal mobility situations as examples below.

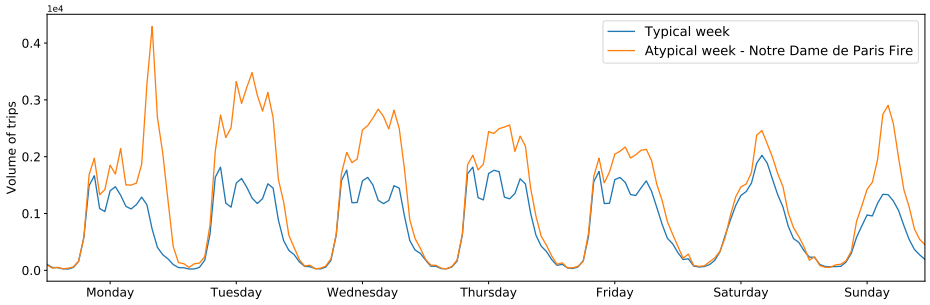
First of all, on Wednesday, the 1$^{\text{st}}$ of May 2019, a bank holiday, the Labor Day march took place near Place d'Italie in Paris. Figure 15a shows in blue the typical attraction profile of this zone and in red the attraction profile of the week that includes the demonstration. Whereas, for all days, the attraction profile was similar to the typical profile, we can see that, on Labor Day, the attraction of the studied zone presents a high peak after midday.

As a second event, we studied the fire of the Notre Dame de Paris cathedral, on Monday the 15$^{\text{th}}$ of April 2019. Figure 15b shows in blue the typical attraction profile of this zone, and in red the attraction profile of the week that includes the abnormal event. We can see a high peak in the attraction profile right after the beginning of the fire on Monday 15$^{\text{th}}$ (around 6:30pm). Contrary to the previous example, this event also affected mobility the following days, when an attraction demand higher than usual is observed in the corresponding area. This attraction demand progressively decreases after the event, but we notice an upsurge on Sunday, the Easter holiday, probably explainable by religious activities and nearby gatherings of tourists and worshipers visiting the area surrounding the cathedral after the fire on this special day.

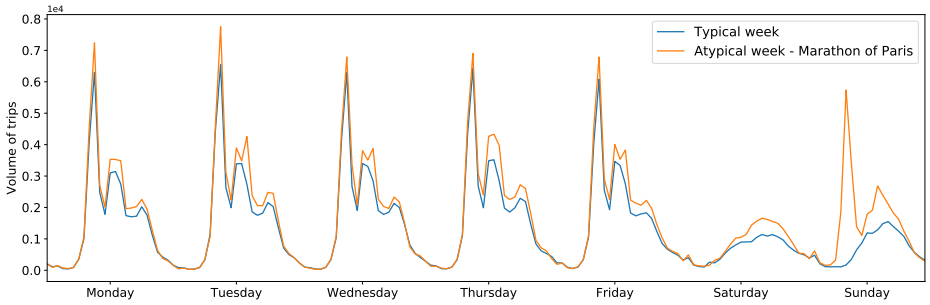
Finally, we study another special event, the Marathon of Paris, on Sunday the 14$^{\text{th}}$ of April 2019, with its start and end in the proximity of the Arc de Triomphe. A high peak on the attraction profile can be observed at the



(a) Attraction of the zone Place d'Italie (Paris)



(b) Attraction of the zone Notre Dame (Paris)



(c) Attraction of the zone Arc de Triomphe (Paris)

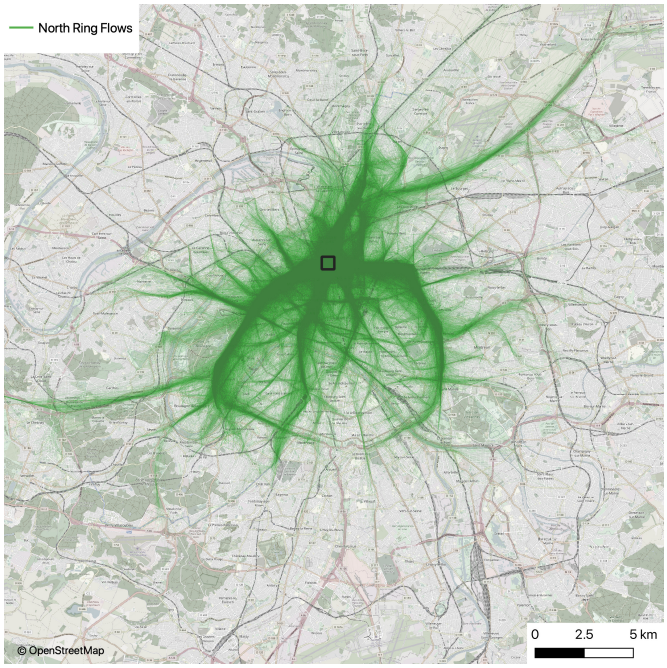
Figure 15: Typical/atypical weekly temporal demand profile during atypical events

departure time of the marathon, at 9am, as shown in Figure 15c. A second peak, is observed few hours later, more spread over time and lower in magnitude compared to the first one, corresponding to the marathon arrival.

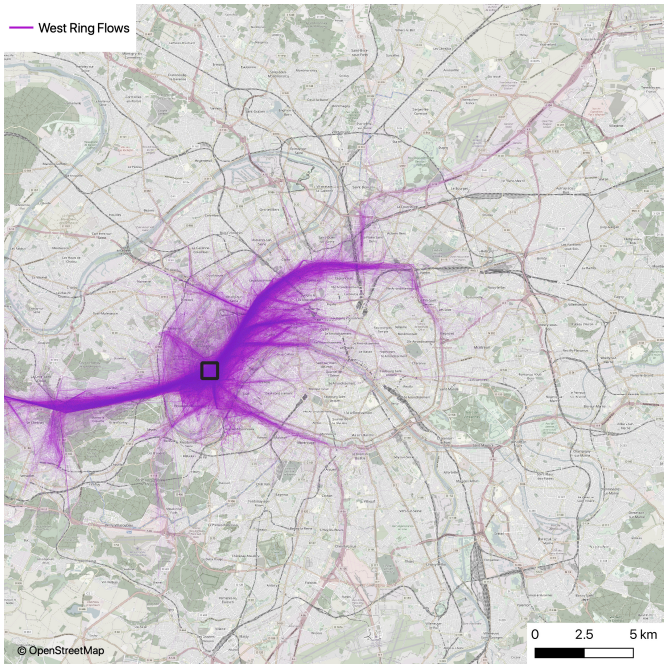
These three examples are representative of the vast potential of TRANSIT towards building mobility profiles of the typical demand attracted by a given zone, as well as detecting and characterizing mobility patterns during abnormal or special events.

6.4. Ring Road Trajectory Analysis

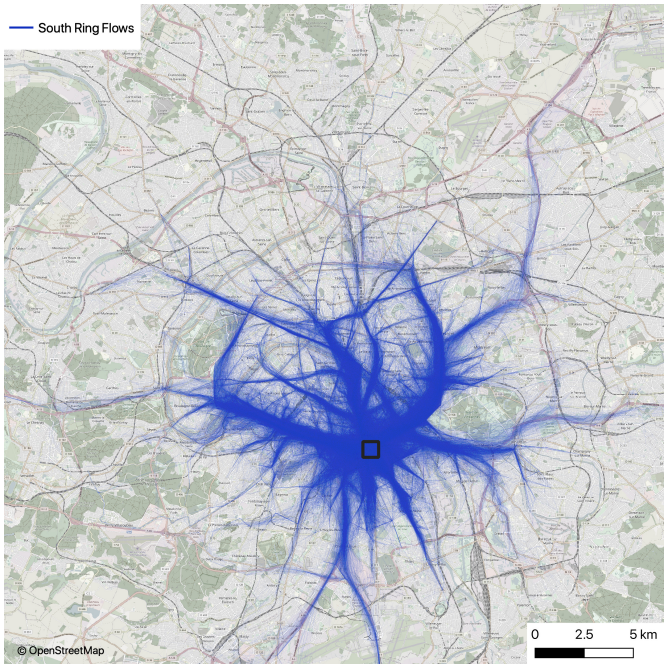
As a fourth application, we leverage TRANSIT to perform a fine-grained trajectory analysis focused on the Paris *périphérique* (ring road). The mobility flow on this urban highway is usually very high, often leading to heavy congestion especially during peak hours. Transport authorities are traditionally very interested in the possibility of tracing and quantifying the flows of people moving along city major road axes. Such studies are necessary for urban planning purposes, infrastructure renewal and road maintenance, and can be extremely cost-demanding. Specifically,



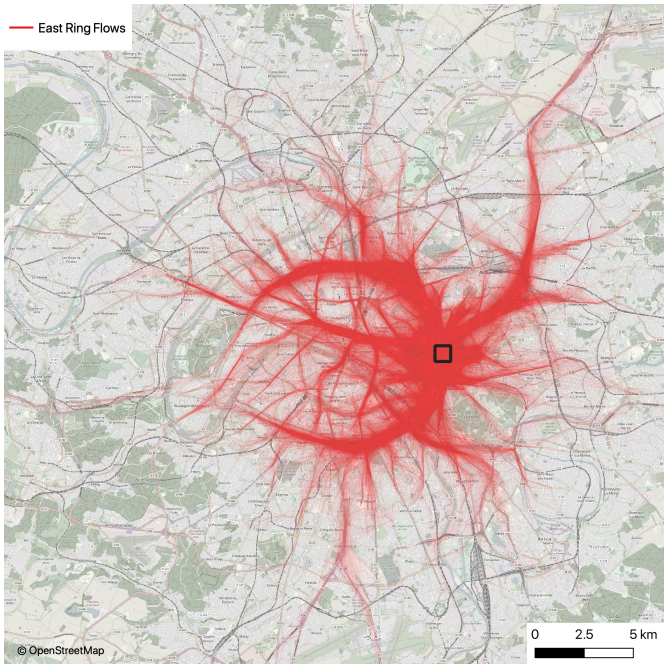
(a) North Ring Road



(b) West Ring Road



(c) South Ring Road



(d) East Ring Road

Figure 16: Heatmap of recurrent trips for the Paris ring-road (the black square shows the catchment area)

they are based on travel diaries or GPS trace collection, and generally end up capturing only a small sample of the flow actually traversing the major axis, with resulting limited accuracy. TRANSIT permits to leverage NSD to access a much larger and more representative sample of this specific population.

In our case study related the Paris *périphérique*, we considered four different zones of interest: the east, west, north and south entries. The idea is to select a spatial zone and study all the trajectories passing by the respective zone. The enhanced trajectories $\widehat{\mathcal{M}}^i$ produced by TRANSIT on $\mathcal{D}_P$ allow us to capture at scale the origin, the destination, and the paths taken by the users passing by the studied zone, the kind of information usually expected in the aforementioned studies. The result for the four zones of the *périphérique* (east, west, north and south) are thus reported in Figure 16b, Figure 16c, Figure 16d and Figure 16a.

The obtained maps underline the major role of the *périphérique* in Paris, allowing people to travel across the city and reach any area of interest. Some interesting patterns can be distinguished as well. For example, the trips coming from the west side of the city show a strikingly different pattern than the three other maps. This can be explained by the fact that the west side of Paris is the richest area of the city, with inhabitants who have a lifestyle involving shorter commuting trips. Moreover, the west side of Paris is also the area with the highest density of offices, including the *La Defense* and *Boulogne* neighborhoods. This could explain why this area attracts a large amount of trips, ever from faraway zones.

These results hint at the numerous perspectives brought by TRANSIT in the study of major road arteries. These include fine-grained temporal analysis, the detection of usage and attraction patterns, origin and destination profiling, etc. Generally speaking, having access to detailed human mobility trajectories at scale, as those produced by TRANSIT, enables the in-depth study of any part of the transportation network.

7. Conclusions

In this paper, we presented TRANSIT, a framework to classify mobile and static phases of human activity and reconstruct fine-grained individual human mobility trajectories from Network Signaling Data. TRANSIT advances the state-of-the-art on human-centric mobility trajectory inference by leveraging dedicated heuristics, consolidation of static activity location and trajectory enhancement via spatial clustering to: *i)* extract useful information from the higher sampling rate at which communication events are collected in NSD; *ii)* perform effective oscillation detection and removal; *iii)* capture the repetitive nature of individual trips over time. This combination of unique features permits to achieve improved classification of mobile and static sessions, as well as increased accuracy of the reconstructed trajectories.

The validation of TRANSIT, performed on a unique small-scale dataset combining NSD and high-resolution GPS trajectories for the same set of individuals, reports a near-perfect recall and comparable precision with respect to related work on mobile-versus-static human activity classification. Concerning the accuracy of the inferred trajectories, TRANSIT largely improves state-of-the-art solutions with unprecedented average spatial errors in the order of 150 meters in urban environments. Based on these results, we leveraged an efficient PySpark implementation of TRANSIT to process a large-scale NSD dataset of millions of mobile phone subscribers. A multitude of relevant city-wide patterns, hard or costly to discover via traditional data sources on human mobility, has been untangled by TRANSIT, such as transport mode shares, popular multi-modal commuting itineraries, human mobility pro-

files during normal/abnormal urban situations and a fine-grained accurate spatial representation of the multi-modal mobility flows traversing specific areas of major urban roads.

Despite its already satisfying results, TRANSIT can be further improved over several dimensions. The reconstructed trajectories could be easily map matched to the different lines and modes of the underlying transportation network to further reduce their spatial error. Other kinds of data on human mobility, such as smart card logs or GPS floating car data, could be jointly leveraged with NSD, *e.g.*, to better estimate the typical duration of similar trips and support a more informed filtering during the trajectory clustering process. Further technical optimizations could be helpful towards a stream-based online implementation of TRANSIT that could support a large variety of real-time applications, such as urban anomaly detection, data-driven dynamic control of transport infrastructures, as well as advanced location-aware caching schemes and scheduling policies for telecommunication networks.

Acknowledgements

The authors acknowledge the support of the French National Research Agency (ANR) grant number ANR-18-CE22-0008 (PROMENADE project) & grant number ANR-18-CE25-0011 (CANCAN project).

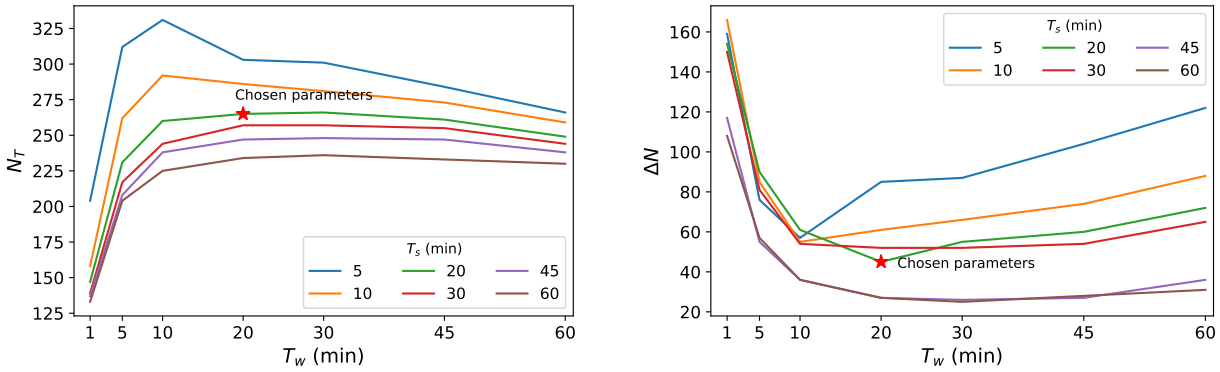
8. Appendix

8.1. Sensitivity analysis on parameters T_w and T_s

The criterion that has been considered for the selection of the T_s and T_w thresholds is twofold. Firstly, to determine T_s we consider the number of mobile sessions (*i.e.,* non-enhanced trips) detected by TRANSIT on our validation dataset $\mathcal{E}_{NSD}$, namely N_T. Secondly, to determine T_w we consider the number of missed trips by TRNSIT, namely ΔN, with respect to a benchmark segmentation method for GPS data [29], used as ground truth. We recall that T_s is the minimum duration of a static session, and it is a shared parameter of the benchmark segmentation method used with GPS data and the trajectory segmentation approach of TRANSIT used with NSD. T_w is the minimum cumulated time for an antenna to be labeled as static, and it is only related to the segmentation approach of TRANSIT.

On the one hand, Fig.17a shows the sensitivity of N_T on T_w and T_s. We can observe that for a given T_w, N_T increases when T_s decreases. In other words, when the minimum duration of a static session T_s decreases, TRANSIT captures more trips, *i.e.,* trips taking place between shorter stationary activities, which can correspond to *e.g.,* leisure stops, public transport connections and modal shifts, taking children at school, stops at traffic lights, etc. Therefore, T_s has to be chosen accordingly to considerations that are specific to the kind of mobility analyses one might want to study. With that regard, the nature of the trips that we aim to reconstruct and enhance in our paper with TRANSIT via NSD is mainly related to recurrent itineraries linked to different kinds of transport motifs, which do not normally include trips between very short stationary activities. For this reason, we set T_s to 20 minutes, which allows detecting a fairly high number of trips (as from Fig.17a) and is also a typical reference values from state-of-the-art literature on stationary activity detection via mobile phone and GPS data for recurrent mobility analyses [50], [22], [29]. On the other hand, Fig.17b shows the sensitivity of ΔN on T_w and T_s. In particular, the number of missed trips ΔN is minimized by larger values of T_w as larger values of T_s are considered. Specifically, the value of T_w that minimizes

the ΔN error is 10 minutes for $T_s = 5$ minutes, 20 minutes for $T_s = 20$ minutes, 30 minutes for $T_s = 60$ minutes, etc. Thus, given our choice of T_s equal to 20 minutes, the value of T_w has been set to 20 minutes in order to minimize the number of trips missed by TRANSIT with respect to the adopted benchmark method.



(a) Number of trips detected by TRANSIT

(b) Number of trips difference between GPS and TRANSIT

Figure 17: Sensitivity of TRANSIT on T_w and T_s in terms of: (a) volume of trips detected by TRANSIT; (b) error in the number of detected trips using [29] applied to $\mathcal{E}_{GPS}$ and TRANSIT applied to $\mathcal{E}_{NSD}$

References

[1] L. Sun, K. W. Axhausen, Understanding urban mobility patterns with a probabilistic tensor factorization framework, Transportation Research Part B: Methodological 91 (2016) 511–524. doi:10.1016/J.TRB.2016.06.011.
URL https://www.sciencedirect.com/science/article/pii/S0191261516300261

[2] X. Y. Yan, X. P. Han, B. H. Wang, T. Zhou, Diversity of individual mobility patterns and emergence of aggregated scaling laws, Scientific Reports 3. doi:10.1038/srep02678.

[3] S. Hasan, X. Zhan, S. V. Ukkusuri, Understanding urban human activity and mobility patterns using large-scale location-based data from online social media, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, New York, USA, 2013, p. 1. doi:10.1145/2505821.2505823.
URL http://dl.acm.org/citation.cfm?doid=2505821.2505823

[4] Y. Xu, R. D. Clemente, M. C. González, Understanding vehicular routing behavior with location-based service data, EPJ Data Science 10 (1) (2021) 1–17. doi:10.1140/epjds/s13688-021-00267-w.
URL https://doi.org/10.1140/epjds/s13688-021-00267-w

[5] C. Chen, J. Ma, Y. Susilo, Y. Liu, M. Wang, The promises of big data and small data for travel behavior (aka human mobility) analysis (2016). doi:10.1016/j.trc.2016.04.005.

[6] M. C. González, C. A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782. doi:10.1038/nature06958.
URL http://www.nature.com/doifinder/10.1038/nature06958

[7] M. S. Iqbal, C. F. Choudhury, P. Wang, M. C. González, Development of origin–destination matrices using mobile phone call data, Transportation Research Part C: Emerging Technologies 40 (2014) 63–74. doi:10.1016/J.TRC.2014.01.002.
URL https://www.sciencedirect.com/science/article/pii/S0968090X14000059/

[8] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, Z. Smoreda, A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas, IEEE Transactions on Mobile Computing 16 (10) (2017) 2682–2696. doi:10.1109/TMC.2016.2637901.
URL http://ieeexplore.ieee.org/document/7779102/

[9] A. Furno, M. Fiore, R. Stanica, Joint spatial and temporal classification of mobile traffic demands, in: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, IEEE, 2017, pp. 1–9. doi:10.1109/INFOCOM.2017.8057089.
URL http://ieeexplore.ieee.org/document/8057089/

[10] R. W. Douglass, D. A. Meyer, M. Ram, D. Rideout, D. Song, High resolution population estimates from telecommunications data, EPJ Data Science 4 (1) (2015) 4. doi:10.1140/epjds/s13688-015-0040-6.
URL http://www.epjdatascience.com/content/4/1/4

[11] Q. Xu, A. Gerber, Z. M. Mao, J. Pang, AccuLoc: Practical localization of performance measurements in 3G networks, in: MobiSys'11 - Compilation Proceedings of the 9th International Conference on Mobile Systems, Applications and Services and Co-located Workshops, ACM Press, New York, New York, USA, 2011, pp. 183–195. doi:10.1145/1999995.2000013.
URL http://portal.acm.org/citation.cfm?doid=1999995.2000013

[12] G. Chen, A. C. Viana, M. Fiore, C. Sarraute, Complete trajectory reconstruction from sparse mobile phone data, EPJ Data Science 8 (1) (2019) 30. doi:10.1140/epjds/s13688-019-0206-8.
URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-019-0206-8

[13] G. Ranjan, H. Zang, Z.-L. Zhang, J. Bolot, Are call detail records biased for sampling human mobility?, ACM SIGMOBILE Mobile Computing and Communications Review 16 (3) (2012) 33–44. doi:10.1145/2412096.2412101.
URL https://dl.acm.org/doi/10.1145/2412096.2412101

[14] D. Bachir, Estimating Urban Mobility with Mobile Network Geolocation Data Mining, Tech. rep. (2018).
URL https://mail.ifsttar.fr/service/home/~/?auth=co&loc=fr&id=2018&part=2.2

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

162–177. doi:10.1016/J.TRC.2015.04.022.
URL https://www.sciencedirect.com/science/article/pii/S0968090X15001631

[16] F. Calabrese, L. Ferrari, V. D. Blondel, Urban Sensing Using Mobile Phone Network Data: A Survey of Research, ACM Computing Surveys 47 (2) (2014) 1–20. doi:10.1145/2655691.
URL https://dl.acm.org/doi/10.1145/2655691

[17] M. Janzen, M. Vanhoof, Z. Smoreda, K. W. Axhausen, Closer to the total? Long-distance travel of French mobile phone users, Travel Behaviour and Society 11 (2018) 31–42. doi:10.1016/j.tbs.2017.12.001.

[18] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, M. C. González, Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities, Transportation Research Record 2526 (2015) 126–135. doi:10.3141/2526-14.
URL https://journals.sagepub.com/doi/10.3141/2526-14

[19] J. L. Toole, M. Ulm, M. C. González, D. Bauer, Inferring land use from mobile phone activity, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, New York, USA, 2012, pp. 1–8. doi:10.1145/2346496.2346498.
URL http://dl.acm.org/citation.cfm?doid=2346496.2346498

[20] L. E. Olmos, S. Çolak, S. Shafiei, M. Saberi, M. C. González, Macroscopic dynamics and the collapse of urban traffic, Proceedings of the National Academy of Sciences of the United States of America 115 (50) (2018) 12654–12661. doi:10.1073/pnas.1800474115.
URL www.pnas.org/cgi/doi/10.1073/pnas.1800474115

[21] G. Khodabandelou, V. Gauthier, M. Fiore, M. A. El Yacoubi, Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata, IEEE Transactions on Mobile Computingdoi:10.1109/TMC.2018.2871156.

[22] S. Jiang, J. Ferreira, M. C. Gonzalez, Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore, IEEE Transactions on Big Data 3 (2) (2017) 208–219. doi:10.1109/TBDATA.2016.2631141.
URL http://ieeexplore.ieee.org/document/7755745/

[23] Z. Zhao, S. L. Shaw, Y. Xu, F. Lu, J. Chen, L. Yin, Understanding the bias of call detail records in human mobility research, International Journal of Geographical Information Sciencedoi:10.1080/13658816.2015.1137298.

[24] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, C. Sarraute, Enriching sparse mobility information in Call Detail Records, Computer Communications 122 (2018) 44–58. doi:10.1016/J.COMCOM.2018.03.012.
URL https://www.sciencedirect.com/science/article/pii/S0140366417309234

[25] M. Li, S. Gao, F. Lu, H. Zhang, Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data, Computers, Environment and Urban Systems 77 (2019) 101346. doi:10.1016/j.compenvurbsys.2019.101346.

[26] S. Choi, H. Yeo, J. Kim, Network-Wide Vehicle Trajectory Prediction in Urban Traffic Networks using Deep Learning, Transportation Research Record 2672 (45) (2018) 173–184. doi:10.1177/0361198118794735.
URL https://journals.sagepub.com/doi/10.1177/0361198118794735

[27] J. Kim, H. S. Mahmassani, Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories, in: Transportation Research Procedia, Vol. 9, Elsevier, 2015, pp. 164–184. doi:10.1016/j.trpro.2015.07.010.

[28] W. Wu, Y. Wang, J. B. Gomes, D. T. Anh, S. Antonatos, M. Xue, P. Yang, G. E. Yap, X. Li, S. Krishnaswamy, J. Decraene, A. S. Nash, Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling, in: 2014 IEEE 15th International Conference on Mobile Data Management, IEEE, 2014, pp. 321–328. doi:10.1109/MDM.2014.46.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6916937

[29] P. Katsikouli, M. Fiore, A. Furno, R. Stanica, Characterizing and Removing Oscillations in Mobile Phone Location Data (2019) 1–9doi:10.1109/WoWMoM.2019.8793034{\"{i}}.
URL https://hal.inria.fr/hal-02110719

[30] B. C. Csáji, A. Browet, V. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, V. D. Blondel, Exploring the mobility of mobile phone users, Physica A: Statistical Mechanics and its Applications 392 (6) (2013) 1459–1473. doi:10.1016/J.PHYSA.2012.11.040.
URL https://www.sciencedirect.com/science/article/pii/S0378437112010059

[31] F. Asgari, A. Sultan, H. Xiong, V. Gauthier, M. A. El-Yacoubi, CT-Mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network, Computer Communicationsdoi:10.1016/j.comcom.2016.04.014.

[32] L. Bonnetain, A. Furno, J. Krug, N.-E. E. Faouzi, Can We Map-Match Individual Cellular Network Signaling Trajectories in Urban Environments? Data-Driven Study, Transportation Research Record: Journal of the Transportation Research Board 2673 (7) (2019) 74–88. doi:10.1177/0361198119847472.
URL http://journals.sagepub.com/doi/10.1177/0361198119847472

[33] M. Forghani, F. Karimipour, C. Claramunt, From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration, Transportation Research Part C: Emerging Technologies 117 (2020) 102666. doi:https://doi.org/10.1016/j.trc.2020.102666.
URL http://www.sciencedirect.com/science/article/pii/S0968090X20305817

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

[35] E. Trevisani, A. Vitaletti, Cell-ID location technique, limits and benefits: An experimental study, in: Proceedings - IEEE Workshop on Mobile Computing Systems and Applications, WMCSA, 2004, pp. 51–60. doi:10.1109/MCSA.2004.9.

[36] R. Ahas, A. Aasa, S. Silm, R. Aunap, H. Kalle, Mark, Mobile positioning in space-time behaviour studies: Social positioning method experiments in Estonia, Cartography and Geographic Information Science 34 (4) (2007) 259–273. doi:10.1559/152304007782382918.
URL https://www.tandfonline.com/doi/abs/10.1559/152304007782382918

[37] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, H. Hlavacs, Cellular Data Meet Vehicular Traffic Theory: Location Area Updates and Cell Transitions for Travel Time Estimation, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 361–370. doi:10.1145/2370216.2370272.
URL https://doi.org/10.1145/2370216.2370272

[38] I. Leontiadis, A. Lima, R. Stanojevic, H. Kwak, D. Wetherall, K. Papagiannaki, From cells to streets: Estimating mobile paths with cellular-side data, in: CoNEXT 2014 - Proceedings of the 2014 Conference on Emerging Networking Experiments and Technologies, Association for Computing Machinery, Inc, New York, NY, USA, 2014, pp. 121–132. doi:10.1145/2674005.2674982.
URL https://dl.acm.org/doi/10.1145/2674005.2674982

[39] Y. Zhao, Z. Zhou, X. Wang, T. Liu, Z. Yang, Urban scale trade area characterization for commercial districts with cellular footprints, ACM Transactions on Sensor Networks 16 (4) (2020) 1–20. doi:10.1145/3412372.
URL https://dl.acm.org/doi/10.1145/3412372

[40] L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, L. Bravo, An individual-level ground truth dataset for home location detection, Tech. rep. (2020).

[41] Y. Zhao, X. Wang, J. Li, D. Zhang, Z. Yang, CellTrans, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3 (3) (2019) 1–26. doi:10.1145/3351283.
URL https://dl.acm.org/doi/10.1145/3351283

[42] Z. Qin, Z. Fang, Y. Liu, C. Tan, W. Chang, D. Zhang, EXIMIUS: A measurement framework for explicit and implicit urban traffic sensing, in: SenSys 2018 - Proceedings of the 16th Conference on Embedded Networked Sensor Systems, Association for Computing Machinery, Inc, New York, NY, USA, 2018, pp. 1–14. doi:10.1145/3274783.3274850.
URL https://dl.acm.org/doi/10.1145/3274783.3274850

[43] Z. Qin, F. Cao, Y. Yang, S. Wang, Y. Liu, C. Tan, D. Zhang, CellPred: A behavior-aware scheme for cellular data usage prediction, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4 (1) (2020) 1–24. doi:10.1145/3380982.
URL https://dl.acm.org/doi/10.1145/3380982

[44] C. Zhao, A. Zeng, C. H. Yeung, Characteristics of human mobility patterns revealed by high-frequency cellphone position data, EPJ Data Science 10 (1) (2021) 5. doi:10.1140/epjds/s13688-021-00261-2.
URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00261-2

[45] Y. Song, Y. Liu, W. Qiu, Z. Qin, C. Tan, C. Yang, D. Zhang, Miff Human mobility extractions with cellular signaling data under spatio-Temporal uncertainty, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4 (4) (2020) 1–19. doi:10.1145/3432238.
URL https://dl.acm.org/doi/10.1145/3432238

[46] Z. Shen, W. Du, X. Zhao, J. Zou, DMM, in: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, ACM, New York, NY, USA, 2020, pp. 1–14. doi:10.1145/3372224.3421461.
URL https://dl.acm.org/doi/10.1145/3372224.3421461

[47] D. Bachir, V. Gauthier, M. E. Yacoubi, G. Khodabandelou, Using mobile phone data analysis for the estimation of daily urban dynamics, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2017, pp. 626–632. doi:10.1109/ITSC.2017.8317956.
URL http://ieeexplore.ieee.org/document/8317956/

[48] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of Predictability in Human Mobility, Science 327 (5968) (2010) 1018–1021. doi:10.1126/science.1177170.
URL https://science.sciencemag.org/content/327/5968/1018

[49] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, M. C. González, Unravelling daily human mobility motifs, Journal of The Royal Society Interface 10 (84) (2013) 20130246. doi:10.1098/rsif.2013.0246.
URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0246

[50] M. Fekih, T. Bellemans, Z. Smoreda, P. Bonnel, A. Furno, S. Galland, A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France), Transportation (2020) 1–32doi:10.1007/s11116-020-10108-w.
URL https://doi.org/10.1007/s11116-020-10108-w

[51] A. A. Taha, A. Hanbury, An Efficient Algorithm for Calculating the Exact Hausdorff Distance, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (11) (2015) 2153–2163. doi:10.1109/TPAMI.2015.2408351.