Multi-modal Fine-grained Map-matching of Mobile Phone Network Signaling Data in Urban Areas

July, 31st 2021

Loïc Bonnetain[1],
[1] Univ. Lyon, IFSTTAR, ENTPE, LICIT UMR_T9401, F-69675, Lyon, France
loic.bonnetain@univ-eiffel.fr

Angelo Furno[1],
[1] Univ. Lyon, IFSTTAR, ENTPE, LICIT UMR_T9401, F-69675, Lyon, France
angelo.furno@univ-eiffel.fr

Nour-Eddin El Faouzi[1]
[1] Univ. Lyon, IFSTTAR, ENTPE, LICIT UMR_T9401, F-69675, Lyon, France
nour-eddin.elfaouzi@univ-eiffel.fr

Submited for the Transportation Research Board 2022 annual conference

Word Count:

Number of words: 6998
Number of tables: 2 (250 words each)
Total: 7498

Abstract

Network signaling data have a great potential for human centric mobility analysis at large scale. But their intrinsic characteristics, such as sparsity (in time and space), noise and large localization error limit their applicability for detailed studies on mobility, especially in urban scenarios. In this paper, we propose a framework able to infer fine-grained spatio-temporal information at both microscopic and macroscopic scales from users' signaling traces. The framework leverages TRANSIT, a pre-processing approach, that outputs mobile sessions (possibly enhanced in space and time) from the raw signaling traces of users. This set of mobile sessions feed an Hidden Markov Model based map-matching capable of inferring the route traveled by the mobile subscriber with high accuracy: a geographical error around 60 m, a matching rate of 77% and a F1 score of 0.77. Such a promising results are made possible by leveraging the trajectory enhancement step of TRANSIT and relying on the assumption of having a coarse transportation mode knowledge before the map-matching algorithm. This assumption is discussed and a preliminary solution is proposed. The whole framework is evaluated in a case-study based on real signaling traces collected by a major French operator in the city of Lyon. We validated our approach at both microscopic and macroscopic levels.

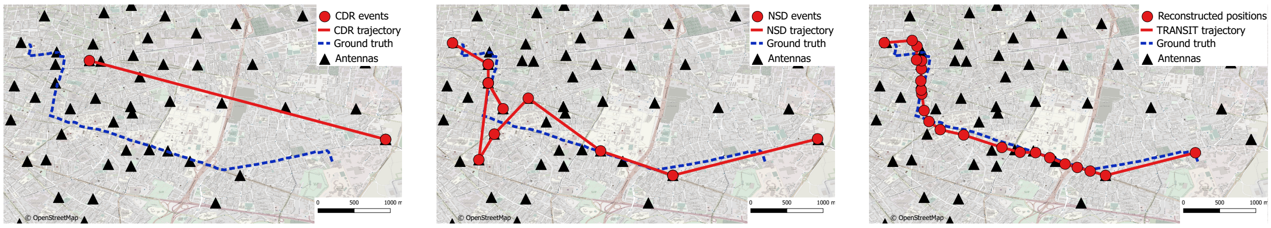
Keywords: **Map-matching, Mobile phone, Hidden Markov Model, Multi-modal Transport**

1. INTRODUCTION

In recent years, mobile phone data and especially Call Detail Records (CDR) have demonstrated a great potential for large scale human mobility studies. This emerging source of data has been leveraged for analyzing human movements at unprecedented spatio-temporal scales [1], modeling the general laws governing human mobility [2], reconstructing Origin-Destination (OD) matrices [3] and understanding urban land use [4, 5]

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 [6]. 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 [7]. In this paper, the objective is to reconstruct fine-grained mobility information from cell phone trajectories with high accuracy, and, more specifically, to infer the route traveled by a mobile subscriber over the multi-modal transportation network. Traditional CDR are not suitable to address such a task, due to their limited spatial resolution and sampling frequency. For instance, Fig. 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 as recorded via a Global Positioning System (GPS) tracking app (dotted blue). The figure makes it clear that inferring the actual movement from CDR is an arduous mission. Given these limitations, extended variants of CDR, namely, Network Signaling logs Data (NSD) are currently collected by network providers and investigated by the research community. Differently from CDR data, NSD report on multiple kinds of events besides calls and text messages (*e.g.*,, IP protocol message exchanges, hand-overs, location updates, etc.) thus increasing the spatio-temporal sampling frequency of mobile phone passive data. Research on this kind of data is however still at early stages. In our previous work [8], we proposed a framework, named TRANSIT, for processing network signaling trajectories and returning augmented individual mobility trajectories. By leveraging the regularity of human mobility, TRANSIT is capable of increasing the spatio-temporal accuracy of the trajectories. For instance, Fig. 1b and Fig.. 1c show respectively the raw NSD trajectory and the trajectory enhanced with TRANSIT. The framework is presented in more details in Sec. 3.2.

In this paper, we extend our previous work by relying on a map-matching solution that allows to infer accurately, from mobile network signaling trajectories, possibly enhanced by TRANSIT, the route traveled on the multi-modal transportation network.



(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.

To achieve this purpose, in the context of this work, we consider the hypothesis of having a coarse knowledge of the transportation mode selected by the user (road, public transport or train) before the map-matching process. A Hidden Markov Model (HMM) based map-matching is investigated. Our results show that the whole approach is capable of map-matching with high accuracy both raw and TRANSIT-enhanced trajectories. The proposed solution achieves a matching rate of 77%, with a spatial accuracy of $60m$. In addition, when the TRANSIT framework is applied before the map-matching step, the matching rate increases by 10%, while the spatial accuracy grows by a factor 2 compared to map-matching directly raw signaling trajectories.

The key contributions of this work are thus the following:

- A novel solution to the challenging problem of mapping cellular trajectories to the multimodal transportation network in urban settings.
- A complete framework for reconstructing and map-matching fine-grained mobility information from raw network signaling data.
- The analysis of a unique dataset composed of both real-world cellular and GPS trajectories related to a group of users in the Lyon metropolitan area, France.

2. RELATED WORK

Map-matching is a well-known operation for improving positioning accuracy by integrating positioning data with spatial transportation data to identify the correct link on which a mobile object is traveling [9]. The problem of map-matching GPS traces to the road network has been largely investigated by the research community. Quddus *et al.* [10] categorize map-matching approaches in four classes. The first one includes geometric approaches where matching is performed based on the spatial distance between the GPS points and the candidates links in the road network [11]. The second class comprises more sophisticated, topological approaches that use geometric information, like in the geometric approaches, but in combination with topological information such as the existence of connectivity between nodes of the network [12]. Very sensitive to noise and outliers, these approaches are not appropriate in presence of highly noisy and sparse data. The third kind of approaches exploits probabilistic methods: a confidence region around the location of the moving object is defined. Then, candidate network links are identified as those present in this confidence region. The evaluation of the candidates is based on the geometrical criteria. Finally, the last category leverages more complex mathematical tools. A non exhaustive list of these methods includes, *i.e.*,, the Kalmam Filter [13], Dempster–Shafer theory [12] or fuzzy logic models [14].

These state-of-the-art algorithms may achieve very high accuracy (location error lower than 10 meters) when used with high-sampling-rate GPS data. Newson *et al.* [15] first introduce HMM-based map-matching dealing with different GPS traces sampling rate. Their approach appears to be more robust and accurate with sparse and noisy trajectories than standard advanced map-matching solutions for high sampling rate data.

With the emergence of large-scale mobile phone data, map-matching of cell phone trajectories has recently became a relevant problem studied by the research community. The peculiar features of mobile network data, such as sparsity (in time and space), noise and large localization error, make the task of map-matching cell phone trajectories highly challenging. Most of the approaches used with cellular trajectories are based on those traditionally designed for GPS map-matching. Schulze *et al.* [16] use a probabilistic approach: their solution restricts the set of admissible routes to a corridor by estimating the area within which a user is allowed to travel and infers path using the shortest path on candidate routes. With only 55% of correct matches, this method has been outperformed by a HMM-based approach recently developed by Jagadeesh *et al.* [17], which reaches 75% of median accuracy.

HMM-based map-matching has thus become the state-of-the-art for noisy and sparse location data and, specifically, for mobile phone data traces. Reham *et al.* [18], Thiagarajan *et al.* [19] and, more recently, Algizawy *et al.* [20] developed supervised HMM models exhibiting good accuracy (75% for Thiagarajan *et al.* approach). Jagadeesh *et al.* [17] proposed an online map-matching algorithm combining HMM-based map-matching and route-choice model. Recently, Shen *et al.* [21] proposed a an approach based on recurrent neural networks obtaining a precision of 80% and a recall of 85%. However, these approaches have two main drawbacks. They leverage supervised machine-learning solutions that require a large amount of labeled cellular trajectories for training the parameters of the models. Ground-truth precise trajectories (*i.e.*, the labels) are very hard to obtain, especially when dealing with highly dynamic and irregular environments, such as urban areas. Moreover, most of the approaches match the cellular trajectories only to road networks, without considering other sub-networks corresponding to alternative transportation modes, such as tramway, subway, bus, etc.

Among the very few exceptions, it is worth mentioning the methodology recently proposed by Asgari *et al.* [22]. Their solution, namely CT-Mapper, relies on an unsupervised HMM model, which aims at mapping sparse cellular trajectories to a multi-layer transportation network. Similarly, in our previous work [23], we also studied unsupervised HMM-based map-matching for solving the same problem. However, differently from [22], we focused on the more complex problem of map-matching mobile phone signaling traces in urban environments considering a larger variety of urban transportation modes, while CT-mapper only focused on three modes: road, subway and railway.

This work differs compared to the aforementioned works. By building on our previous work, we propose an unsupervised HMM-based map-matching solution for network signaling traces in urban settings. However, we modified the HMM-based framework so that both raw signaling trajectories (a sequence of antennas) and TRANSIT-enhanced trajectories (a sequence of reconstructed positions) could be map-matched with high accuracy. As discussed in the evaluation section, the integration of TRANSIT-enhanced trajectories allows to improve significantly the performance of our map-matching solution. Moreover, we collected a unique set of GPS and signaling trajectories where transportation mode has been manually labeled. Such a rich dataset allows validating the map-matching output more precisely compared to previous works.

3. FRAMEWORK

3.1. Case study and Network Modeling

This paper considers the multi-modal transportation network of the city of Lyon (France) as a case study. This network includes multiple transportation modes, *i.e.*, road, subway, tramway, bus and train. In the following, we make the assumption that the transportation modes available to travelers can be generally classified into three categories: road, public transport (subway, tramway and bus) and train. Thus, the network is modeled as a graph G composed of three sub-graphs G_{road}, G_{tc} and G_{train} that are assumed to be not connected between each other. In order to move from one graph to another the user will need to spend a period of *immobility* at a given location (*e.g.*, in the proximity of a bus stop, *i.e.*, under the coverage of a limited subset of cellular antennas), which, if long enough, will be detected by TRANSIT as a static activity. As a consequence, our framework will consider two different trips (before and after the modal-shift static session) that can be separately matched to the specific sub-graph. Inter-modal trips are therefore not possible across the three sub-graphs, but can occur within the transit network, *i.e.*, G_{tc}, which covers three different transportation modes (*i.e.*, subway, tramway and bus). For the rest of the paper, we will denote as G_j the generic sub-graph of G related to a specific category of transport modes, *i.e.*, either G_{road}, G_{tc} or G_{train}.

The graph and its different sub-graphs are built using multiple data sources and programming tools. The road sub-network G_{road} is generated via OSMnx [24], a Python library which creates NetworkX graphs from OpenStreetMap (OSM) data. Public transport sub-network G_{tc} has been generated using GTFS (Google Transit Feed Specification) data. The public transport sub-network G_{tc} is modeled as a multi-layer graph including three layers corresponding to the three transportation modes (subway, tramway and bus). Between public transport layers, cross-layers edges are defined as connections at transfer stops between public transport lines (this information is contained in the GTFS transfer file). Finally, the train sub-network G_{train} is derived using the geometry of train lines available as open data[1]. The nodes of the G_{train} sub-network correspond to stations of the railway system of the city of Lyon.

Similarly to what is proposed in the work of Putra *et al.* [25], whenever the distance between a pair of adjacent nodes (from both the transit and the train sub-networks) is larger than a given threshold D_{inter}, additional nodes have been added via linear interpolation of the x-y coordinates of the considered pair of adjacent nodes. Such an interpolation can be considered as a reasonable approximation of the real geometry of the link connecting the pair of nodes, which is hard to take into account during the map-matching process. In particular, D_{inter} is set to $200m$ for G_{tc} and $500m$ for G_{train} so that the distance between adjacent nodes is, on average, in the same order of magnitude for the three sub-graphs G_{road}, G_{tc} and G_{train}. More details in the importance of adding interpolated nodes in the transportation network can be found in the work of Putra *et al.* [25]. Some statistics of the final multi-modal transportation network G is given in Table 1.

Layer	Mode	$\|V\|$	$\|E\|$	$\langle k \rangle$	$\langle l \rangle$ (km)
G_{road}	Road	27213	58593	4.3	0.13
G_{tc}	Bus	31072	41755	2.7	0.15
	Subway	636	669	2.3	0.17
	Tramway	2239	2790	2.6	0.16
	All modes	34033	46458	2.7	0.15
G_{train}	Train	1657	1725	2.1	0.46
G	All modes	59942	102073	/	0.14

Table 1: Main characteristics of each transportation layer of G: number of nodes $|V|$, number of edges $|E|$, average node degree $\langle k \rangle$ and average edge length in kilometer $\langle l \rangle$.

3.2. TRANSIT

TRANSIT is a framework capable of processing raw cell phone NSD trajectories to accurately distinguish mobility phases from stationary activities for individual mobile devices, and reconstruct fine-grained human mobility trajectories. In the following, we provide a short summary of the TRANSIT framework for the sake of readability. The interested reader can refer to our previous work [8] for more details.

TRANSIT receives as input the set of NSD events, *i.e.*, a *NSD trace*, of a mobile device i that we assume to correspond to a single user. A user's NSD trace is thus denoted by $\mathcal{T}^i = \{e^i_1, \ldots, e^i_n, \ldots, e^i_{N_i}\}$, where e^i_n is the n^{th} NSD event recorded for device i. Each NSD event is the result of a communication activity between 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

[1] https://www.data.gouv.fr/fr/datasets/fichier-de-formes-des-lignes-du-reseau-ferre-national/

$e_n^i = (c_n^i, t_n^i)$, 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, our approach processes $\mathcal{T}^i$ to produce two outputs in succession, as follows.

Trajectory segmentation. 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. This step leverages a combination of thresholds and heuristics to: (i) identify candidate antennas for static activities; (ii) detect sufficiently long sequences of events that take place at candidate antennas. The approach also includes an oscillation removal process and a spatial clustering method to refine the location of adjacent static sessions. This phase 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*.

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 spatio-temporal accuracy of the trajectories. The approach exploits a spatio-temporal DBSCAN clustering procedure, which leverages the well-known Hausdorff distance to identify and merge spatially similar traces collected over multiple days of a subscriber's observation. Based on the result of DBSCAN, $\mathcal{M}^i$ can be thus divided 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 in two steps. Firstly, the trajectories in a same cluster are temporally scaled (*i.e.*, stretched or compressed) in time so as to match the average travel duration for the cluster. Secondly, the scaled trajectories within the same cluster 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 for each cluster. If there is no event within a particular time slot, the resulting enhanced trajectory will have missing positions. All original trajectories in a given cluster are then mapped 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. The resulting set of mobile, possibly augmented, trajectories is denoted as $\widehat{\mathcal{M}^i}$ with $\widehat{\mathcal{M}^i} = \widehat{\mathcal{M}_R^i} \cup \mathcal{M}_O^i$. An example of such augmented trajectories is reported in Fig. 1c.

Ultimately, the output of TRANSIT is: (i) the set $\mathcal{A}^i$ of static activity sessions of user i, and (ii) the set $\widehat{\mathcal{M}^i}$ of mobile sessions with augmented trajectories. This set $\widehat{\mathcal{M}^i}$ is the input of our HMM based map-matching.

3.3. HMM based map-matching

The following section reports on the main methodological background characterizing our solution to perform map-matching of cellular network trajectories from mobile phone passive data, with the assumption of knowing, beforehand, a coarse approximation of the transport mode (*i.e.*, road, train, public transport sub-network) for a given user's mobility trace. A Hidden Markov Model can be defined by a 5-tuple $\langle S, O, I, T, E \rangle$, with $S = \{s_0, \ldots, s_{N-1}\}$ representing a finite set of states; $O = \{o_0, \ldots, o_{M-1}\}$ corresponding to a finite set of observations; I being the probability distribution of the initial states; T representing a set of transition probability. The probability to transit from hidden state s_m to hidden state s_n is denoted as $t(s_m, s_n)$. Finally, E is a set of emission probability. The probability to emit observation o_k from hidden state s_m is denoted as $e(s_m, o_k)$.

Our map-matching problem on $\widehat{\mathcal{M}^i}$ can be modeled as a Hidden Markov Model with the following formulation. Hidden states are modeled as the set of vertices (nodes) of the generic transportation sub-network G_j. Emissions are modeled as the unique set of x-y coordinates in $\widehat{\mathcal{M}^i}$. This set is composed of antennas coordinates from cellular network in $\mathcal{M}^i_O$ and reconstructed positions in $\widehat{\mathcal{M}^i_R}$.

Ultimately, the Hidden Markov Model allows solving the following problem: given a sequence of observations, *i.e.*, sequence of antennas for $\mathcal{M}^i_O$ and reconstructed positions for $\widehat{\mathcal{M}^i_R}$, the model finds the most likely sequence of hidden states, *i.e.*, sequence of nodes on the transportation sub-network G_j.

In the following, we will define the main (hyper-)parameters of the HMM: the initial, transition and emission probabilities.

3.3.1. Initial Probability

All the nodes of the transportation network are initially equally assigned with a probability of $1/N$ with N representing the total number of nodes in the transportation network:

$$\pi(s_m) = \frac{1}{N} \tag{1}$$

3.3.2. Transition Probability

The transition probability corresponds to the probability that a mobile phone user moves on the underlying transportation network from hidden state s_m at time $t-1$ to hidden state s_n at time t. In the following, we choose the definition proposed by Putra et al. [25], *i.e.*, the transition probability depends on the travel time over an edge. For the public transport and railway sub-networks, the travel time of each edge is calculated by multiplying the speed (the speeds are defined by mode[2] and the edge distance (geodesic distance between the two nodes of the edge). For the road network, the travel time corresponds to the free flow travel time on each road segment, as available from the OpenStreetMap information. Additionally, for public transport, cross-layers edges connecting the different lines and modes are associated to a travel time that corresponds to a typical connecting time, which is set to 5 minutes.

Finally, the transition probability $t(s_m, s_n)$ between the generic pair of nodes s_m and s_n is defined to be exponentially decreasing according to the travel-time weighted shortest path between the two nodes s_m and s_n. Formally:

$$t(s_m, s_n) = \exp^{-\beta \cdot tt_{s_m,s_n}}, \quad tt_{s_m,s_n} = \textstyle\sum_{\forall (s_u,s_v) \in SP_{mn}} tt_{s_u,s_v} \tag{2}$$

where (s_u, s_v) is the generic edge on the travel-time weighted shortest path SP_{mn} connecting the two nodes s_m and s_n in sub-graph G_j, computed via the Dijkstra algorithm. The length of the weighted shortest path SP_{mn} corresponds to the sum of the travel time over each edge (s_u, s_v) belonging to SP_{mn}. tt_{s_u,s_v} denotes the travel time between each two nodes s_u and s_v. β is a damping factor to control the effect of the travel time.

[2] Speeds on the road network depend on the OpenStreetMap type of route, it varies from 30 km/h to 90km/h. For the subway, the tramway and the bus the speed is respectively 30 km/h, 15 km/h and 15 km/h.

3.3.3. Emission Probability

When the trajectory is a sequence of reconstructed positions, the map-matching problem can be viewed as a map-matching problem with noisy GPS points. Similarly to [15], we model the emission probability as a Gaussian noise centered on the hidden state s_m and an empirically estimated standard deviation of the distance error between hidden states and observations:

$$e(s_m, o_k) = \frac{1}{\sqrt{2\pi}\alpha} e^{-0.5\left(m\frac{d_{s_m,o_k}}{\alpha}\right)^2} \tag{3}$$

where d_{s_m,o_k} is the geodesic distance between the generic observation o_k and the generic node s_m, while α is the standard deviation of a Gaussian random variable associated to the error distance between the reconstructed and the real position of the mobile. More details on the estimation of this parameter are given in Sec. 4.1.2.

3.4. Map-matching algorithm

As a pre-processing step, for the set of raw signaling trajectories not enhanced by TRANSIT, we re-sample the network signaling traces with a three minutes frequency. In space, we calculate the centroid of the coordinates of the signaling events falling in the three minutes time-window. In time, we associated each time window to its starting time. This aggregation step aims at reducing the oscillation effect on the cellular trajectory.

After the pre-processing step, our approach performs a two-steps map-matching procedure. The first phase consists in an optimized Viterbi algorithm [26]. The inputs of the Viterbi process are the following: the generic transportation sub-network modeled as a graph G_j, the possibles states (set of the nodes of G_j), the emissions (the unique set of x-y coordinates in $\widehat{\mathcal{M}^i}$), the previously defined HMM parameters and the input trajectory from $\widehat{\mathcal{M}^i}$. By calculating all possible paths given the input trajectory, the Viterbi process output is the most likely sequence of graph nodes, one for each time instant in the input. For real-time application, due to a large number of states and emissions, the execution time of the Viterbi algorithm is critical [20]. To improve performance, we implemented an optimized version of the Viterbi algorithm as done by Algizawy *et al.* [20] which consists in eliminating all multiplications by zero thus reducing the search space by keeping only emittable states from each observable state. Moreover, to further reduce computation time, the following approximations are considered: (i) if the distance between state s_m and observation o_k is larger than $2km$, the emission probability $e(s_m, o_k)$ is rounded to 0; (ii) similarly, if the distance between state s_m and state s_n is larger than $5km$, transition probability $t(s_m, s_n)$ is rounded to 0.

It is worth observing that, after inferring the most likely states sequence using the optimized Viterbi implementation presented above, the output sequence of hidden states (nodes on a given sub-network G_j) do not necessarily form a connected path on the specific transport sub-network. Therefore, as the second step of the map-matching procedure, the final trajectory is further completed by applying a traditional shortest path (Dijkstra) detection algorithm on the underlying transportation graph between any two consecutive nodes of the most likely states sequence. The final completed sequence of nodes on sub-network G_j represents the map-matched trace for the processed trace from $\widehat{\mathcal{M}^i}$ for user i.

4. EVALUATION

4.1. Microscopic validation

4.1.1. Datasets

The dataset used in our microscopic validation was collected for four Orange subscribers who voluntarily agreed to be monitored by a GPS tracking application 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 pseudonymized, and accessed by authorized personnel of the research team only. The combined GPS and NSD data of the four users were collected during a continued period of three months, March 15^{th} and June 15^{th} 2019, in the city of Lyon, France.

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.

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 subscribers at the time of the data collection campaign, and (*ii*) they were explicitly invited to maintain their regular mobile communication and service consumption habits during the measurement period.

Overall, the validation datasets $\mathcal{E}_{GPS}$ and $\mathcal{E}_{NSD}$ provide corresponding GPS and NSD data. We applied a recent segmentation approach for spatio-temporal GPS data [27] for the traces in $\mathcal{E}_{GPS}$. The resulting set of trajectories is denoted as M_{GT}. TRANSIT is applied on $\mathcal{E}_{NSD}$ and outputs the set $\widehat{\mathcal{M}}$ of mobile sessions with augmented trajectories which is the input of our map-matching approach. Then, we manually labeled the transport mode of all trajectories in M_{GT} by associating one sub-graph G_j of G for each trajectory. In total, ground-truth data contain 111 trajectories related to public transport, 72 to car and 12 to train, for a total of 195 trajectories.

It is worth highlighting that the choice of the parameters of the Hidden Markov Model presented in Sec. 3.3 makes our map-matching approach suitable for GPS trajectories according to recent works on GPS map-matching [15, 25]. The only difference with the map-matching of signaling trajectories concerns the definition of emissions. For GPS trajectories, they are defined as the unique set of x-y coordinates in M_{GT}. Based on the extremely high accuracy (above 95%) of the map-matching process on GPS trajectories [15, 25], we consider the set of map-matched GPS trajectories as ground-truth in the evaluation.

Finally, we apply our map-matching approach on different NSD-based sets of trajectories, specifically: M, $\mathcal{M}_R$, $\widehat{\mathcal{M}}_R$ and $\widehat{\mathcal{M}}$, defined as follows. M is the set of signaling trajectories that the trajectory segmentation step of TRANSIT outputs, prior to any further processing. $\mathcal{M}_R$ is the set of recurrent trajectories identified by TRANSIT, without any trajectory enhancement. $\widehat{\mathcal{M}}_R$ is the set of recurrent trajectories that have received a trajectory enhancement by TRANSIT. $\widehat{\mathcal{M}}$ is the whole set of trajectories, possibly augmented, produced at the end of TRANSIT. These sets of trajectories covers the four volunteer users considered in the microscopic validation.

4.1.2. Parameter choice

Our map-matching approach depends on two parameters, namely α and β, respectively associated to the emission and transition probabilities of the Hidden Markov Model. In order to choose the best values for such parameters, we apply our approach on M and $\widehat{\mathcal{M}}_R$ and then compute the average F1 score for different combinations of values for α and β, using the corresponding map-matched GPS trajectories as ground-truth. The sensitivity analysis on M and $\widehat{\mathcal{M}}_R$ is respectively considered for choosing the best parameters for applying map-matching on raw signaling trajectories and TRANSIT enhanced trajectories. The F1 score is a metric for evaluating the performance of the map-matching at the trajectory level. To evaluate the performance on a set of trajectories, we average the F1 score obtained for each NSD-based trajectory. This score is defined as follows:

$$\text{Precision} = \frac{TP}{(TP+FP)}, \quad \text{Recall} = \frac{TP}{(TP+FN)} \quad \text{and} \quad \text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{4}$$

where: (i) the number of true positives TP is the number of edges in common between the ground-truth GPS and NSD map-matched trajectories; (ii) the number of false positives FP represents the number of edges in the NSD map-matched trajectory that do not belong to the corresponding ground-truth GPS map-matched trajectory; (iii) the number of false negatives FN represents the number of edges from the ground-truth GPS map-matched trajectory that do not belong to the NSD map-matched one.

Fig. 2 shows the sensitivity analysis on the parameters α and β of our approach on the two sets of NSD-based trajectories $\mathcal{M}$ and $\widehat{\mathcal{M}_R}$ for two transportation sub-networks: road and public transport. We do not dispose of enough trajectories to conduct a sensitivity analysis on train trips. We recall that $\mathcal{M}$ is the set of raw signaling trajectories identified after the TRANSIT segmentation step and $\widehat{\mathcal{M}_R}$ is the set of recurrent trajectories as enhanced by TRANSIT. From the figure, it can be noted that both the nature of the transportation mode and the enhancement performed by TRANSIT have a relevant impact for the optimal choice of α and β. Thus, the sensitivity analysis appears necessary for choosing the most appropriate combination of values for parameters α and β and thus for optimizing the performance of the map-matching procedure. In addition, all the figures exhibit a very similar trend: performance globally grows when both alpha and beta grow. Finally, the optimal values of β and α are located around the yellow diagonal of the two heatmaps in Fig. 2 (higher values of F1 score). Based on such results, we choose the following settings for the parameters: for road raw signaling trips, $(\alpha, \beta) = (0.75, 250)$; for public transport raw signaling trips, $(\alpha, \beta) = (0.5, 500)$; for road TRANSIT enhanced trips, $(\alpha, \beta) = (0.5, 250)$; and for public transport TRANSIT enhanced trips, $(\alpha, \beta) = (0.25, 100)$. For the train sub-network, we make the assumption that this transport mode is more similar to the public transport one than to the car-mode. Thus, we decided to adopt, for this mode, the same parameters as those used for the public transport one.

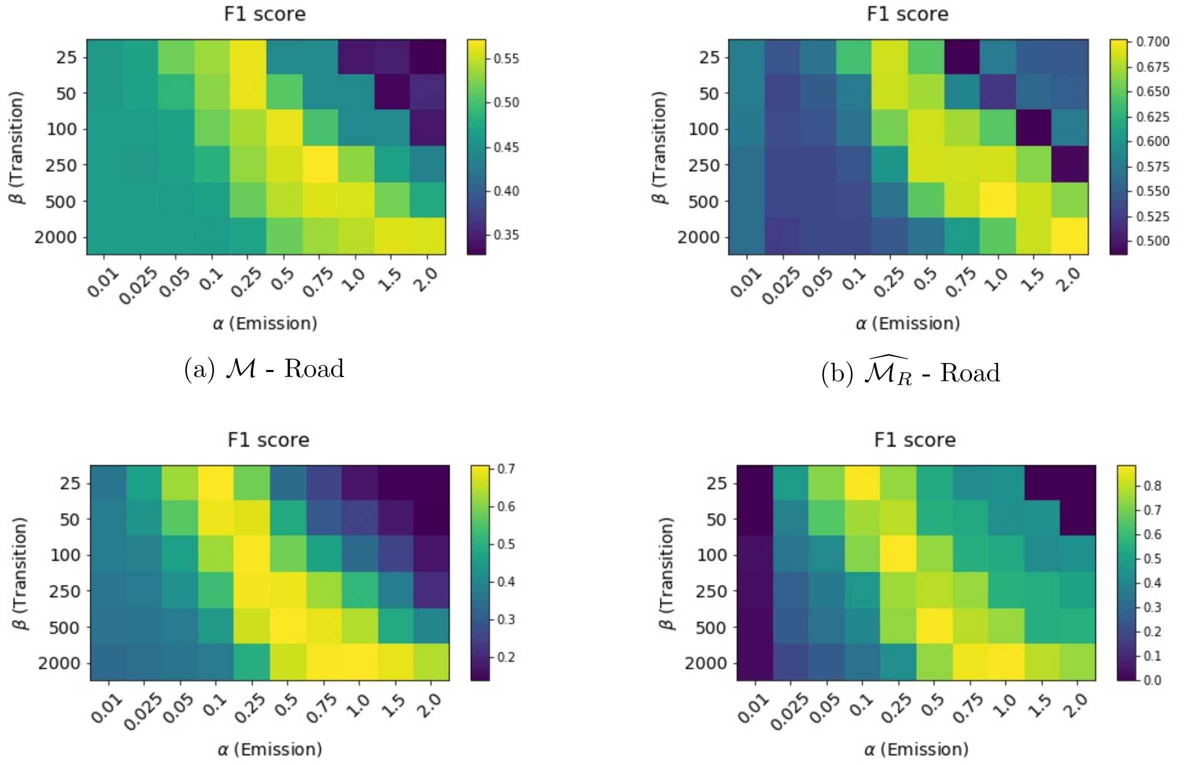
4.1.3. Map-matching performance

Once the parameters set, to evaluate the map-matching performance, we use two additional metrics, which complement the information provided by the F1 score *i.e.*, the matching rate and the geographical error. The matching rate, MR is the percentage of correctly map-matched edges by our approach. The geographical error, G_e is the distance between the NSD-based map-matched trajectory and the GPS-based one: it is computed as the average geodesic distance between each node in the inferred trajectory from NSD data and its closest node in space from the GPS map-matched trajectory. Formally:

$$MR = \frac{TP}{TP + FN + FP} \quad \text{and,} \quad G_e = \frac{1}{|m_{NSD}|} \sum_{e_n \in m_{NSD}} \min_{e_{n'} \in m_{GPS}} d(l_n, l_{n'}) \tag{5}$$

where TP, FP and FN correspond respectively to the number of true positives, false positives and false negatives as defined above. m_{GPS} and m_{NSD} are, respectively, two map-matched trajectories (sequence of nodes in the transportation network) from GPS and mobile network data, respectively. The operator $|\cdot|$ denotes the cardinality of the argument set, *i.e.*, the number of samples contained in the trajectory, while the operator $d(\cdot, \cdot)$ denotes the geodesic distance.

In our evaluation, we compare the result of the map-matching procedure with and without prior knowledge on the transportation mode. In case the map-matching is done without any prior knowledge on transportation mode, we map-matched the trajectories to all the sub-graphs of G, and we output the one with the highest probability from the Viterbi algorithm. Table 2 reports on the performance of our map-matching approach on 5 different input sets of trajectories, namely $\widehat{\mathcal{M}}$ without prior knowledge on the transportation mode and M, $\mathcal{M}_R$, $\widehat{\mathcal{M}_R}$ and $\widehat{\mathcal{M}}$ with prior



(a) $\mathcal{M}$ - Road

(b) $\widehat{\mathcal{M}_R}$ - Road

(c) $\mathcal{M}$ - Public transport

(d) $\widehat{\mathcal{M}_R}$ - Public transport

Figure 2: Sensitivity analysis on parameters α and β for raw signaling trajectories with road (a) and public transport (c); for transit enhanced trajectories with road (b) and public transport (d)

knowledge on the transportation mode. The results clearly highlights the importance of adding a prior information on transportation knowledge in order to improve the overall performance of the map-matching approach. The improvement is particularly relevant in relation to the matching rate, allowing an increase of 13% with respect to the case without any prior knowledge. Similarly, the F1 score increases from 0.59 without prior transportation mode knowledge, to 0.77 when considering it. In the following, we thus assume to dispose of transportation mode information before applying our map-matching approach. We can observe that despite the large uncertainty of NSD, our approach can map-match the NSD trajectories rather accurately. For the whole set of mobile sessions, the geographical error is in fact equal to only $60m$, matching attains a 77% success rate and the F1 score equals 0.77. By comparing the performance of the map-matching on $\widehat{\mathcal{M}_R}$ and $\mathcal{M}_R$ we are also able to appreciate the positive impact of TRANSIT on the map-matching performance. The results show that the enhancement step performed by TRANSIT on $\mathcal{M}_R$ allows improving significantly the map-matching process: the geographical error is divided by a factor 2, the matching rate increase by 10% and the F1 score reaches 0.80 on $\widehat{\mathcal{M}_R}$ (with TRANSIT) instead of 0.64 on $\mathcal{M}_R$ (without TRANSIT). Then, in the worst case, *i.e.*, for the set M of trajectories which are not enhanced by TRANSIT, the performance of the map-matching remains good with a geographical error inferior than $100m$ and a matching rate of 71%. Finally, it can be noticed that the result of the map-matching approach is superior, with respect to all considered metrics, in the public transport sub-network case than the road one. Such a result is explained by the more complex topology of

Set of trajectories	Transportation mode knowledge	Mode	Ge (km)	MR	F1 score
$\widehat{\mathcal{M}}$	No	All modes	0.11	63	0.59
M	Yes	Road	0.14	63	0.57
		TC	0.05	78	0.71
		All	0.09	71	0.65
$\mathcal{M}_R$	Yes	Road	0.14	60	0.56
		TC	0.05	78	0.70
		All modes	0.09	70	0.64
$\widehat{\mathcal{M}_R}$	Yes	Road	0.08	68	0.69
		TC	0.02	86	0.89
		All modes	0.05	78	0.80
$\widehat{\mathcal{M}}$	Yes	Road	0.10	67	0.67
		TC	0.03	86	0.86
		All modes	0.06	77	0.77

Table 2: Result of the map-matching approach on different sets of trajectories: $\widehat{\mathcal{M}}$ without prior knowledge on the transportation mode and M, $\mathcal{M}_R$, $\widehat{\mathcal{M}_R}$ and $\widehat{\mathcal{M}}$ with prior knowledge on the transportation mode.

the road network compared to that of the public transport one, which makes the map-matching problem harder in the former case.

4.1.4. Impact of sampling rate

To further evaluate the performance of our approach, we also quantify in the following the impact of spatio-temporal sparsity of NSD data that are fed to TRANSIT before map-matching. We do this by randomly sub-sampling the NSD of each user down to a fraction of the original mobile events in every sessions in $\widehat{\mathcal{M}}$; we then run our map-matching approach on the sparser trajectories, after performing or not trajectory augmentation with TRANSIT. Due to the stochastic nature of the sub-sampling, we averaged the F1 score over 10 trials (random samples selected with a given frequency) for each distinct sampling rate.

We can observe that the average value of the F1 score follows different trends in the case of the raw signaling trajectories and the TRANSIT-enhanced ones. Values of the F1-score increase for small growing values of the sampling frequency, both in the case of raw NSD and in the TRANSIT one. The difference of the performance between TRANSIT and raw NSD is constant and close to 0.15 for the smaller values. Differently, for higher sampling frequency, the F1 score remains approximately constant for raw NSD, regardless of the sampling frequency, whereas it keeps increasing for TRANSIT. Indeed, there is no reason why an increased number of raw NSD events would improve the intrinsic spatial uncertainty proper to such kind of data, as the map-mathing error is linked to the geographical sparsity of the antennas in the raw NSD case. In fact, the distance between NSD and the closest GPS position stays constant, and, consequently, the map-matching process cannot rely on additional spatial information to improve its performance when a higher sampling frequency is available. On the contrary, 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. As TRANSIT achieves to better

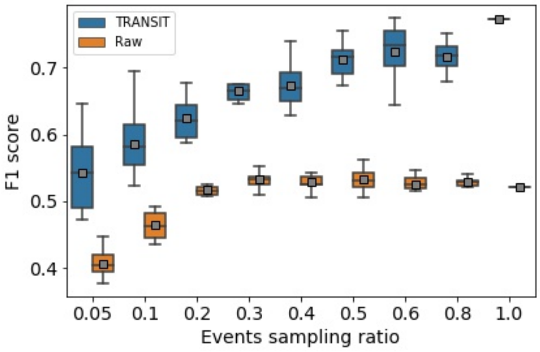


Figure 3: Performance of the map-matching with and without TRANSIT with varying events sampling rate.

reconstructing spatial information, the map-matching on TRANSIT-enhanced trajectory attains increased performance even with higher sampling frequency. This lets TRANSIT increase its F1 score up to 0.8 as the sampling rate grows.

4.2. Macroscopic validation

4.2.1. Dataset

The raw dataset used for macroscopic validation covers all the Orange subscribers' network signaling data for the city of Lyon, France. Data have been pseudonymized and treated with TRANSIT by the mobile phone network operator to extract (possibly enhanced) trips for a limited subset of Origin-Destination (OD) pairs. No personal information has been made available and trips from different users have been treated together, thus preventing any form of individual re-identification. Specifically, we only keep in our analysis TRANSIT trips whose origin (first position) and destination (last position) respectively fall into the origin and destination of the considered OD-pair. In particular, we focus our macroscopic analysis on three OD-pairs in Lyon: 4653 trips from the neighborhood of Confluence to the one of Cordeliers (the set of trips is denoted as C_1), 3222 trips from Part Dieu to Cailloux-sur-Fontaines (the set of trips is denoted as C_2) and 402 trips from Bellecour to Saint-Priest (the set of trips is denoted as C_3). These OD-pairs have been chosen to stress our map-matching approach in complex settings: multiple transportation modes (car, train and public transport), multimodal trips (public transport case). In addition, these OD-pairs are interesting to study because they connect zones of interest in the city of Lyon that contain high numbers of transport stations and hubs, touristic attractions, residential areas or commercial zones.

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 funded 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)[3], which provides an exemption for research, in particular for scientific and research purposes. Additionally, sensitive data have been processed by authorized personnel only at Orange facilities.

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

4.2.2. Results

To validate our map-matching approach at macroscopic level, we apply our approach on three different pairs of OD, namely: C_1, C_2 and C_3. It is worth highlighting that C_1, C_2 and C_3 are composed of raw signaling trips as well as TRANSIT enhanced trips related to the whole subscribers' base that have traversed these zones. As discussed in Sec. 4.1.3, the performance of the map-matching process is superior when considering prior knowledge on the transportation mode. Thus, we apply a simple, yet effective, speed-based heuristic to infer the transportation mode (either car, public transport or train) of each trip in C_1, C_2 and C_3. This heuristic is based on the assumption that public transport trips speed is lower than car trips speed which is lower than train speed. Ideas for improving such a basic inference approach are given in Sec. 5. Then, we apply our map-matching approach on C_1, C_2 and C_3. For evaluating the result of the map-matching, we compare the reconstructed paths obtained for each OD pair via our approach with reference paths, that we call ground-truth popular paths in the following. The latter have been obtained using a variety of route planners[4]. The former have been obtained by summing the number of occurrences of each edge of the transportation network from the map-matched trajectories obtained by means of our approach. Results are graphically presented in Fig.4, while the performance of our approach is assessed via visual inspection, as discussed in the following.

Concerning the OD pair C_1, related trips belong either to the road or to public transport sub-networks. For public transport trips, in Fig. 4a and Fig. 4b, we can observe that our approach correctly inferred the main two ground-truth popular paths as obtained via commonly-used route planners. The first one is a bus itinerary and the second one is a multi-modal public transport itinerary, consisting in a tramway segment followed by a subway one. Regarding car trips, in Fig. 4c and Fig. 4d, our approach completely retrieved ground-truth itineraries 1, 2, while retrieving only a portion of itinerary 3. In particular, our approach seems to wrongly infer a popular itinerary in the center of the figure. This itinerary is located between the retrieved itineraries 1 and 2. It is possible that, our speed-based heuristic approach failed at inferring the correct transportation mode for some trips. As a result, our approach map-matched trips to the wrong transportation network. In our case, it is likely that some public transport trips have been wrongly matched to the road network thus generating a fake road-based popular path, which is spatially close to a well-known public transport segment (included in itinerary 1 from Fig. 4b).

For the OD pair C_2, trips are associated either to the road or to the train sub-networks. For train trips, in Fig. 4e and Fig. 4f the only popular itinerary is correctly retrieved by our approach. Concerning car trips, in Fig. 4g and Fig. 4h, two main popular paths are present in our ground-truth data. The first one (itinerary 1) is correctly detected by our approach, whereas the second one (itinerary 2) is not. It is worth highlighting that ground-truth popular paths proposed by the route planners give some reasonable indications on popular paths, but may not necessarily be representative of actual route-choice preferences of users. This can explain some of the differences observed when inferring popular paths via our approach that relies on large-scale fresh data describing actual movements of large crowds of people, as observed through the lens of the mobile phone communication network.

Finally, for the C_3 OD pair, as reported in Fig. 4i, Fig. 4j, Fig. 4k and Fig. 4l, trips are associated either to the road or to the public transport sub-networks. For both car and public transport trips, a good match for popular paths can be observed: our approach retrieves almost all the popular paths detected via route planners. We also highlight that popular itinerary 1 for public transport is a multi-modal one.

All these aggregate results show a very promising application of our approach for inferring fine-

[4] https://www.google.fr/maps, https://www.viamichelin.fr/web/Itineraires

grained mobility information, *i.e.*, popular paths, by transportation mode at macroscopic level. Reported results also prove the capability of our solution to perform accurate map-matching even in the case of complex urban and multi-modal settings.

5. DISCUSSION AND CONCLUSION

In this paper, we performed an empirical study, based on real signaling traces collected in the city of Lyon, France, by a major telecommunication operator, aimed at investigating the potential of network signalling data to provide fine-grained spatio-temporal information to reconstruct users' mobility. We developed a HMM-based map-matching algorithm for mapping sparse and noisy cellular trajectories to the underlying multi-modal transportation network. The map-matching approach is coupled with TRANSIT, a network signaling data pre-processing framework developed by our team, able to enhance in both space and time the raw signaling trips.

To validate our approach, we have analyzed an original case study, related to French city of Lyon, by leveraging both real cellular traces collected by a major network operator and GPS data collected via a mobile phone application. This data has been leveraged to perform a microscopic validation, aimed at both fine-tuning the parameters of the HMM-based map-matching step and at showing the capability of our approach to accurately map-matching cellular trajectories on multiple transportation mode. We also demonstrated the importance of having prior rough transportation mode knowledge before applying the map-matching process to improve the performance of the latter.

Finally, using simple coarse transportation mode inference, we have demonstrated the possibly to retrieve popular paths by transportation mode for multiple OD-pairs. We underline the fact that, by relying on our approach and network signaling data, such a knowledge can be provided at very large scale (an entire country), with a temporal description (popular paths can be different at given moments of the day or during week-ends), and at much higher spatial resolution (covering also peripheral areas, or regions hardly observed via GPS-based data) than the one provided via simple traditional route planners, thus proving the utility of our approach and interest of the analyzed case study.

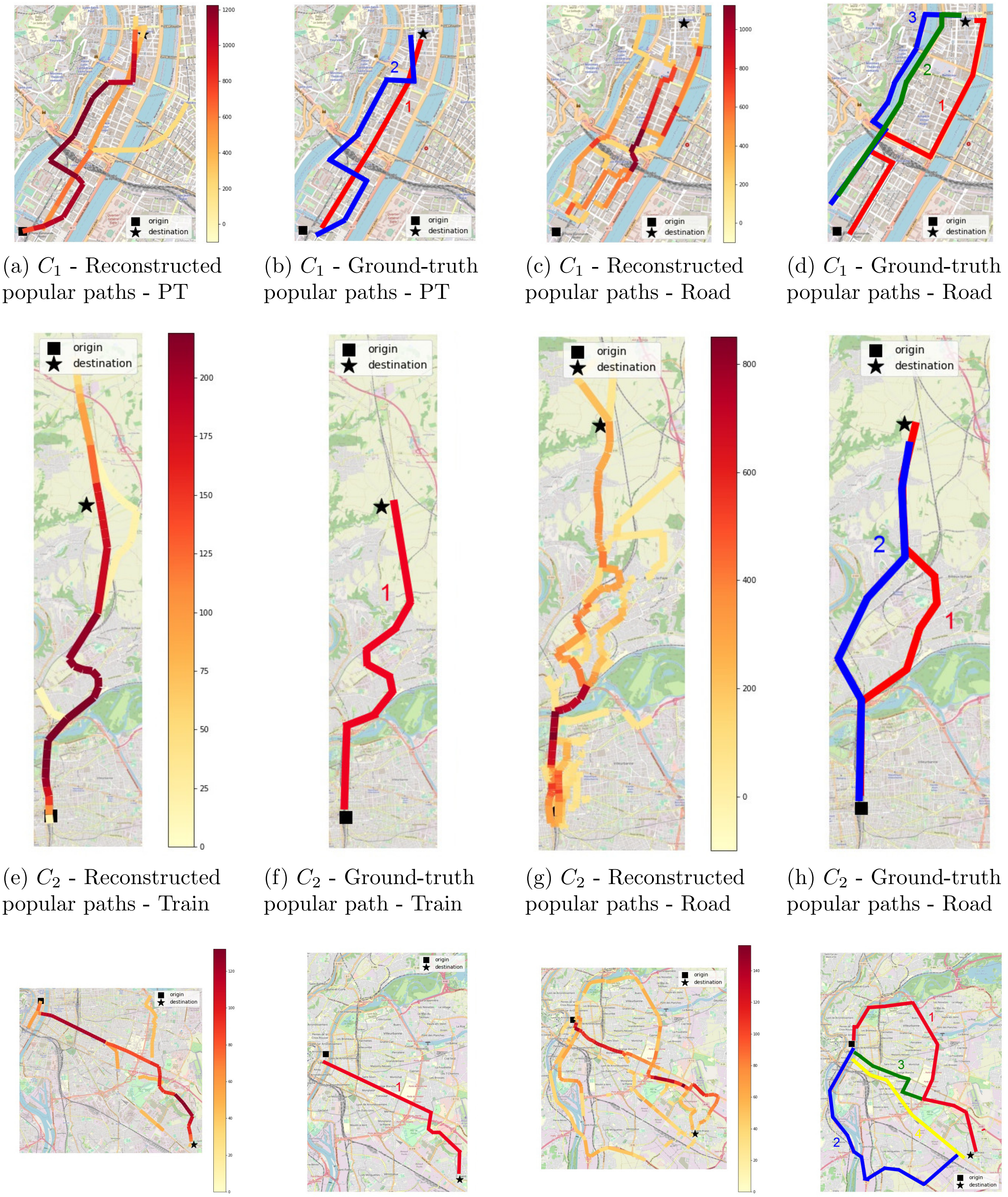
Future work includes improvement on the transportation mode inference technique. Indeed, instead of using only the speed, a lot of features could be used for the inference such as: the probability that the virterbi algorithm outputs, start time/duration of the trip and total static activity duration within the trip. Detailed analysis of the results deriving from a country-scale application of our solution could be explored as future directions.

ACKNOWLEDGMENTS

The authors would like to thank Orange for providing the mobile phone data used in this study. 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).

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study, conception and design: LB, AF, NEEF; analysis and interpretation of results: LB, AF, NEEF; LB was the lead writer of the manuscript. All authors reviewed the results and approved the final version of the manuscript.



(a) C_1 - Reconstructed popular paths - PT

(b) C_1 - Ground-truth popular paths - PT

(c) C_1 - Reconstructed popular paths - Road

(d) C_1 - Ground-truth popular paths - Road

(e) C_2 - Reconstructed popular paths - Train

(f) C_2 - Ground-truth popular path - Train

(g) C_2 - Reconstructed popular paths - Road

(h) C_2 - Ground-truth popular paths - Road

(i) C_3 - Reconstructed popular paths - PT

(j) C_3 - Ground-truth popular path - PT

(k) C_3 - Reconstructed popular paths - Road

(l) C_3 - Ground-truth popular paths - Road

Figure 4: Comparison between reconstructed popular paths reconstructed by our approach and ground-truth popular paths for 3 case studies: C_1, C_2 and C_2.

References

[1] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. Large-Scale Mobile Traffic Analysis: A Survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161, 21 2016.

[2] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 6 2008.

[3] Md. Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 3 2014.

[4] Angelo Furno, Marco Fiore, Razvan Stanica, Cezary Ziemlicki, and Zbigniew Smoreda. A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas. *IEEE Transactions on Mobile Computing*, 16(10):2682–2696, 10 2017.

[5] Angelo Furno, Marco Fiore, and Razvan Stanica. Joint spatial and temporal classification of mobile traffic demands. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9. IEEE, 5 2017.

[6] Qiang Xu, Alexandre Gerber, Zhuoqing Morley Mao, and Jeffrey 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*, pages 183–195, New York, New York, USA, 2011. ACM Press.

[7] Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1):30, 12 2019.

[8] Loïc Bonnetain, Angelo Furno, Nour-Eddin El Faouzi, Marco Fiore, Razvan Stanica, Zbigniew Smoreda, and Cezary Ziemlicki. TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. *Transportation Research Part C: Emerging Technologies*, 130:103257, 9 2021.

[9] An Luo, Shenghua Chen, and Bin Xv. Enhanced Map-Matching Algorithm with a Hidden Markov Model for Mobile Phone Positioning. *ISPRS International Journal of Geo-Information*, 6(11):327, 2017.

[10] Mohammed A. Quddus, Washington Y. Ochieng, and Robert B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

[11] Christopher E White, David Bernstein, and Alain L Kornhauser. Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8(1-6):91–108, 2 2000.

[12] Meng Yu. Improved positioning of land vehicle in its using digital map and other accessory information. *The Hong Kong Polytechnic University*, 2006.

[13] Dragan Obradovic, Henning Lenz, and Markus Schupfner. Fusion of Map and Sensor Data in a Modern Car Navigation System. *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 45(1-2):111–122, 11 2006.

[14] Mohammed A. Quddus, Robert B. Noland, and Washington Y. Ochieng. A High Accuracy Fuzzy Logic Based Map Matching Algorithm for Road Transport. *Journal of Intelligent Transportation Systems*, 10(3):103–115, 9 2006.

[15] Paul Newson and John Krumm. Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, page 336, New York, New York, USA, 2009. ACM Press.

[16] Gunnar Schulze, Christopher Horn, and Roman Kern. Map-Matching Cell Phone Trajectories of Low Spatial and Temporal Accuracy. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2015-Octob:2707–2714, 2015.

[17] George R. Jagadeesh and Thambipillai Srikanthan. Online Map-Matching of Noisy and Sparse Location Data with Hidden Markov and Route Choice Models. *IEEE Transactions on Intelligent Transportation Systems*, 18(9):2423–2434, 2017.

[18] Reham Mohamed, Heba Aly, and Moustafa Youssef. Accurate Real-Time Map Matching for Challenging Environments. *IEEE Transactions on Intelligent Transportation Systems*, 18(4):847–857, 4 2017.

[19] Arvind Thiagarajan, Lenin Ravindranath, Hari Balakrishnan, Samuel Madden, and Lewis Girod. Accurate, low-energy trajectory mapping for mobile devices, 2011.

[20] Essam Algizawy, Tetsuji Ogawa, and Ahmed El-Mahdy. Real-Time Large-Scale Map Matching Using Mobile Phone Data. *ACM Transactions on Knowledge Discovery from Data*, 11(4):1–38, 7 2017.

[21] Zhihao Shen, Wan Du, Xi Zhao, and Jianhua Zou. DMM: fast map matching for cellular data. *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 9 2020.

[22] Fereshteh Asgari, Alexis Sultan, Haoyi Xiong, Vincent Gauthier, and Mounîm A. El-Yacoubi. CT-Mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network. *Computer Communications*, 2016.

[23] Loïc Bonnetain, Angelo Furno, Jean Krug, and Nour-Eddin El 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):74–88, 7 2019.

[24] Geoff Boeing. OSMNX: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. *SSRN Electronic Journal*, 5 2016.

[25] Raymond H. Putra, T. Morimura, T. Osogami, and Noriaki Hirosue. Map matching with Hidden Markov Model on sampled road network. *undefined*, 2012.

[26] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 4 1967.

[27] Panagiota Katsikouli, Marco Fiore, Angelo Furno, and Razvan Stanica. Characterizing and Removing Oscillations in Mobile Phone Location Data. pages 1–9, 2019.