Can we map-match individual cellular network signaling trajectories in urban environments? A data-driven study.

Submission Date: August, 1st 2018

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

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

Jean Krug[1],
[1] Univ. Lyon, IFSTTAR, ENTPE, LICIT UMR_T9401, F-69675, Lyon, France
jean.krug@entpe.fr

Nour-Eddin El Faouzi[1,2]
[1] Univ. Lyon, IFSTTAR, ENTPE, LICIT UMR_T9401, F-69675, Lyon, France
[2] Queensland University of Technology, STRC, Gardens Point Campus, 2 George Street, G.P.O. Box 2434, Brisbane, Queensland 4001, Australia.
nour-eddin.elfaouzi@ifsttar.fr

Submitted to the 98th Annual Meeting of the Transportation Research Board for publication and presentation

---

Word Count:

Number of words: 6541
Number of tables: 2 (250 words each)
Total: 7041

Abstract

Mobile phone data collected by network operators can provide fundamental insights on individual and aggregate mobility of people, at unprecedented spatio-temporal scales. However, traditional call detail records (CDR) have fundamental issues due to low accuracy along both the spatial and the temporal dimensions, which limit their applicability for detailed studies on mobility, especially in urban scenarios. In this paper, we focus on a new generation of mobile phone passive data, individual cellular-network signaling data, characterized by higher spatio-temporal resolutions than traditional CDR data. We design a framework based on unsupervised Hidden Markov Model (HMM) for map-matching such kind of data on multi-modal transportation network, aimed at accurately inferring the complex multi-modal travel itineraries and popular paths people follow in their urban daily mobility. This information, especially if computed at large spatio-temporal scales, can represent a solid basis for studying actual and dynamic travel demand, to properly dimension multi-modal transport systems and even perform anomaly detection and adaptive network control. We evaluate our approach in a case-study based on real cellular traces collected by a major French operator in the city of Lyon, and propose a validation study at both microscopic and macroscopic levels. The results show that our approach can properly handle sparse and noisy cell phone trajectories in urban complex environments. Besides, the results are promising concerning popular paths detection and reconstruction of Origin-Destination matrices.

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

1. INTRODUCTION

In recent years, the widespread diffusion of mobile devices and the exploding consumption of Internet traffic via 3G and 4G technologies have made mobile phone data a crucial source of information in multiple domains. This is especially true in the field of transportation, as these data, usually including spatio-temporal information related to mobile phone users, can provide fundamental insights on people's mobility both at individual and aggregate scales.

For instance, Call Detail Records (CDR), also referred to as *mobile phone passive data*, have fed plenty of large-scale studies on human mobility, given the possibility to study urban mobility at unprecedented spatio-temporal scales [1]. Relevant work based on CDR data comprises i.e., modelling the general laws governing human movements [2], reconstructing Origin-Destination (O-D) matrices [3], understanding urban land use [4], [5] and inferring population density [6]. Mobile phone passive data are increasingly used also in operational contexts by mobility service providers and traffic authorities, in conjunction with - or even at the place of - more traditional data sources on mobility like census data, local travel surveys and logs from road-side units (e.g., loop detectors, LI-DAR or acoustic sensors, Bluetooth scanners, etc.). In fact, the latter suffer from very high deployment costs, extremely poor spatio-temporal resolutions, and are rarely informative in terms of individual mobility [7], [8].

However, despite significant benefits, CDR still have fundamental issues that need to be addressed due to low accuracy along both the spatial dimension (i.e., user location is only known at the cell sector or base station coverage levels) and the temporal one (i.e., events are recorded only when the user performs a voice call or texts a message), which limits their applicability for detailed studies on mobility, especially in urban settings.

In such scenarios, Global Positioning System (GPS) logs still represent the preferred choice, since they allow for obtaining data with higher degree of accuracy (i.e., meters) and temporal frequency (i.e., seconds). Such measures can be relatively easily analyzed and mapped to mobility patterns by relying on machine learning techniques [9], [10]. However, a huge overhead exists in collecting detailed GPS datasets at statistically relevant scales, being such data mostly retrieved on voluntary basis or via special agreements involving only a small sample of users or vehicles [7]. Given these limitations, extended variants of CDR (namely, network signaling logs and Internet session reports) are currently collected by network providers and investigated by the research community. Differently from CDR data, network signaling data 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 this paper, by building on related work from the field of GPS map-matching and CDR analysis, we focus on the possibility of inferring relatively accurate measures of both individual and aggregate mobility flows from cellular network signaling data.

In fact, in the context of next-generation intelligent transportation systems, inferring individual trips with a certain degree of accuracy, even in urban environment, will enable a better and more precise understanding of both microscopic and macroscopic mobility. Such knowledge is expected to be leveraged in many applications such as multi-modal transportation network analysis and optimization, traffic routing and adaptive control.

Map-matching of GPS traces has been widely studied in the literature [11] and state-of-the-art approaches can achieve high accuracy in the presence of large-sampling rate data (e.g., sampling rate of 1 Hz) [12]. Although, it is worth to remark that, in terms of penetration rate and energy consumption, mobile phone data represent much better candidates than GPS data to track users in a large-scale and suitable way [13]. In this paper, we deal with the sparsity (in time and space), the noise and large localization error associated to cell phone trajectories, that make the task of reconstructing trips very challenging [14]. A methodology based on Hidden Markov Model (HMM) is presented as the core of a map-matching algorithm engineered for cellular network signaling data. The algorithm infers the most likely path of a mobile phone user, given a sequence of network signaling events emitted by her/his smartphone during a trip.

The network modeling of the transportation graph and the cellular network, key elements of the proposed approach, are also presented. A study case using the HMM-based map-matching is performed with two different datasets from the city of Lyon (France).

In summary, the key contributions of this work are the following:

- The main solution for the challenging problem of mapping cellular trajectories to the multi-modal transportation network instead of only considering the road network.

- Unsupervised HMM-based map-matching approach allowing to infer trajectories on physical network from any sparse (spatially and temporally) cellular trajectory in dense urban context. This is made possible by a more fine-grained modeling of both transportation and cellular networks, compared to state-of-the art approaches [15].
- Dataset collection of real-world cellular trajectories related to a group of users in the Lyon metropolitan area. The dataset has been collected by Orange, the major French mobile network operator. Despite the sparsity of the available data, we analyze our approach in two case studies, for both macroscopic and microscopic mobility analysis.

The rest of the paper is organized as follows. In Sec. 2, we present related work. In Sec. 3, we formulate key definitions to define the map-matching problem. In Sec. 4, network modeling is presented. In Sec. 5, we discuss about the methodology of our HMM-based model. In Sec. 6, we test our approach on the considered dataset. We conclude in Sec. 7 by discussing the limits of our approach and future directions.

2. RELATED WORK

Map-matching is a basic 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 [16]. Several approaches exist in the literature to solve the problem of map-matching GPS traces to a transportation network. Quddus et al. [17] categorize map-matching approaches in four classes.

Geometric approaches only use the spatial geometry of the network: the most simple and popular map-matching algorithm consists in matching each position point to the closest node in the network [18].

Topological approaches use geometric information as well as topological information like the existence of connectivity between nodes of the network [19]. Very sensitive to noise and outliers, these approaches are not appropriate to solve map matching problem in presence of highly noisy and sparse data.

The third kind of approaches exploit *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, *advanced map-matching approaches* use more complex mathematical tools. A non exhaustive list of these methods includes, i.e., the Kalmam Filter, its Extended Kalman version [20], Dempster–Shafer theory [19], fuzzy logic models [21], or the application of Bayesian inference [22]. These state-of-the-art algorithms may achieve a quasi-perfect accuracy (location error lower than 10 meters) with high sampling rate GPS data. Newson et al. [12] first introduce HMM-based map-matching dealing with different GPS traces sampling rate. Their approach turned out to be much more robust and accurate with sparse and noisy trajectory compared to standard advanced map-matching approaches for high sampling rate data.

As a consequence of the growing availability of large-scale mobile phone data collected by network operators, map-matching cell phone trajectories is recently becoming a challenging task for researchers. Most of the approaches used with cellular trajectories are based on those traditionally designed for GPS map-matching. Sculze et al. [23] 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. [24], which reaches 75% of median accuracy.

Finally, HMM-based map-matching has become state-of-the-art approach for noisy and sparse location data and, a fortiori, mobile phone trajectory. Thiagarajan et al. [13] and, more recently, Algizawy et al. [25] developed supervised HMM models exhibiting good accuracy (75% for Thiagarajan et al. approach). However, such an approach needs to train the HMM model with a large amount of labeled cellular trajectories, which are very hard to obtain, especially when dealing with highly dynamic and irregular environments, such as urban areas. Instead, we prefer to focus on unsupervised models that do not require collecting and labeling any trajectory. Moreover, we state that additional information such as signal strength of observation are relatively hard to obtain from mobile network operators and therefore should not be required by the map-matching approach, as for example it's the case in [13]. Jagadeesh et al. [24] proposed an online map-matching algorithm combining HMM-based map-matching and route choice model.

Finally, it is worth to remark that most of the approaches match cellular trajectories only to road networks, without considering other transportation modes. Among the very few exceptions, it is necessary to mention the methodology recently proposed by Asgari et al. [15]. The authors have developed a framework, namely CT-Mapper, which has been designed with very similar objectives to those of our work. CT-Mapper is an unsupervised HMM model which aims at mapping sparse multi-modal cellular trajectories by using a multilayer transportation network. Yet, CT-Mapper has some limitations: the multilayer network allows for unrealistic paths (each subway station is connected to its closest road intersection for simplification matters). In addition, CT-Mapper requires already cleaned cellular trajectories. Dealing with noisy mobile phone data requires an advanced cleaning process which is not further specified in CT-Mapper. Finally, Asgari et al. [15] filtered trajectories, whose lengths are shorter than 5 kilometers and validated with trajectories with an the average length of 26.5 kilometers. Hence, CT-Mapper has been validated only in inter-urban mobility scenarios, thus seeming not to handle urban mobility. Our model aims at investigating and overcoming these limitations, using a more sophisticated approach especially concerning network modeling.

3. PROBLEM STATEMENT

The section presents the main definitions, and a formal conceptualization of the problem of map-matching sparse cell-phone trajectories to the underlying multi-modal transportation network. The definitions reported in the following are based on those used in strictly related recent work[15, 23]:

Definition 1 (Signaling event) *A signaling event is defined as any observation resulting of a communication activity between a cell phone and a base station. Each observation o is defined as a tuple $(\phi, \lambda, z, t) \in \mathbb{R}^3 \times \mathbb{N}$ consisting of the latitude ϕ, the longitude λ, the azimuth z and the timestamp t of the event.*

Definition 2 (Cell phone trajectory) *A cell phone trajectory $T = (o_1, \ldots, o_n)$ is defined as a sequence of network signaling events, ordered by their timestamps and related to the same mobile phone user. We consider the following as typical kinds of signaling events:* i) *communication events (i.e., calls and SMS);* ii) *handover events (i.e., cell changes during an established communication) and Location Area (LA) updates;* iii) *network attachment/detachment events;* iv) *data/internet connections.*

Definition 3 (Multi-Layer Transportation Graph) *A Multi-Layer Transportation Graph is defined as a directed graph $G = (V, E, L, \Psi)$ where E, V represent the vertices and the edges, respectively, and L is the set of possible layers related to different transportation modes. In our study, we focus on four layers only: road, bus, tramway and subway. Function Ψ indicates the layer associated to a given node, i.e., $\Psi : V \to L$ in G. Transportation Layer $G^l = (V^l, E^l)$ is a subset of G where $V^l = \{v | v \in V, \Psi(v) = l\}$ and $E^l = \{\langle v_i | v_j \rangle \in E, \Psi(v_i) = \Psi(v_j) = l\}$. Each node v_i is characterized by its latitude and longitude (i.e., the geographical position $v_i = \langle lat, lon \rangle_i$). CrossLayer edge set $E^{cl} \subset E$ defines the edges with pair of nodes not belonging to the same layer: $E^{cl} = \{\langle v_i | v_j \in E | \Psi(v_i) \neq \Psi(v_j)\}$.*

Definition 4 (Cellular Network) *The cellular network is defined as a set of cellular towers $C = (c_0, c_1 \ldots c_p)$, where each cell tower $c_p = (\phi, \lambda, z)$ is characterized by its latitude and longitude in the geographical coordinate system and the direction of the antennas called azimuth.*

Definition 5 (Path) *A path P between two nodes $v, w \in V$ is a sequence of edges $(e_1, \ldots, e_n) \in E^n$ such that $e_1 = (v, \cdot), e_n = (\cdot, w)$ and $\forall i \in [\![1, n-1]\!], \exists u \in V, e_i = (u, \cdot), e_{i+1} = (\cdot, u)$.*

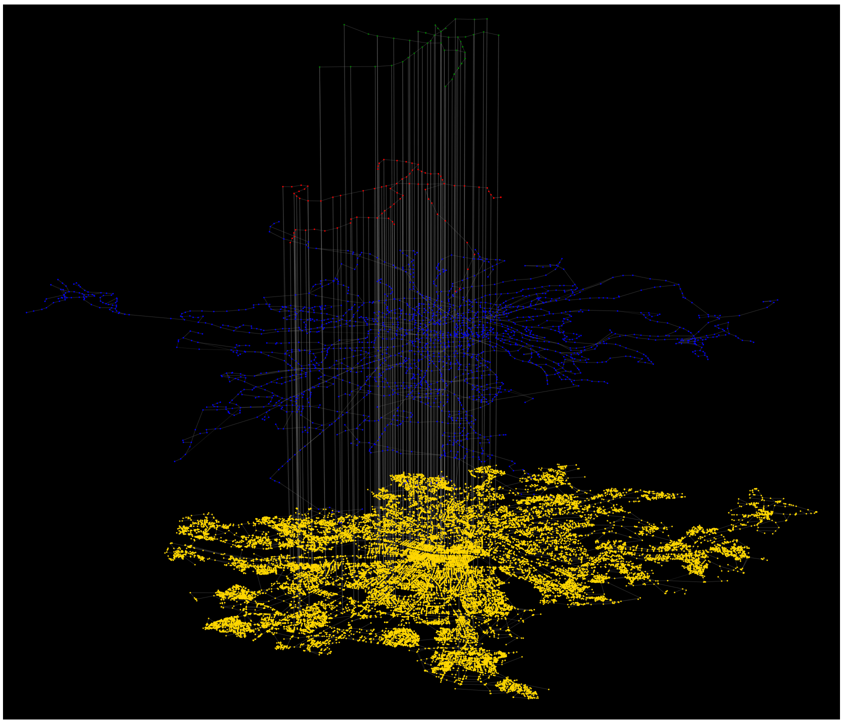
Finally, using the above definitions, the current work problem can be defined as follows: given a cellular trajectory T and the Multi-Layer graph G, the aim is to find the path P in G that leads to the observation o in T. This is obviously a map-matching problem.

4. NETWORK MODELING

4.1. Multi-Layer Transportation Graph

The transportation network studied in the paper is the multimodal transportation system of the city of Lyon, France. The network is designed as a multiplex network G composed of four graph layers representing four transportation modes: road, bus, tramway and subway. The whole Multi-Layer network and its different layers is shown in fig. 1a. The Python NetworkX library is used for multilayer modeling [26]. The graph and its different layers are built using multiple data sources and programming tools. The road network is generated via OSMnx [27], a Python library which creates NetworkX graphs from OSM data. Simplification of the road network topology derived from OSM is integrated as a facility in the library. The resulting road network corresponds to all drivable routes, representing the finest level of granularity that can be reached in road modeling.

Public transport layers have been generated using GTFS (Google Transit Feed Specification) data. We have performed some preprocessing steps (such as merging same public transport stops, which are in different directions) to obtain a reliable graph. Finally, cross-layers are added between layers to obtain the final multiplex graph structure. Between public transport layers, cross layers are defined as connections at transfer stops between public transport lines (this information is contained in the GTFS transfer file). In Asgari et al. [15], each subway station is connected to its closest road intersection for simplification matters. In light of a more realistic modeling, we prefer instead linking the road and public transport layers by using parking locations derived from Lyon OpenData [28]. The closest node of each parking location is thus connected to the closest public transport node.



(a) Visualization of Multi-Layer transportation network. Four transportation modes are considered: subway (green nodes, upper layer), tramway (red nodes, mid layer), buses (blue nodes, mid layer) and road (yellow, bottom layer). Cross-Layers (vertical grey edges) connect the different layers. See fig. 1b for statistics related to each of these layers.

Layer	$\|N\|$	$\|E\|$	$\langle k \rangle$	$\langle l \rangle$ (km)	Source
Multi-Layer	29012	63676	4.39	0.14	OSM/GTFS
Subway	46	80	3.47	0.78	GTFS
Tramway	86	173	4.03	0.60	GTFS
Bus	2023	4495	4.44	0.46	GTFS
Road	26853	58340	4.34	0.11	OSM

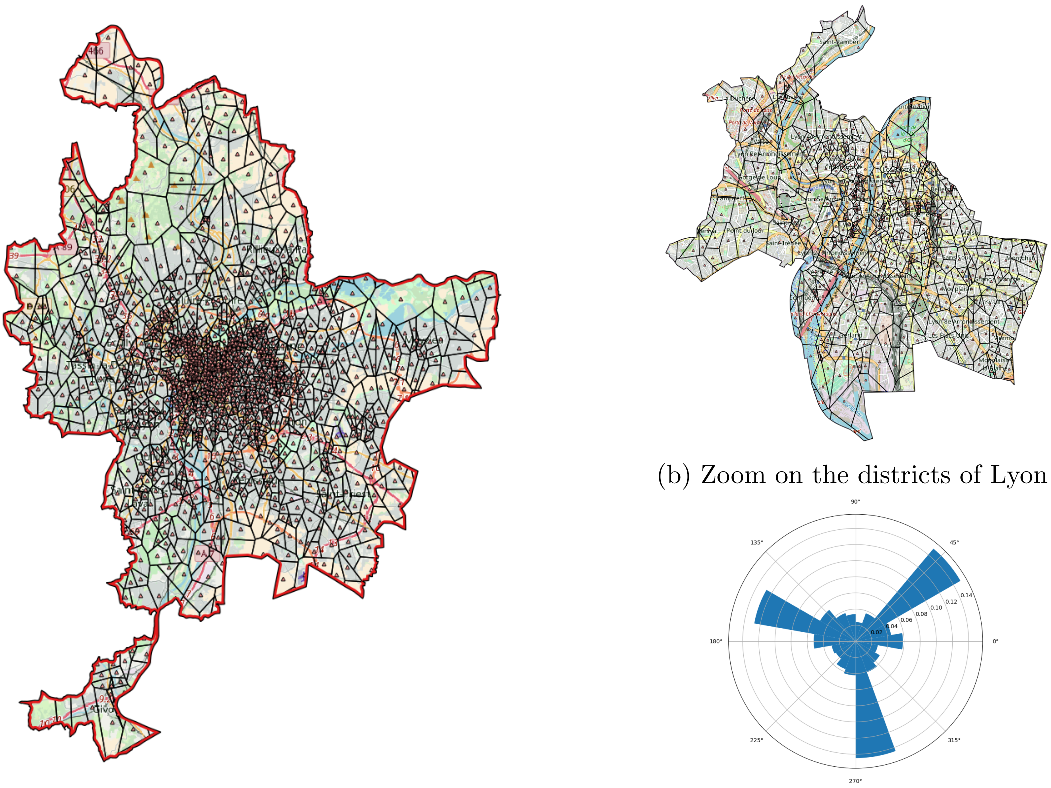
(b) Main characteristics of each transportation layer and Multi Layer network: number of nodes $|https://preview.overleaf.com/public/dgwdksqjgkzc/images/773f58be26b9877cc376c8b219f02635f183f679.pngN|$, number of edges $|E|$, average node degree $\langle k \rangle$ and average edge length in kilometer $\langle l \rangle$.

Figure 1: Lyon multimodal network: graphical representation (a) and main features (b)

4.2. Cellular Network

The cellular network considered in this study is composed of 13,306 antennas of the Orange mobile network operator, in the Metropole of Lyon region. Due to the overlapping of 2G, 3G and 3G+ cellular networks, antennas from different layers have the same characteristics (longitude, latitude and azimuth). After filtering duplicates, the result is a cellular network of 3,706 antennas. However, there are still antennas with the same spatial location (longitude and latitude) but different azimuths. In order to improve the modeling of the cellular network, we propose a method joining traditional Voronoi tessellation with the azimuth information to remove spatial overlapping. Specifically, each antenna is translated of an infinitesimal distance in the direction of its azimuth.

After applying Voronoi tessellation to the azimuth-corrected set of antennas, we consider the new location of the antenna as the barycenter of the polygon representing the Voronoi cell fig. 2a. Compared to the simple Voronoi tessellation, as applied in [15], our coverage model is about three times more segmented by taking into account the azimuth of the antennas (i.e., the area covered by each antenna is on average three times lower in our approach than in a traditional Voronoi tessellation). fig. 2c shows the azimuth distribution of the set of antennas from the cellular network. Three main directions can be observed.



(a) Cellular network in Metropole of Lyon

(b) Zoom on the districts of Lyon

(c) Azimuth of antennas in the cellular network

Figure 2: Cellular network

5. METHODOLOGY

The following section reports on the main methodological background characterizing our solution to perform map-matching of cellular network trajectories, as issued from individual anonymized network signaling mobile phone passive data.

5.1. Hidden Markov Model

A Hidden Markov Model can be defined by a five-fold $\langle V, C, \pi, A, B \rangle$, where:

- $V = \{v_1, \ldots, v_N\}$ is a set of states.

- $C = \{c_1, \ldots, c_M\}$ is a finite state of possible observations (also called emissions)
- π is the probability distribution of the initial state, given that π is a probability distribution: $\sum_{i=1}^{N} \pi(i) = 1$
- A is a set of transition probability . The probability to transit from hidden state v_i to hidden state v_j is denoted as $\{a(v_i; v_j)\}$. Besides, $\forall v_i \in V$, $\sum_{v_j \in V} a(v_i, v_j) = 1$
- B is a set of emission probability . The probability to emit observation o_j from hidden state v_j is denoted as $\{b(v_i; o_j)\}$. Besides, $\forall v_i \in V$, $\sum_{o_j \in C} b(v_i, o_j) = 1$

Our map-matching problem can be modeled with a Hidden Markov Model: hidden states are modeled as the set of vertices (nodes) V from the Multi-Layer Transportation Graph. Observations are modeled as the set of antennas C from Cellular Network. Hidden Markov Model allows to solve the following problem: given a sequence of observations (sequence of antennas on a cellular trajectory), the model finds the most likely sequence of hidden states (sequence of nodes on the transportation network).

5.2. HMM parameters

5.2.1. Initial Probability

As the definition of the initial probability, all the nodes in the transportation network are equally assigned with a probability of $1/N$ with N representing the number total of nodes in the transportation network:

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

5.2.2. Transition Probability

The transition probability corresponds to the probability that a mobile phone user moves, on the underlying transportation network from hidden state v_i at time $t-1$ to hidden state v_j at time t. Various transition probabilities have been proposed in the literature. For instance, as in the definition by Luo et al. [16], transition probability only depends on the network connectivity. The one used by Thiagarajan et al. [13] depends instead on the distance between transportation nodes. However, all of these approaches use road transportation network to define transition probability. Thus, these definitions require to be adapted to the case of a multilayer network, in order to take into account the attributes of each layer. Hence, we choose the definition proposed by Asgari et al. [15], i.e., the transition probability depends on the average speed over an an edge and the edge length.
Weights which depend on the average speed over an edge are defined as follow:

$$W_{ij} = \begin{cases} w_{ij} & \text{if } v_i \text{ and } v_j \text{ are adjacent in } G \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

value of w_{ij}	Condition
1/80	$\Psi(v_i) = \Psi(v_j) = subway$
1/25	$\Psi(v_i) = \Psi(v_j) = tramway$
1/15	$\Psi(v_i) = \Psi(v_j) = bus$
1/50	$\Psi(v_i) = \Psi(v_j) = road$
1/10	$\Psi(v_i) \neq \Psi(v_j)$

Table 1: Edge classification and weights for multilayer transportation network G

Finally, the transition probability is defined as the inverse of the shortest path cost between two nodes v_i

and v_j:

$$a(v_i, v_j) = \left(\sum_{\forall (mn) \in SP_{v_i v_j}} w_{mn} \cdot d(v_m, v_n) \right)^{-1} \tag{3}$$

where (mn) is an edge between v_m and v_n belonging to $SP_{v_i v_j}$ the shortest path between two nodes v_i and v_j in graph G. The shortest path cost of $SP_{v_i v_j}$ is the sum of distances over each edge (mn) belonging to $SP_{v_i v_j}$ weighted by w_{mn}. $d(v_m, v_n)$ is the geodesic distance between each two nodes v_m and v_n.

5.2.3. Emission Probability

The emission probability corresponds to the probability that an individual user is in the hidden state v_i at time t given that an observation (e.g communication event at an antenna) o_j is observed on the cellular network at time t. In the literature, in the mobile phone data context, various emission probability have been proposed. Luo et al. [16] define a score inversely proportional to the distance between the hidden state and the observation. Jagadeesh et al. [24] prefer to use a Gaussian distribution with zero mean and an empirically estimated standard deviation of the measurement error between hidden states and observations. Similarly to Asgari et al. [15], since detailed information regarding the underlying cellular network is unavailable, we use Voronoi tessellation to model the area covered by each antenna. Finally, the emission probability is defined as a decreasing function of the distance between the antenna location and the hidden state:

$$b(o_t, v_j) = \begin{cases} 1 & \text{if: } d_{tj} < r_{max} \\ \left(\frac{r_{max}}{d_{tj}} \right)^{\beta} & \text{if: } r_{max} < d_{tj} < \tau \cdot r_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where d_{tj} is the euclidean distance between o_t and intersection v_j, and $\beta = \frac{\ln(10)}{\ln(\tau)}$ is the decreasing factor which has been defined to obtain an emission score ten times lower for $d_{tj} = \tau \cdot r_{max}$ and $\tau \cdot r_{max}$ is a threshold corresponding to the maximum distance at which a cell phone can be covered by a given cellular antenna. Considering the fact that the communication power is generally proportional to inverse square of the distances [25], coefficient $\beta = 2$ that leads to $\tau = 3$ is chosen.

5.3. Preprocessing

This preprocessing step aims at reducing the noise in cellular phone trajectory. This a key step to improve map-matching accuracy process. The cleaning algorithm of our approach follows these three sequential steps:

- apply a recursive look ahead filter [29]. This filter is based on the mobile phone travel speed on the cellular network. If the speed is higher than a given parameter, the outlier record is removed. In the algorithm, this speed is set at 500 km/h.
- through investigation into the data, we have decided to aggregate records with a given threshold of two minutes to reduce the oscillation effect (also called ping-pong effect) on the cellular trajectory. Moreover, this value of two minutes is lower enough to avoid loosing information on the cellular trajectory. The antennas detected within the threshold are replaced by a single antenna, i.e., the closest one to the coordinates of the barycenter of the diverse antennas.
- remove consecutive records detected at the same antennas. We consider in this case that the user is static, thus no information is lost by simply removing the record.

5.4. Map-Matching algorithm

After applying the cleaning algorithm described above, map-matching can be used on cleaned cellular trajectory. Our approach is a two-steps map-matching algorithm 1. First, an optimized Viterbi algorithm [30] is run. The inputs of the Viterbi process are the following: the transportation network modeled as a multiplex network G, the possibles states (set of the nodes of G, the emissions (set of antennas from the cellular network), the HMM parameters defined and the cellular trajectory. By calculating all possible paths given the cellular trajectory, the Viterbi process output is the 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 [25]. The standard Viterbi algorithm applied in a case with 6,110 states (less complex network than the one used in the study), 3,706 antennas results in around two hours to the reconstruction of a set of 2,300 observation sequences. The algorithm runs on a server machine equipped with an Intel Xeon E5 2,640 2.4 GHz multi-core machine, equipped with 56 virtual cores and 128 GB of DDR4 RAM. Using the sparseness of cellular trajectory, the main optimization processes for real time application used by Algizawy et al. [25] are applied. The major process consists in "eliminating all multiplications by zero and reduces the search space by keeping only with emittable states from each state observable". The execution time of the optimized Viterbi algorithm is 3 seconds instead of 2 hours to reconstruct a set of 2,300 observation sequences.

Due to extremely long execution time on a traditional PC hardware, we have used the server machine for our computation. The server has been used for running the map-matching algorithm using a transportation network of 29,012 nodes. In order to reduce time execution, multiprocessing Python libraries such as joblib have been used.

Finally, after inferring the most likely states sequence using the optimized Viterbi implementation presented above, the final trajectory is inferred by applying a traditional shortest path (Dijkstra) detection algorithm on the underlying transportation graph between two consecutive nodes.

Procedure 1 Map-Matching algorithm

Input:

Graph, G

States, $V = \{V_0, \dots, V_{N-1}\}$

Emissions, $C = \{c_0, \dots, c_{N-1}\}$

Cell phone trajectory, $T = (o_0, o_1, o_2, ..., o_l)$ where $o_i \in C$ and l is the length of the sequence

Initial probabilities, $\pi_i \in V$

Transition probabilities, a_{ij} such that $i, j \in V$ and $0 < i, j < n - 1$

Emission probabilities, b_{ij} such that $i \in C$ and $j \in V$

Output:

Maximum probability, OutputProb

Edge sequence, $OutputPath =< V_{o_0}, V_{o_1}, \dots, V_{o_{l-1}} >$

First step: Optimized Viterbi Algorithm

$V \leftarrow \{\}$
$Path \leftarrow \{\}$
for all y in V **do**
  $V[0][y] = \pi_y \cdot b_{o_i,y}$
  $Path[y] \leftarrow y$
end for
for $t \leftarrow 1$ to $l - 1$ **do**
  for all y in $V | b_{o_t,y} \neq 0$ **do**
    $V[0][y] = \pi_y \cdot b_{o_0,y}$
    $(prob, state) = \max_{y_0 \in V | a_{y_0,y} \neq 0} (V[t-1][y_0] \cdot a_{y_0,y} \cdot b_{o_0,y}, y_0)$
    $V[t][y] \leftarrow prob$
    $NewPath[y] \leftarrow Path[state] + y$
  end for
end for

Second step:

$(prob, state) = \max_{y \in V} (V[l-1][y], y)$
$OutputProb \leftarrow prob$
$OutputPath \leftarrow Path[State]$
$FinalPath \leftarrow OutputPath[0]$
for all k in $OutputPath \setminus \{OutputPath[0]\}$ **do**
  $FromNode \leftarrow OutputPath[k-1]$
  $ToNode \leftarrow OutputPath[k]$
  $IntPath \leftarrow ShortestPath(FromNode,\ ToNode,\ G)$
  $FinalPath \leftarrow FinalPath + IntPath$
end for

6. STUDY-CASE: LYON

6.1. Datasets

In order to test our approach, two datasets related to Lyon metropolitan area are used:

- Anonymized individual mobile phone data provided by Orange, the major French telecom operator covering the week from 9/9/2015 to 15/9/2015. The latter are records of all users who visit, in a same day, at least one base station in two areas of Lyon, i.e., the Part Dieu (PD) and Sainte-Foy (SF) areas. It is worth mentioning that the identifiers of such users are not the same across different days for privacy issues. Only timestamps, user id, antennas id information are provided. This dataset is used for the macroscopic validation of our approach 6.2.2.

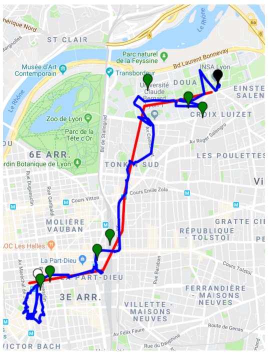
- Both GPS traces and mobile phone records are collected for a group of users in Lyon metropolitan area. Mobile phone data have the same characteristics as described above. This dataset is used for the microscopic validation of our approach, GPS traces being used as ground truth 6.2.1.

6.2. Result

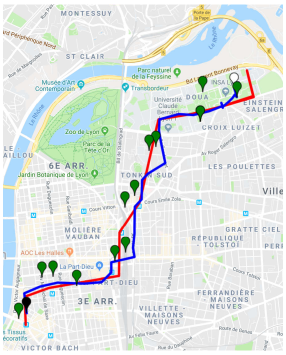
6.2.1. Microscopic validation

In order to validate our model for microscopic user mobility, we applied our HMM-based map-matching on the cellular trajectory and compare the inferred trajectory with GPS traces, considered as ground truth. In fig. 3, six results from the proposed approach are shown. In fig. 3a and fig. 3b, thanks to a fine-grained multimodal network, our algorithm shows a good accuracy, despite sparse trajectory, in an urban context.

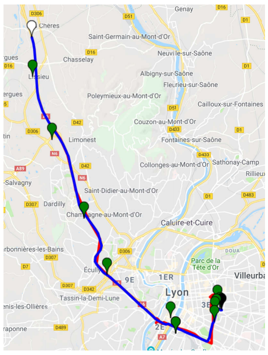
Moreover, in fig. 3a our algorithm is able to properly infer a trip on the public transportation network (the transportation mode used is the tramway). The algorithm is particularly effective in accurately map-matching cellular trajectories along major roads in inter-urban contexts, as clearly shown in fig. 3c. Indeed, the complexity of the map-matching problem is reduced when a user is moving in a non-urban environment, compared to an urban context. In such situations, our model is highly accurate and able to fairly reconstruct such trajectories. Finally, in fig. 3f, an example of reduced accuracy of our map-matching solution is shown. In case of multiple events in a short spatial range, our approach considers the user as mobile and attempts to infer a path whereas she/he is static. This explains why some loops appear on inferred cellular trajectory.



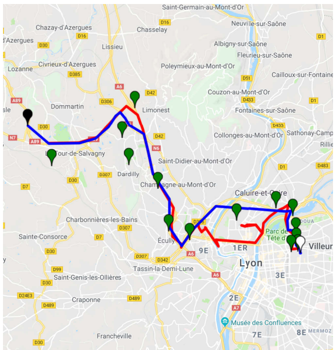
(a) Cellular trajectory: urban trip



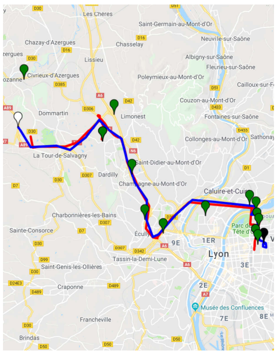
(b) Cellular trajectory: urban trip



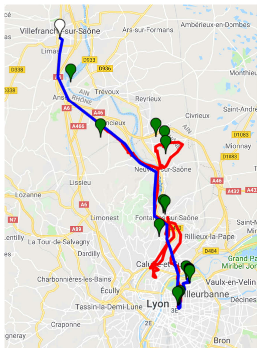
(c) Cellular trajectory: inter-urban trip



(d) Cellular trajectory: home to work trip



(e) Cellular trajectory: work to home trip

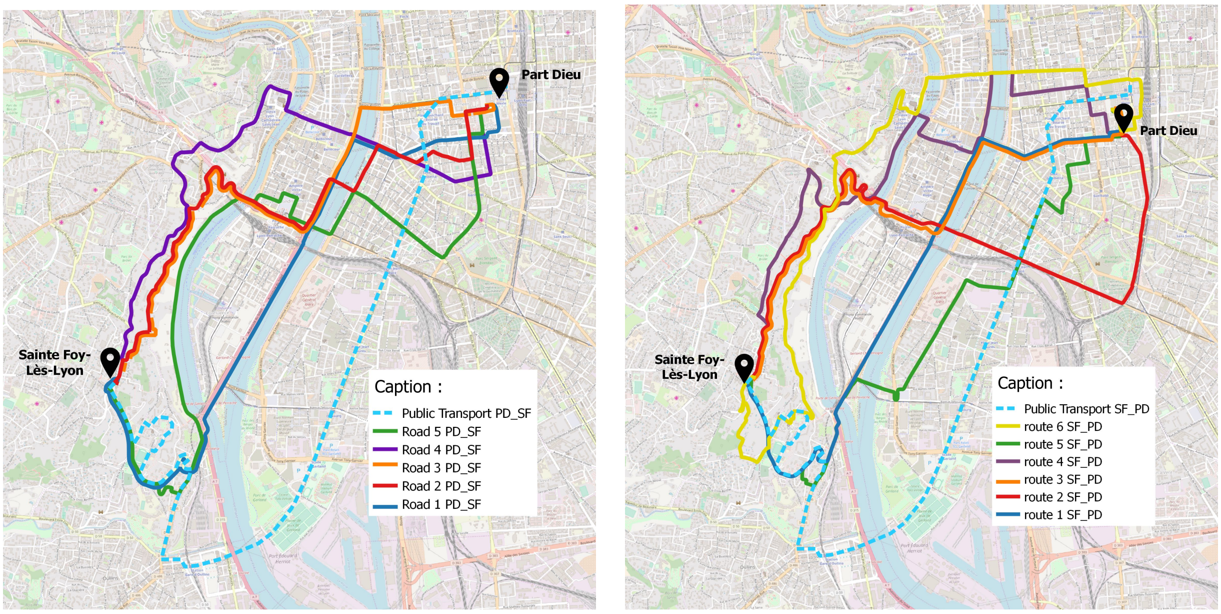


(f) Cellular trajectory: multiple events issue

Figure 3: Set of cellular trajectory. The blue line represents the GPS trace (Ground Truth), the red line is the trajectory inferred by our approach (Output of the map-matching algorithm), the green markers correspond to the cellular trajectory (input of the map-matching algorithm), the white marker represents the beginning of the trip and the black one, the end.

6.2.2. Macroscopic validation

To validate our approach according to a more aggregate and larger-scale perspective, we propose a macroscopic evaluation, which aims at answering the following question: is our algorithm able to properly infer the distribution of flows over the most-traversed paths between the two considered areas of Part-Dieu and Sainte-Foy? In order to determine, in a fairly realistic way, a sort of ground truth describing the common paths between Sainte-Foy (SF) and Part-Dieu (PD), we used a combination of several tools. First, we used a A* shortest-path (SP) algorithm, based on the work of [31]. This SP algorithm uses a heuristics-directed search and it includes link penalties for multi-path search. It also incorporates a link penalty depending on a hierarchical description of the network. This method provided us with a set of routes efficient for cars only. We completed these results with the Google Map itineraries calculation, in order to confirm the results produced by the SP algorithm and to add supplementary routes for public transportation and bicycles. For public transportation, we also relied on the website of the SYTRAL, the public transportation authority in Lyon. At last, our choices were confirmed by our knowledge of the city. Especially, for the SF to PD direction, we added a supplementary route which seemed to be reliable, even if it was not proposed neither by our A* nor by Google Map algorithm.



(a) Popular paths from Part-Dieu to Sainte-Foy

(b) Popular paths from Sainte-Foy to Part-Dieu

Figure 4: Popular paths from Sainte-Foy to Part-Dieu

Assigning users to the different alternative paths is not straightforward. In our case, we derived the assignment coefficients results from a length-based C-logit approach following the work of [32]. The C-logit model solves a Stochastic User Equilibrium problem by considering both the cost of each alternatives and the commonality factor between alternatives. The cost is the mean travel time, provided as a static data. A numerical parameter β, presented in the above-mentioned article, was set to 70. The θ parameter on which the logit formula relies was set to $\theta = 0.009$ (see [32]). To determine θ, we ran a static traffic simulation on the city of Lyon-Villeurbanne in which we tried to minimize the difference between observed and modeled flow, while calibrating θ. Observations were taken from loop detectors and furnished by the city authorities.

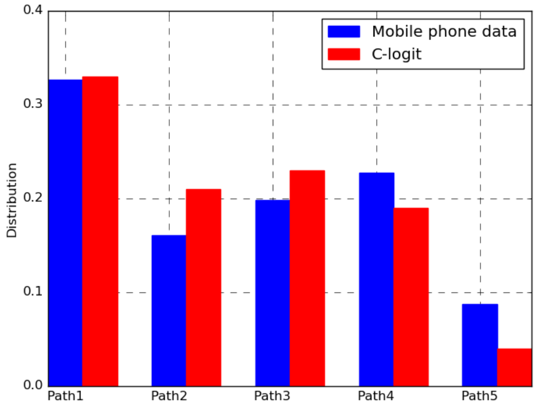


Figure 5: Macroscopic flow between Part-Dieu and Sainte-Foy

In fig. 5, we report on the comparison between macroscopic flows, as inferred from static traffic simulations (in red) and the aggregated results retrieved by using our cellular-trajectory-based map-matching approach (in blue). It is worth highlighting that, given the biases and incertitude present in both approaches to compute path distribution, expecting a perfect match between the two approaches is rather unrealistic. However, we believe this comparison can provide qualitative and global insights on the capacity of our solution to properly match trajectories on the multi-modal transport network. Also, it complements the previously described microscopic validation, which has already proven a good hit-rate at an individual scale.

As a first interesting result, our approach doesn't lead to completely unrealistic and unexpected traffic flows, like for instance all cellular trajectories matching with only one or two expected popular paths from simulations. Besides, the two approaches lead to consistent results in terms of popular/unpopular paths: *Path 1* is the most used in both cases, while *Path 5* has the lowest score in the two approaches as well.

7. CONCLUSION AND DISCUSSION

Cellular-network signaling data have the great potential to provide fine-grained spatio-temporal information to reconstruct users' mobility at both microscopic and macroscopic scales. In this paper, we performed an empirical study, based on real cellular traces collected in the city of Lyon, France, by a major telecommunication operator, aimed at investigating such potential. We developed a HMM-based map-matching algorithm for mapping sparse and noisy cellular trajectories to the underlying transportation network.

As a practical basis for our approach, we developed automatic tools to build a large multi-modal transportation network.

Taking into account the azimuth of antennas allows to increase cellular network segmentation. Network modeling at a fine-level of granularity allows for properly applying map-matching in urban complex environment.

By providing a formal definition of the HMM parameters, our methodology follows three main steps: the cleaning process, an optimized implementation of the Viterbi algorithm, and the determination of the shortest path on the sequence of nodes returned by Viterbi algorithm.

To validate our approach, we have analyzed an original case study, related to the 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 proving the accurate map-matching capability of our approach, even in a complex urban context. Moreover, we have demonstrated the possibly to retrieve popular paths between two areas by comparing the spatial distribution of flows as computed by both our approach and simulations.

Future directions should consider improvement with dynamic HMM parameters, in order to build a transition matrix depending on actual traffic conditions. In addition, some limitations of our algorithm have been shown in relation to oscillations (or ping-pong effect) in the user's communication activity. Therefore, a better understanding of this recurrent phenomenon is required. The latter should allow to create an advanced filtering approach to remove this oscillation effect from cellular trajectories and further improve the map-matching accuracy.

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] Rex W Douglass, David A Meyer, Megha Ram, David Rideout, and Dongjin Song. High resolution population estimates from telecommunications data. *EPJ Data Science*, 4(1):4, 12 2015.

[7] Hugo Barbosa-Filho, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. Human Mobility: Models and Applications. 9 2017.

[8] John R. B. Palmer, Thomas J. Espenshade, Frederic Bartumeus, Chang Y. Chung, Necati Ercan Ozgencil, and Kathleen Li. New Approaches to Human Mobility: Using Mobile Phones for Demographic Research. *Demography*, 50(3):1105–1128, 6 2013.

[9] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the Levy-Walk Nature of Human Mobility. *IEEE/ACM Transactions on Networking*, 19(3):630–643, 6 2011.

[10] Yu Zheng, Xing Xie, and Wei-Ying Ma. Understanding Mobility Based on GPS Data, 9 2008.

[11] Y U Zheng. Trajectory Data Mining: An Overview. *ACM Trans. On Intelligent Systems and Technology*, 6(3), 2015.

[12] 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.

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

[14] Wei Wu, Yue Wang, Joao Bartolo Gomes, Dang The Anh, Spiros Antonatos, Mingqiang Xue, Peng Yang, Ghim Eng Yap, Xiaoli Li, Shonali Krishnaswamy, James Decraene, and Amy Shi Nash. Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling. In *2014 IEEE 15th International Conference on Mobile Data Management*, pages 321–328. IEEE, 7 2014.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Meng Yu. Improved positioning of land vehicle in its using digital map and other accessory information. 2006.

[20] 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.

[21] 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.

[22] Jong-Sun Pyo, Dong-Ho Shin, and Tae-Kyung Sung. Development of a map matching method using the multiple hypothesis technique. In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585)*, pages 23–27. IEEE.

[23] 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.

[24] 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.

[25] 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.

[26] Aric A Hagberg hagberg, lanlgov Los, Daniel A Schult, and Pieter J Swart swart. Exploring Network Structure, Dynamics, and Function using NetworkX. Technical report, 2008.

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

[28] Données métropolitaines du Grand Lyon.

[29] Spécialité : Informatique et Réseaux. Technical report.

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

[31] Peter Hart, Nils Nilsson, and Bertram Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[32] Zhong Zhou, Anthony Chen, and Shlomo Bekhor. C-logit stochastic user equilibrium model: formulations and solution algorithm. *Transportmetrica*, 8(1):17–41, 1 2012.