Unlocking the potential of mobile phone data for large scale urban mobility estimation

Loic Bonnetain

▶ **To cite this version:**

Loic Bonnetain. Unlocking the potential of mobile phone data for large scale urban mobility estimation. Data Analysis, Statistics and Probability [physics.data-an]. Université de Lyon, 2022. English. NNT : 2022LYSET003 . tel-03920673

HAL Id: tel-03920673

https://theses.hal.science/tel-03920673v1

Submitted on 3 Jan 2023

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

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

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opéré au sein de
L'École Nationale des Travaux Publics de l'Etat (ENTPE)

École Doctorale MEGA ED162
Mécanique – Energétique – Génie Civil – Acoustique

Spécialités : **Génie Civil**

Soutenue publiquement le 01/02/2022, par :
Loïc Bonnetain

Potentialité des Données de Téléphonie Mobile pour l'Estimation de la Mobilité Urbaine à Large Echelle

---

Devant le jury composé de :

Christine SOLNON, Professeur, INSA Lyon ..... Présidente de jury

Francesco VITI, Assistant Professeur, Université du Luxembourg ..... Rapporteur
Edward CHUNG, Professeur, Université de Hong-Kong ..... Rapporteur
Marco FIORE, Assistant professeur, IMDEA ..... Examinateur
Zbigniew SMOREDA, Chercheur, Orange Labs ..... Examinateur
Nour-Eddin EL FAOUZI, Directeur de recherche, UGE-ENTPE ..... Directeur de thèse
Angelo FURNO, Chargé de recherche, UGE-ENTPE ..... Directeur de thèse

DISSERTATION OF THE UNIVERSITY OF LYON
at
Environmental & Urban Engineering School (ENTPE)

Doctoral school MEGA ED162
Mechanics – Energy – Civil engineering - Acoustics

PhD specialities: **Civil Engineering**

Defended publicly on 01/02/2022, by:
Loïc Bonnetain

Unlocking the Potential of Mobile Phone Data for Large Scale Urban Mobility Estimation

---

Composition of the doctoral committee:

Christine SOLNON, Professor, INSA Lyon ..... Jury president

Francesco VITI, Associate Professor, University of Luxembourg ..... Reviewer
Edward CHUNG, Professor, University of Hong-Kong ..... Reviewer
Marco FIORE, Associate professor, IMDEA ..... Examinator
Zbigniew SMOREDA, Researcher, Orange Labs ..... Examinator
Nour-Eddin EL FAOUZI, Research Director, UGE-ENTPE ..... Thesis director
Angelo FURNO, Researcher, UGE-ENTPE ..... Thesis director

Abstract

Nowadays, cities deal with multiple issues such as rapid urbanization, pollution or congestion. A fundamental step for facing these raising challenges consists in accurately characterizing how people move within the city. Such a fundamental knowledge can be leveraged to improve and optimize the transportation system. Until recent times, transportation authorities have mainly relied on mobility surveying to capture information on human mobility. However, this method suffers from multiple drawbacks. Surveys are expensive to run, get quickly outdated, are unavoidably based on relatively small samples of the population and cannot capture fine-grained mobility dynamics in space and time. In the last decades, with the digitization of the society, the emergence of new data sources such as smart cards, the Global Positioning System (GPS), location-based social media, or mobile phone records have attracted researchers' and practitioners' attention for mobility estimation. The research community has started to investigate the use of these data in the domain of mobility and transportation, where they could allow analyses at unprecedented scales compared to traditional surveys. Among the emerging sources, mobile phone network data are currently one of the most promising human mobility data sources. Such kind of data presents a unique combination of desirable properties: (i) they offer unprecedented penetration, as they are available for the whole subscribers base of a network provider; (ii) they are recorded continuously over long time periods, thus allowing fine-grained longitudinal studies over months or years; and, (iii) they are passively collected and maintained in curated databases for billing purposes, which makes them a very cost-efficient source of data for secondary use and analysis. However, only limited knowledge exists on the use of large-scale mobile network data to analyze human mobility in urban environments. This dissertation aims at filling this gap. In this thesis, we show that despite some limitations that are typical for the mobile phone data (the sparsity in time and space, the noise and the large localization error), such data contain rich spatiotemporal information that can be used for various purposes in the transportation domain such as OD matrix construction, travel demands patterns, land use analysis or popular paths inference. In addition, we present a new framework TRANSIT (*TRAjectory inference from Network SIgnaling daTa*) processing large scale mobile network data. On the one hand, the framework is able to reconstruct mobility information in especially tell apart movement intervals from stationary activity periods for each mobile device. This allows to reconstruct travel demand from mobile network data at an unprecedented spatiotemporal scale. On the other hand, TRANSIT is capable of inferring fine-grained human mobility trajectories during the associated movement intervals. Thereby, the frameworks aims at overcoming the above mentioned limitations. TRANSIT exploits the recurrent patterns of human mobility, i.e., the same individual typically performs many trips between two same given locations over time, generally following very similar paths. This creates redundancy in the mobility information that TRANSIT uses to increase the spatiotemporal accuracy of the trajectories. To totally unlock the potential of the mobile phone data, we also developed map-matching approaches that can be applied on the top of TRANSIT. The latter allow to retrieve the path a traveller follows on the multimodal transportation system from the reconstructed trajectories. Relying on the result of TRANSIT, we study the problem of mobility patterns discovery along multiple dimensions at aggregated scale that we solved using a data-agnostic method based on tensor decomposition. Thus, we propose a new set of applications such as the OD-matrix anonymization, the modeling of the COVID-19 propagation or the regional-scale travel patterns analysis that are

made possible by the use of mobile network data. As a conclusion, the results of this thesis demonstrates that fine-grained mobility information can be inferred from mobile phone data at large scale. This paves the way to new applications that could be further investigated by the research community. In order to make the mobile phone data totally operational, approaches have to be designed to deal with privacy issues and bias in the results obtained with mobile phone data.

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

Résumé

De nos jours, les villes sont confrontées à de multiples enjeux tels que l'urbanisation croissante, la pollution ou la congestion. Une étape fondamentale pour faire face à ces défis consiste à caractériser avec précision la façon dont les gens se déplacent dans la ville. Cette connaissance fondamentale peut être utilisée pour améliorer et optimiser le système de transport urbain. Jusqu'à aujourd'hui, les autorités en charge du transport utilisent des enquêtes de déplacements pour recueillir des informations sur la mobilité des populations. Toutefois, ces méthodes présentent de nombreux inconvénients. Les enquêtes sont coûteuses à réaliser, deviennent rapidement obsolètes, sont inévitablement basées sur des échantillons relativement petits de la population et ne peuvent pas capturer de manière dynamique la mobilité des personnes. Au cours des dernières décennies, avec la digitalisation de la société, l'émergence de nouvelles sources de données telles que les données billetiques, les données GPS, les données géolocalisés issues des réseaux sociaux ou encore les données de téléphonie mobile a attiré l'attention des chercheurs et des acteurs opérationnels pour l'estimation de la mobilité. Les chercheurs ont commencé à étudier l'utilisation de ces données dans le domaine du transport et particulièrement de la mobilité, où elles pourraient permettre des analyses à une échelle sans précédent (tant au niveau spatial que temporel) par rapport aux enquêtes traditionnelles. Parmi les sources de données émergentes, les données de téléphonie mobile sont l'une des sources les plus prometteuses. Ce type de données présente en effet une combinaison unique de propriétés souhaitables : (i) elles offrent un taux de pénétration sans précédent, car disponibles pour l'ensemble des abonnés d'un fournisseur de réseau ; (ii) elles sont enregistrées en continu sur de longues périodes, ce qui permet des études longitudinales fines sur des mois voir des années ; et (iii) elles sont déjà collectées passivement par l'opérateur à des fins de facturation, ce qui en fait une source de données peu coûteuse pour une utilisation ultérieure. Cependant, il n'existe que peu de connaissances sur l'utilisation des données de téléphonie mobile à large échelle pour analyser la mobilité des populations en milieu urbain. Cette thèse vise à apporter des contributions à ce sujet. Dans cette thèse, nous montrons qu'en dépit des limitations bien connus des données de téléphonie mobile (données éparses dans le temps, ayant une large incertitude spatiale et soumises à des phénomènes d'oscillation récurrents), ces données contiennent de riches informations spatio-temporelles qui peuvent être utilisées pour divers applications : la construction de matrices Origine Destination, l'analyse de densité de population ou encore l'inférence des chemins populaires du réseau de transport. Par ailleurs, nous présentons une nouvelle approche TRANSIT (*TRAjectory inference from Network SIgnaling daTa*) qui porte sur l'analyse de données de téléphonie mobile à grande échelle. D'une part, notre approche est capable de distinguer les sessions mobiles des sessions statiques pour un utilisateur donné. Cela permet de reconstruire la demande de déplacement à une échelle spatio-temporelle fine. D'autre part, TRANSIT est capable de réduire fortement l'erreur spatiale des trajectoires de téléphonie mobile. Ainsi, notre approche arrive à surmonter les principales limitations déjà mentionnées. TRANSIT exploite la récurrence de la mobilité humaine: le fait qu'un individu effectuant des déplacements réguliers entre deux zones va généralement prendre toujours le même chemin. Cela crée une redondance dans les informations de mobilité que notre approche TRANSIT utilise pour augmenter la précision spatio-temporelle des trajectoires. Afin d'exploiter pleinement le potentiel des données de téléphonie mobile, nous avons également développé une approche de map-matching qui peut être couplée à TRANSIT. Cette dernière

permet de retrouver le chemin suivi par un utilisateur dans le système de transport multimodal à partir des trajectoires reconstruites par TRANSIT. Pour démontrer le potentiel des approches développées dans cette thèse, nous avons construit quelques applications qui sont rendus possible par nos travaux : l'analyse de la mobilité lors d'évènements atypiques à une échelle spatio-temporelle fine, l'analyse des trajectoires des véhicules roulant sur le périphérique de Paris ou encore un modèle épidémiologique pour l'étude de la propagation du COVID-19. En conclusion, les résultats de cette thèse démontrent que l'analyse de la mobilité urbaine à large échelle est possible avec les données de téléphonie mobile. Cela ouvre la voie à de nouvelles applications qui pourraient être étudiées par la suite. Aussi, afin de rendre les données des téléphones mobiles totalement opérationnelles, deux problématiques principales doivent être traitées : le respect de la vie privé des utilisteurs et les biais que peuvent contenir les résultats issus des données des téléphones mobiles.

Mots Clés : Données de téléphone mobile, Mobilité urbaine, Trajectoires individuelles, Données massives

Acknowledgements

I would like to take this opportunity to thank all the people who contributed in one way or another to this thesis.

Angelo Furno, my supervisor who accompanied me during these three years of PhD. I thank him for his availability (especially at night before article submissions), his scientific clarifications and his precious advices.

I would also like to thank Nour-Eddin El Faouzi, my thesis director, who gave me the opportunity to enter in the research domain. His guidance and the scientific discussions we had were very valuable.

I would also like to thank all the members of my PhD committee that accepted to evaluate my work: Professor Christine Solnon, Doctor Francesco Viti, Professor Edward Chung, Doctor Marco Fiore and Doctor Zbigniew Smoreda.

I also thank all the members of the LICIT for their warm welcome and the pleasant working ambiance.

I thank the Ministère de la Transition Ecologique (MTE) for the funding and for giving me the opportunity to do this thesis. I also want to thank the French ANR research project PROMENADE, which contribute to fund some of the thesis missions.

I would like also to thank Orange, who provided the data used in this thesis and especially Zbigniew Smoreda and Cezary Ziemlicki for their time and help when i worked in the production infrastructure of Orange.

A thought for my family without whom I would not have reached this point.

Last but not least, affectionate thoughts are for my friends who have always supported me.

Contents

List of Figures

List of Tables

List of Abbreviations

BSC	**B**ase **S**tation **C**ontroller
BTS	**B**ase **T**ransceiver **S**tation
CDR	**C**all **D**etail **R**ecords
CP	**C**andecom **P**arafac
CWMA	**C**umulative **W**eighted **M**oving **A**verage
DECRE	**D**etect **E**xpand **C**heck **RE**move
GPS	**G**lobal **P**ositionning **S**ystem
GTFS	**G**oogle **T**ransit **F**eed **S**pecification
HMM	**H**idden **M**arkov **M**odel
LA	**L**ocation **A**rea
LBSN	**L**ocation **B**ased **S**ervice **N**etwork
NSD	**N**etwork **S**ignaling **D**ata
NTF	**N**on-Negative **T**ensor **F**actorization
OSM	**O**pen **S**reet **M**ap
PD	**P**art **D**ieu
SF	**S**aint **F**oy
RMSE	**R**oot **M**ean **S**uare **E**rror
TA	**T**racking **A**rea
TRANSIT	**TRA**jectory inference from **N**etwork **S**ignaling da**T**a

Chapter 1

Introduction

In our society which is becoming more and more digital, the emergence of new data sources has been observed in the transportation domain. These data offers great opportunities to understand and monitor human mobility as well as build future intelligent transportation systems. Among these new data sources, mobile phone data are particularly promising in the transportation research area. Collected at large scale by mobile network operators, the mobile phone data provide the mobility dynamics of millions of mobile phone subscribers, mobility information that is impossible to grasp at such a scale with other data sources. However, mobile phone data still have fundamental issues such as low accuracy in space, sparsity in time and sensitive to oscillation effects which limit their applicability for fine-grained mobility studies. For unlocking the potential of this data, the thesis proposes a set of methodologies and approaches able to overcome these limitations at scale. As a proof of this potential, we develop multiple applications including mobility pattern inference, epidemic model of COVID-19 propagation or fine-grained mobility anomaly detection. Our research makes multiple scientific contributions to the field of computer science and applied mathematics for solving transportation and mobility problems.

In this opening chapter, we first discuss the research context in Section 1.1 and the field of data-driven human mobility analysis in Section 1.2. Then, core data used in this thesis, the mobile phone data, are presented in Section 1.3. Then, in Section 1.4 the research questions that this thesis addresses are stated. We elaborate on our scientific contributions in Section 1.5. Finally, the manuscript organization is presented in Section 1.6.

1.1 Context

The world population has grown significantly and our economies have become more industrialized over the past few hundred years. As a result, many people have moved into cities. The urban population of the world has grown rapidly from 751 million in 1950 to 4.2 billion in 2018. Current forecasts for urban population growth predict an increase, from the actual 55% of the world's population living in urban areas, to 68% by 2050 [42].

With such rapid urbanization, the cities are facing new challenges. From the mobility perspective, there is a growing transportation demand handled by an already saturated transportation system which is hardly expandable. This creates a pressure between the supply and the demand causing several inconveniences. This includes heavy congestion in the road networks or overcrowding in the public transport systems, phenomena which are even more important during peak hours when the demand is high. In addition of a deteriorated level of service, these consequences represent a high social-economic cost, for instance, the congestion cost can amount to several billion dollars in developed cities[1]. To handle properly this additional demand and maintain a good level of mobility service, urban planners and transportation authorities need to optimize the multimodal transportation system. To achieve this goal, a fundamental step is to capture knowledge on the transportation demand. At city scale, transportation authorities have to characterize the urban mobility along multiple dimensions: why, who, when, where, how people move in the city. This knowledge is necessary for building and improving the urban transportation system.

So far, transportation authorities have mainly relied on mobility surveys to capture information on human mobility. However, this method suffers from multiple drawbacks. Surveys are expensive to run, get quickly outdated (around one mobility survey per decade), are unavoidably based on relatively small samples of the population (generally, less than 5% penetration rate of the population in the studied area) and cannot capture fine-grained mobility dynamics in space and time.

At the same time, as our society is becoming more and more digital and with the technology advances, new data sources such as smart cards, Global Positioning System (GPS), Location-Based Social Network (LBSN), or mobile phone data have emerged. These data have gathered the attention of practitioners and researchers for human mobility modeling. In fact, the research community has largely demonstrated the potential of these data in the context of mobility and transportation research [106, 133, 43], [132], where they allow analyses at unprecedented scales compared to traditional surveys [22].

1.2 Data-Driven Human Mobility Analysis

Data-driven approaches for studying human mobility have recently become a new area of research. In the following, we present the main sources of data investigated by the research community. We briefly describe each data source, their strengths, weaknesses and their scope for human mobility studies.

[1] https://www.lapresse.ca/actualites/grand-montreal/201809/13/01-5196357-les-couts-de-la-congestion-evalues-a-42-milliards-pour-2018.php

1.2.1 Global Positioning System (GPS)

GPS use information from multiple satellites to provide the precise geo-localization of moving objects (like smartphones, vehicles) equipped with a GPS receiver. This data can record the trajectories of user's movement with a high degree of spatial accuracy (around 5 meters) and high temporal resolution (in the order of a few seconds). This high spatio-temporal precision allows to analyze human mobility at a very fine scale in space and time. However, due to the high consumption usage of the GPS system on the smartphone, the large volume of collected data and privacy matters, GPS data have only been collected on a small sample of individuals [141]. This limits their applicability for city-scale human mobility studies. In the research area, GPS have been however used for several research studies such as demonstrating the Levy walks nature of human mobility [95], traffic congestion estimation [66], anomaly detection [65] or zones of interest mining in a city [142].

1.2.2 Location-Based Social Network (LBSN)

LBSN data have emerged with the worldwilde expansion of social media. Particularly, the posts made by a user on platforms like Twitter, Facebook and Flickr are associated with geolocation. All the geotagged posts on social networks by the users provide human trajectories that can be analyzed to study human mobility. These trajectories are very precise in space (like GPS data) but very sparse in time (the interevent time between social posts can vary from minutes to hours). Compared with other data sources, social media data has its unique characteristics of associating contextual information, *i.e.*, the social content to the user trajectory. These context-enriched trajectories allow to study human mobility along dimensions that cannot be covered by the other data sources. LBSN data have been used to study human mobility patterns of different social communities through the social interactions of the individuals [73]. These data have also been used to feed a real time monitoring system that capture travel behaviour from Twitter data and can detect abnormal situation in a subway network [51].

1.2.3 Smartcard Data

The public transport systems are becoming more and more developed and popular in the cities. To travel in the public network, people usually carry a smartcard that the user has to check in and/or check out for boarding and alighting public transport stops. Such a system generates, at large scale, massive travel data which include *i.e.*, card ID, stop origin, boarding time, stop destination, alighting time. The strength of smartcard data is to cover quasi-exhaustively the whole public transport trips with very high accuracy both in space and time. Nevertheless, the data cannot provide mobility information on other transportation modes such as car, bike, walk, train. Smartcard data have been largely used for the study of human mobility. Sun *et al.* [107] propose a probabilistic tensor factorization framework to understand aggregated urban human mobility patterns of public transport in Singapore. This approach provides knowledge on mobility patterns in space and time. Egu *et al.* [32] explore the day-to-day variability of public transport users mobility during one month in the city of Lyon (France).

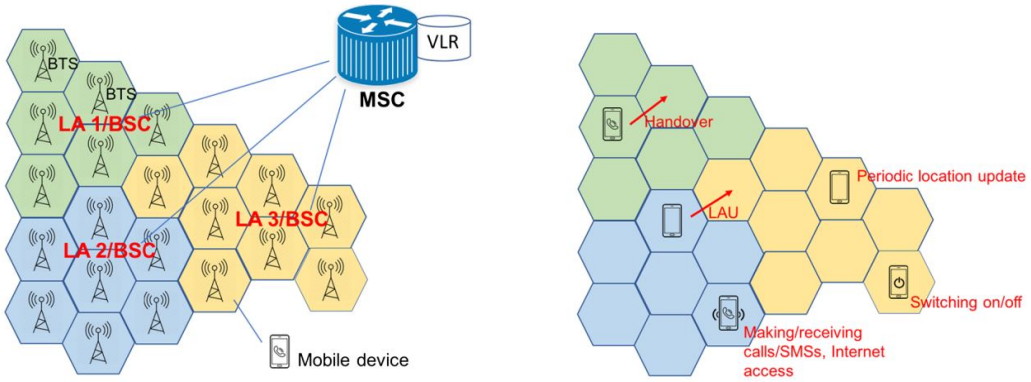
1.2.4 Mobile Phone Data

In recent years, the widespread diffusion of mobile devices and the exploding consumption of Internet traffic via 2G, 3G and 4G technologies have made mobile phone data a crucial source of information in multiple domains. These data are passively collected from mobile network operators for billing purposes and network management. Thus, for research purpose, these data can be used, with no additional cost for studying human mobility. Compared to traditional surveys and above mentioned data sources, mobile phone data present a unique combination of desirable properties. First of all, they offer unprecedented penetration, as they are available for the whole subscriber base of a network provider; then, they are recorded continuously at a large geographical scale and over long time periods, allowing fine-grained longitudinal studies over months or years. Finally, this data covers the mobility of mobile phone subscribers over all transportation modes so that the whole mobility can be covered at large scale (including multimodality). However, despite significant benefits, mobile phone data still have fundamental issues such as large spatial error and sparsity in time that need to be addressed and which limits their applicability for detailed studies on mobility and especially in urban settings. Despite these limitations, mobile phone data have fed plenty of studies related to human mobility. Indeed, they have been employed to derive and validate general laws that govern human movements [38], reconstructing static origin-destination matrices [33], understanding urban land use dynamics [36, 35], or inferring population density shifts in time [31]. A detailed literature review on the use of mobile phone data for human mobility studies and its limitation are discussed in detail in Section 1.3.2.

1.3 Mobile Phone Data

1.3.1 Typology of mobile phone data

First, let us describe how mobile phone data are collected. The mobile phone data are generated by the communications between the communications of the users' mobile phone and the cellular network. The cellular network is composed of a set of base transceiver stations called BTS. Each base station is composed of multiple antennas. Each antenna covers one technology (either 2G, 3G or 4G at the moment) and one azimuth (usually an antenna covers an angle of 120 degree). Each BTS covers a defined area, known as a cell, which is the smallest spatial entity in the cellular network. Besides, there are base stations controllers called BSC which are managing a group of base stations. The area covered by a BSC is called Location Area (LA). In order to ensure the quality of communication services, mobile network operators have to monitor locations of subscribers' mobile phones. Thus, mobile phones are constantly and frequently communicating with the cellular network. These communications have two main characteristics: the technology of the antenna on which the communication is done and the event responsible of the communication triggered by the mobile device or the cellular network. There are multiple kinds of events: *i)* communication events (*i.e*, calls and SMS); *ii)* handover events (i.e., base station change during an established communication); *iii)* network attachment/detachment events; *iv)* data/internet connections; *v)* Location Area (LA) updates (i.e., base station change during a communication resulting in a change of Location Area) and *vi)*



(a) Cellular network structure

(b) Network events captured by the cellular network

Figure 1.1: Illustration of cellular network (from Huang *et al.* [46])

Tracking Area updates (i.e., the mobile phone did not generate any event for a period of times, typically few hours). An illustration of the cellular network structure is given in Figure 1.1.

Based on their temporal and spatial properties, different kinds of mobile phone data can be observed. On the temporal dimension, the data can differ on the set of events collected and available by the mobile network operator. The most popular kind of mobile phone data is Call Detail Records (CDR); events considered are only communication events. Different from CDR data, another kind of mobile phone data is Network Signaling Data (NSD) which report on multiple kinds of events *(i)* communication events; *ii)* handover events; *iii)* network attachment/detachment events; *iv)* data/internet connections; *v)* Location Area (LA) updates and *vi)* Tracking Area updates). For a given mobile phone operator, NSD corresponds to all events generated by the mobile phone of the subscribers. An event is characterized by four attributes: the mobile phone ID which is pseudo-anonymized, the timestamp of the event, the antenna communicating with the mobile phone and the event type. A sample of NSD is given in table 1.1.

This additional event considered increases the temporal sampling frequency of NSD compared to CDR. The spatio-temooral granularity of CDR and NSD is discussed in Chapter 3.

Mobile Phone ID	Timestamp	Cell ID	Event Type	Technology
922**********2440000	2019-06-01 9:00:05	123456	Voice	3G
922**********2440000	2019-06-01 9:10:15	234567	Text	2G
922**********2440000	2019-06-01 9:12:23	234567	Text	3G
	...			
922**********2440000	2019-06-01 9:32:23	234567	Data connection	4G
922**********2440000	2019-06-01 9:38:54	345678	Handover	4G

Table 1.1: Example of rows of NSD

On the spatial dimension, the event can either be mapped to the position of the antenna or can be a reconstructed position of the mobile phone relying on additional

information such as signal strength, triangulation of the signal received by nearby antennas. The last kind of data are referred to as sighting data. The reconstructed positions based trajectory from sighting data are much more accurate in space compared to the antenna based trajectory. In this work, we will mainly focus on CDR and NSD which are the most popular sources of mobile phone data. Sighting data are not considered.

Finally, a sequence of events related to the same mobile phone can be considered as a trajectory called mobile phone trajectory or cell phone trajectory. By assuming that the mobile phone's user connects antennas which are close to the position of the user, the mobile phone trajectory can be seen as an approximation of user's mobility. Available with large penetration rate, these trajectories can capture human mobility at large scale. Considering mobile phone data as a set of individual cellular trajectories, each cellular trajectory being an approximation of the user mobility is a core concept of this thesis.

1.3.2 Mobile Phone Data in Human Mobility Studies

In recent years, the mobile phone turned out to be a crucial source for human mobility modelling. According to Naboulsi *et al.* [80], the works leveraging mobile network data for human mobility studies can be classified into two main categories. The first one includes the works that study physical models able to reproduce typical mobility patterns at individual and aggregated levels. The second one include works processing mobile phone data to provide knowledge on transportation demand. In the first category, some works study the fundamental laws that govern human mobility. By studying the number of cells visited by mobile subscribers, Halepovic *et al.* and Paul *et al.* [41, 86] show that this distribution is heavy tailed. As a direct result, they show that a lot of the users (around half) visited one cell whereas there are few users who visit hundred of cells in a week. This phenomenon reveals high heterogeneity in terms of mobility behaviour of the users. Besides, the number of visited locations, the travel distance of the users distribution have also been studied by the research community. Gonzales *et al.* [38] and Song *et al.* [103] show that the distribution of the total travel distance is a truncated power law. This result has been validated by similar works over different regions: USA [19], Europe [38] or Africa [84]. Another important result that has been demonstrated using mobile phone data is the high spatio-temporal regularity of human mobility. Indeed, users movements exhibit strong periodicity over time and this periodicity is multi-scale (daily and weekly scales) [44]. In space, a strong regularity is also observed over the sequence of the cells visited in the users [136]. Moreover, a direct result from the repetitive nature of human mobility is the predictability of individual movements. Song *et al.* [102] show that 93% of individual movements are predictable. Other works aim at developing physical models able to reproduce the above mentioned laws of human mobility. At individual scale, we can quote the main models proposed in the literature: the Lévy Flights [95], CTRW [103] and preferential returns [103]. A more exhaustive presentation of these models as well as a discussion on the strengths and limits of each model can be found in [11]. At the aggregated level, models are proposed to characterize mobility flows at city-scale or even country-scale. The most popular model is the gravity model [144] which stipulates that the flow between two zones is proportional to the population of each zone and inversely proportional to the distance between these zones.

In the second category, some works aim at inferring transportation demand information from mobile phone data. For instance, the work [10] uses mobile phone data to estimate travel times with low error compared to those obtained with traditional loop detectors. Janecek *et al.* [53] also used network signaling data to characterize congestion on highways. The good results obtained with highway traffic can not be easily reproduced in the more complex and heterogeneous urban environment. Their result shows that the mobile phone data was able to detect traffic anomalies more efficiently compared to other sources of data such as GPS data from taxis. Caceres *et al.* [17] used mobile phone data for travel volume estimation and were able to infer traffic volume with relative error around 20% compared to those obtained with loop detectors. Another important task that received a lot of attention from the research community is estimating Origin-Destination matrices [33, 8]. Besides, other research study the question of understanding demand by transportation mode. The transportation mode inference is often tackled in simplified cases. Approaches aims at distinguishing motorized vs not motorized modes, road vs rail mode, private vs public modes. In such simplified cases, the approaches can exhibit rather good performance (around 80% accuracy). Another field of study is the transportation system itself. Berlingerio *et al.* [12] demonstrate using cellular data that adding 4 lines to public transport system would allow an improvement of overall travel times as high as 10%. Finally, several works show that mobile phone were able to accurately estimate the population density [30, 8].

Our thesis focuses on the second category of work and especially travel demand estimation at large scale and in urban scenario.

1.3.3 Limitations

Despite significant benefits and promising applications, mobile phone data still have fundamental issues that need to be addressed due to low accuracy along both the spatial and temporal dimensions which limits their applicability for detailed studies on mobility and especially in urban settings. In order to fully unlock their potential, there are still main issues that have to be addressed.

The main limitations are the following:

- The mobile phone data are passively collected by the mobile network operator: they are not collected for transportation purposes like smartcard or GPS data. As a result, they do not contain any direct mobility information. The mobility information has to be inferred from the raw data which is a challenging task given the other limitations.
- NSD are highly subject to noise. One source of noise is the so-called ping-pong effect. While static, the mobile phone of a user can connect to multiple antennas. Thus, it can be tricky to successfully infer that a user is actually static while he is moving from the perspective of the cellular network.
- NSD have large spatial error. The spatial granularity available with NSD is the base station level. The base stations can have a range from hundred meters in urban areas to a few kilometers in rural areas.
- NSD are sparse in time. NSD sampling frequency is in order of magnitude of minutes which is sparse and makes the mobility inference of the user challenging since the information level is low.

1.4 Research Gaps and Open Research Questions

Given this overview of the use of mobile phone for human mobility studies and the above mentioned limitations of mobile phone data, multiple gaps have been identified in the literature and will be studied in this thesis.

First, the thesis aims at investigating the question on the representativeness, biases and suitability of the mobile phone data for travel demand estimation. Due to lack of comparative mobile phone and survey data, this question, yet fundamental, is rarely discussed in the literature. Then, we notice that the above mentioned limitations of mobile phone limit the use of NSD for fined grained mobility study in urban environment. Thus, the thesis studies approaches to overcome these limitations. Finally, after overcoming NSD limitations, our thesis focuses on advancing current state of the art applications based on mobile phone data.

In summary, the objective of this thesis is to develop data-driven approaches on large scale mobile phone data capable of estimating human mobility at unprecedented spatio-temporal granularity. Centering around this research objective, we further formulate the following research questions (RQs) that have only been partially addressed by existing studies.

- RQ1: To what extent are mobile phone data suitable for estimating human mobility? - Chapter 2
- RQ2: Can the repetitive nature of human mobility be used to improve in space and time human trajectories as observed through the bias of mobile phone data? - Chapter 3
- RQ3: How can we estimate very fine mobility information *i.e.*, the path traveled on a multimodal transportation network, from mobile phone data? - Chapter 4
- RQ4: How can we derive aggregated mobility patterns along the main dimensions that characterize human mobility? - Chapter 5
- RQ5: What kind of applications are made possible by our approach when applied on mobile phone data? - Chapter 6

1.5 Research Contributions

This thesis provides several contributions to research on data-driven modeling for estimating urban mobility with mobile phone data. These contributions are listed below :

- Dataset collection and analysis of rich real-world mobile phone datasets. We lead an experimentation that allowed to gather both GPS and, thanks to the collaboration with Orange France, mobile phone traces related to a group of users in the Lyon metropolitan area, France. Given the scarcity of ground truth data in mobile phone related studies, the collected dataset is quite unique. We have tested our approaches on real-world, massive mobile phone data provided by Orange including Terabytes of data, millions of subscribers and billion of network signaling logs from 2G, 3G and 4G events. To the best of our

knowledge, this mobile phone dataset is the largest in terms of population covered and number of mobile phone events generated.

- TRANSIT *(TRAjectory inference from Network SIgnaling daTa)*, a new framework that processes mobile phone data to (*i*) tell apart movement intervals from stationary activity periods for each mobile device, and (*ii*) infer fine-grained human mobility trajectories during the associated movement intervals. The validation on the ground-truth dataset showcases the superior performance of TRANSIT (80% precision and 96% recall) with respect to state-of-the-art solutions in the identification of movement periods, as well as an average 190 m spatial accuracy in the estimation of the trajectories.
- An approach for the challenging problem of mapping cellular trajectories to the multimodal transportation network at scale and in urban settings. The latter is based on Hidden-Markov Model.
- A bunch of new large-scale applications such as the anonymization of OD-matrix, the modeling of the COVID-19 propagation built on the use of mobile phone data. These applications are made possible by the data-driven approaches proposed in this thesis and have not been or scarcely explored in the literature of the field.

1.6 Manuscript Organization

The thesis is structured as follows :

Chapter 2 is dedicated showing the potential of mobile phone data as a source of data to study human mobility. Preliminary approach based on the literature and comparative analysis with surveys show that mobile phone data have a good potential to estimate travel demand at large scale.

Chapter 3 aims at overcoming the limitations discussed in Chapter 2. We propose a novel framework TRANSIT capable of processing mobile phone data to accurately distinguish mobility phases from stationary activities for individual mobile devices, and reconstruct, at scale, fine-grained human mobility trajectories, by exploiting the inherent recurrence of human mobility. With such approach we are able to study human mobility at unprecedented spatio-temporal granularity

Chapter 4 builds upon Chapter 3. In this chapter, we develop a Hidden Markov Model based map-matching approach to infer the exact path travelled by the subscriber on the transportation network given its mobile phone trajectory. The approach is applied on the result of TRANSIT, *i.e.*, the enhanced human mobility trajectories extracted from mobile phone data from Chapter 3.

Chapter 5 uses the result of the approach from Chapter 3 to derive an aggregated mathematical representation, a mobility tensor which is leveraged to infer aggregated mobility patterns in a city. The proposed representation, which is data agnostic, is applied on mobile phone data. The use of TRANSIT output allows to derive these mobility patterns at unprecedented spatio-temporal granularity.

Chapter 6 proposes a non-exhaustive overview of applications that are unlocked by the use of mobile phone data and the data-driven approaches proposed in this thesis. The open challenges that still need to be tackled in future works are also discussed.

Chapter 2

Travel Demand Estimation at Regional Scale and Comparison with Surveys

---

As an entry point of this thesis, this chapter study the capability of existing approaches of the literature to estimate travel demand with NSD at large scale. Thus, we develop a comparative analysis between the demand profile obtained with NSD and those obtained with surveys. While encouraging, the results show spatio-temporal biases that have to be tackled for inferring more accurately travel demand with NSD. Based on this observation, we propose a simple yet effective approach for debiasing NSD. The latter allowed to further analyze travel demand at finer granularity. Particularly, we extract travel demand patterns at regional scale with NSD. The results obtained have been validated with surveys and external data sources.

The chapter is structured as follows. Section 2.1 discusses mobile phone data and surveys as sources for studying human mobility. Section 2.2 presents the literature on travel demand estimation with mobile phone data and its validation. Then, the core of this chapter is presented in Section 2.3. This section includes the comparative analysis between travel demand inferred with NSD and surveys, an approach for debiasing NSD and finally extraction of travel demand patterns. In Section 2.4, we discuss the results as well as the perspectives for the rest of this thesis. Finally, Section 2.5 presents the conclusion of this chapter.

This chapter contains parts of the article [34]:

Fekih M., **Bonnetain L.**, Furno A., Bonnel P., Smoreda Z., Galland S., Bellemans T., (2021), "Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: a large-scale comparative study with travel survey and land use data". In: *Transportation Letters*.

2.1 Introduction

Spatiotemporal information about people movements are extremely valuable for human mobility analysis and transportation development purposes [11, 5]. Emerging forms of data generated by pervasive communication systems such as cellular networks are offering new opportunities to track individual-level movements and enhance our understanding of travel behavior patterns [22, 96]. Indeed, mobile phone records are characterized by a low collection cost since they are produced automatically and passively by telecom operators. More interestingly, the existing network mechanism provides continuous temporal and spatial information about individuals' whereabouts. Therefore, massive cellular network data provide a promising source for acquiring information about travel demand, exploring the various factors that might impact community travel flows and supporting long-term policy decisions on large-scale mobility.

The traditional human mobility research relies on household travel surveys that typically record one day of travel diaries per household. Yet, there are notable limitations associated with the classical travel survey process [128, 105]. Collected survey data can be useful to capture cross-sectional snapshots of daily journeys. However, they do not allow considering fine-grained temporal analysis of e.g., the hourly, weekly, or special-events related variability of individual trip flows [67].

Understanding the dynamics of human mobility patterns is a core notion in transportation studies related to, e.g., traffic congestion management and transport infrastructure planning [100, 134]. Among all possible human mobility patterns, dynamic origin-destination flows remain the most used by practitioners. A number of studies have been conducted to extract this pattern using different forms of mobile phone data. The majority of these studies have explored Call Detail Records, called CDR, (i.e., billing data) and developed techniques to figure out temporal distribution of user trips in limited geographical areas. However, it has been shown that these methods perform rather poorly, especially in urban zones, due to the very low spatio-temporal resolution of CDR data [123, 140]. Moreover, few research works have validated the results against external mobility data sources ([3, 14, 55]. Yet, the validation process allows to identify possible biases and to have a clearer idea about the potential of cellular data.

Fekih *et al.* [33] developed a full workflow to transform cell phone network logs into individual trip flows and showed the potential of the method to generate static origin-destination flow matrices. Based upon this work, the focus here is to explore network signalling data collected from 2G and 3G networks to extract dynamic travel patterns of mobile phone users within large-scale area. The aim of this research is therefore to assess whether these massive signalling traces could act as reliable data source to capture real-world temporal mobility behaviours.

As a case study related to the Rhône-Alpes region, France, we conduct a comparative analysis of the hourly trip flows estimated via Fekih *et al.* on NSD approach and those obtained from the latest travel survey performed in the same region. Along the comparison, new techniques are introduced to cope with the spatio-temporal biases detected in the signalling data-based demand estimation. Moreover, we leverage the proposed methodology to capture groups of zones that consistently behave in a similar way with respect to the emitted, estimated time-varying demand. To this end, a spatial clustering process is applied resulting in identifying comprehensive

different temporal demand patterns within the study area. Then, advanced analyses are conducted by combining the obtained travel demand profiles with land use data available in the observed region. That helps to highlight the correlation between land use characteristics and trip generation as well as to reveal meaningful and significant dynamic mobility patterns of mobile phone users.

2.2 Related Work

Considerable efforts have been devoted to extract OD flows and estimating time-dependent travel demand from cell phone communication logs. Significant attempts have been made to study trip distribution differences over weekdays and weekends [21], to generate O-D flows by purpose and time of day [3] and to reconstruct the travel mode and flows in each link of the transportation network to perform traffic assignment [116]. There have been several limited-scale research aimed at identifying temporal movements in urban areas. In 2010, Ahas *et al.* [2] carried out a research work using cell phone positioning data of a random sample of 277 respondents. They analysed the diurnal rhythms of the city life and its spatial differences in Tallinn, Estonia and showed that the majority of users had a similar temporal rhythm. Kang *et al.* [59] proposed to study how mobility patterns inside eight cities in China, are affected by the compactness and the size of the area. The results obtained from CDR data analyses indicate that the distribution of intra-urban travel follows the exponential law and that individuals living in large cities need to travel farther on a daily basis. More recently, in Trasarti *et al.* [117], CDR data have been used to extract interconnections between different city areas that emerge from correlated temporal variations of local population densities. In the same perspective, study on the dynamic urban activity patterns and interaction between areas has been performed in Dakar, Senegal [76]. The authors highlighted high interactions between areas with similar land use characteristics. Based on activity-based modelling approach, Widhalm *et al.* [127] have extracted activity behavioural patterns based on trip departure time, activity types and frequencies combined with spatial typologies and land use data. By leveraging CDR data, they applied the method in the cities of Vienna and Boston showing similarities between conurbations. The resulting trip chains and activity patterns match well with data from surveys even though the inferred activity classes do not directly correspond to those of surveys.

Moreover, mobile phone data have been explored to generate traffic origin destination flows and estimate relevant temporal mobility metrics within different urban areas. Following a trip-based approach, Gundlegard *et al.* [40] proposed a process for dynamic travel demand estimation using two CDR datasets collected in Ivory Coast and Senegal. They computed relevant mobility metrics such as route and link travel flows and travel time. However, the derived estimations were not evaluated due to the lack of validation data. The travel demand scaling for the full populations of the two studied areas is not discussed. Similarly, Wang *et al.* [124] have studied CDR data to estimate dynamic OD traffic flow and traffic demand by time-of-day in the Kansas Metro corridor, US. They conclude that the used cell phone data would be more suitable for long distance or inter-city trips' extraction due to the low location resolution. Notably, the travel demand dynamics of the whole population have not been addressed in detail. Instead, most of the research have focused on dynamic road traffic demand estimation by combining/validating cell phone data with available road traffic counts, which are only measuring vehicle traffic volume and not

moving travellers during a time interval (Huang and Xiao 2018) [47].

Furthermore, the fundamental question on the representativeness and biases of the analysed data is rarely discussed. Indeed, cell phone data have key attributes that are different from travel surveys and which should be carefully interpreted during the processing step. While the existing state-of-art research employs several types of mobile phone data with different sample sizes and characteristics, they still did not provide satisfactory rules to properly deal with these passive travel data contents [68]. In fact, additional work is needed to implement adequate standards and guidelines for data cleaning and processing. Also, more focus should be on addressing the intrinsic existing biases (e.g., sampling, temporal or spatial) and assessing their impacts to address the concerns of transport modelers. Moreover, the expansion and evaluation of the results against external sources need to be addressed extensively to fully check the relevance of estimations for travel demand prediction and decision-making purposes.

2.3 Case Study : Lyon

The idea is to study to which extent NSD capture accurately the travel demand dynamics at regional scale compared to survey and extract meaningfull insights from the data. One the one hand, the explored NSD dataset includes 2G and 3G signaling records from June 2017 of over 2 million mobile phone users and covers the entire Rhône-Alpes region in France. We have analyzed the 24-hour period data collected from 1st June 3:00 am to June 2, 2017 3:00 am. Figure 2.1a presents the spatial distribution of 2G/3G cell towers and the administrative sector zoning considered in the region. On the other hand, the regional travel survey, called EDR 2015, was conducted in the Rhône-Alpes region between 2012 and 2015 (EDR-RA, Conseil régional Auvergne Rhône-Alpes 2016). Specifically, 37,450 individuals, aged over 11 years, have been surveyed, and 143,000 trips have been reported. The region has a population totaling 5.2 million inhabitants aged over 11 years and covers an area of 43,700 km2. The survey sample has been constructed according to a geographical stratification which corresponds to the 77-sector zoning system shown in Figure 2.1a.

2.3.1 Preprocessing Steps

The work developed in this chapter is built upon recent work made by Fekih *et al.* [33]. The idea of their work is to transform raw network signaling data into dynamic OD-matrices. Their framework include 4 main steps:

- NSD filtering: this step aims at keeping data which are exploitable to analyze human mobility. They analyzed the number of events per day generated by the mobile phones. A mobile phone is removed from the dataset if its number of events per day is lower than 4 (there are not enough events to perform trip extraction) or if the number of events per day is greater than 1000 (such a high frequency of event is not imputable to human behaviors, but very likely caused by device anomalies).
- Home detection and resident filtering: for each user, the most frequent observed cell tower is calculated and assigned to the corresponding sector. The

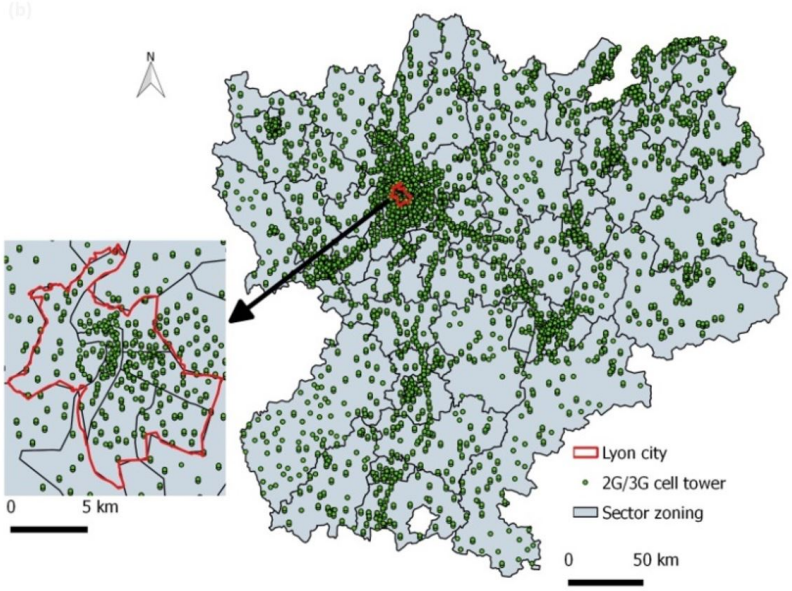


Figure 2.1: Cell tower (2 G and 3 G) distribution and administrative sector zoning in the Rhône-Alpes region with zoom on Lyon city

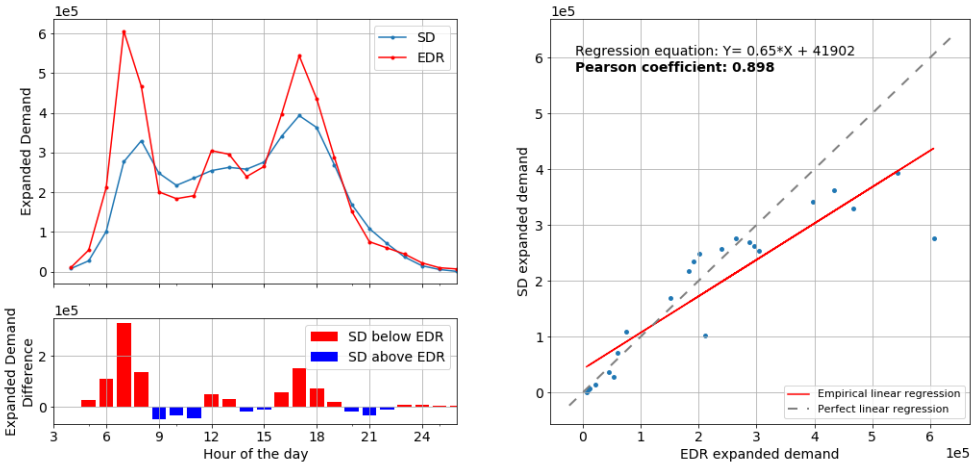
latter is considered as the home location zone of the user. Once the home detection done, all users whose home location is outside the study area are filtered.

- Trips extraction: first for each mobile device (user), the events are sorted by timestamps and the coordinates of the antennas are mapped to sectors. Then, the stationary activities of each user are identified as a set of consecutive events observed at the same sector level based on a minimum stationary time threshold. A trip, between two stationary activities, has a start time calculated as the average between the last time timestamp of the event in the previous stationary activity and the first timestamp event of the next stationary activity. The sector level associated to the previous stationary activity is the origin of the trip and those associated to the next stationary activity is the destination of the trip.

- Trips scaling: as the user samples involved in the signalling datasets represent only a fraction of the population, therefore, the identified trips need to be properly scaled in order to be representative of the full population mobility. Using the resident estimations obtained from home detection step, an expansion factor can be calculated for each filtered user as the ratio of the census population and the number of residents estimated in his home sector

More details about this methodology can be found in the work by Fekih *et al.* [33]

In this study, we retain an activity time threshold of 30 minutes to detect trips. It does not appear recommendable to consider time thresholds which are much lower than 30 minutes, as multiple false-positive stationary detections may occur, yielding false-positive trips. Also, given this assumption, we pre-processed the EDR data and applied the same time threshold to make our comparison fair and realistic, by considering trips taking place between activities with duration more than 30 minutes. It is worth recalling that the following analyses are based on signaling data



(a) (b)

Figure 2.2: (a) Temporal demand profile (number of hourly trips generated from all zones) from signaling data (SD) and survey data (EDR) and demand difference between EDR and SD (red bars when the hourly demand from EDR is higher than that from SD and blue bars otherwise) (b) correlation between hourly demand estimations from SD and EDR

collected during one typical working day -Thursday- which is traditionally considered (in transportation surveys) as representative of an average weekday. The analyses are presented at sector level by removing the intra-zone trips, given our focus is on inter-zone flows.

2.3.2 Travel Demand Estimation

Based on the above of mentioned methodology we are able to compute from NSD, the temporal travel demand profile which corresponds to the number of hourly trips generated from all zones. The temporal travel demand profiles for signaling and survey data are shown in Figure 2.2a. As a first insight from our analyses, the signaling-based demand profile exhibits less sharp morning and afternoon peaks compared with the survey. The total demand observed from signaling data is thus lower than the one reported in the survey, as shown in Figure 2.2b. This result could be explained by the existence of a certain large fraction of users in our mobile phone data, referred to as "static people" in the following, for whom it is possible to detect the home sector but no trip can be observed, as the only stationary activity produced is performed at the home sector. The proportion of such static people in our mobile phone dataset amounts to 46%. It is noteworthy that signaling data-based travel demand is to some extent correlated to the number of events generated from resident users. Therefore, even though a certain portion of static users could be actually stationary (e.g., elderly people), it appears highly likely that another large portion of them could be mobile, but due to their very low mobile phone activity (e.g., during morning hours), no associated trips have been identified. Such reduced device usage patterns inevitably lead to an underestimation of the travel demand especially during morning period and requires a proper de-biasing procedure which is further detailed in the next section. However, despite this underestimation, the

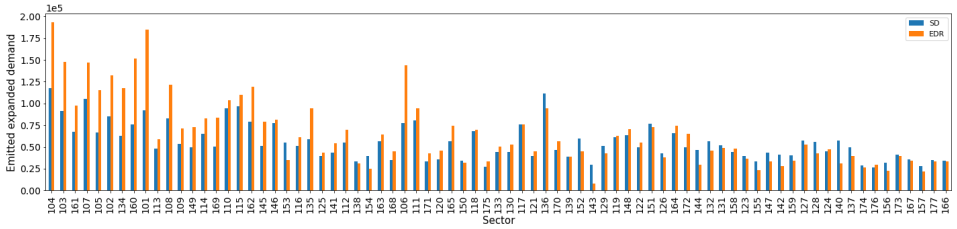
hourly global demand profiles estimated from both data sources are highly correlated (Pearson coefficient equal to 0.94) as shown in Figure 2.2b. This confirms the fact that signaling data can provide a travel profile comparable to the well-known typical demand profile for a working day.

After analysing the demand estimation difference between signaling and survey data at temporal level, the difference between the demand emitted from each geographical zone has been studied. Figure 2.3a shows the number of trips emitted from each sector (as the trip origin). For 45 sectors, the number of emitted trips is higher when compared to survey (median relative difference of +0.20%). It is instead lower for 32 areas (median relative difference of -0.21%). These differences can be better interpreted by relying on the map shown in Figure 2.3c, which is a spatial representation of the absolute difference between the demand generated by each zone from mobile phone and survey data. The emitted demand estimated with mobile phone data tends to be higher in rural areas and lower in urban dense areas. In rural areas we can reasonably assume that signaling records provide more consistent estimations, since long-distance trips from/to these areas are typically better-captured with mobile phone passive data than surveys and for a larger sample of the population (Janzen *et al.* [54]). Instead, in urban areas, it seems that the proposed trip extraction method is unable to capture short-distance trips, which are expected to occur with higher frequency in urban areas than in rural one. Indeed, it is not obvious to differentiate noise from short-distance trips with mobile phone data. Despite such limitation, the total number of trips emitted by each zone based on cellular and survey data remains highly correlated (Pearson coefficient equal to 0.86) as shown in Figure 2.3b.

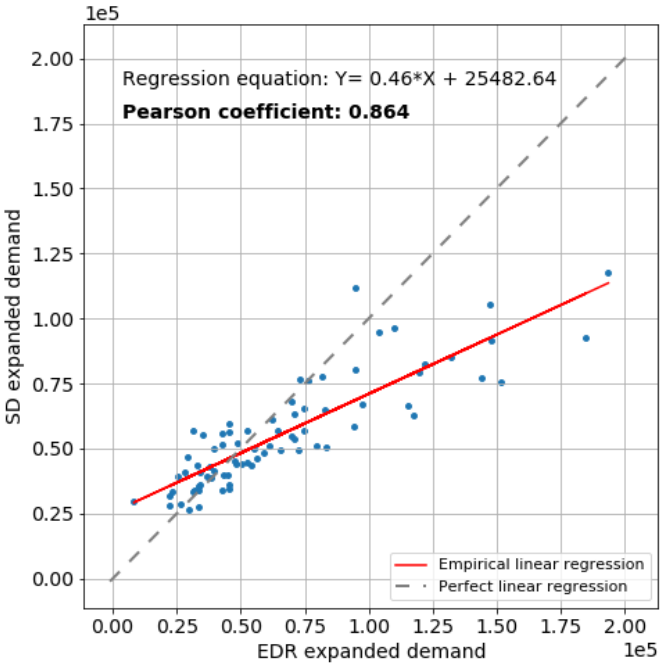
Figure 2.4a, representing the emitted demand difference per zone and per hour, confirms the spatial and temporal bias previously observed. In the figure, zones are sorted from left to right by decreasing density of urban land use, as retrieved from the (*CORINE Land Cover 2012*) dataset. On the one hand, in highly-dense urban zones (on the left of the Figure 2.4a), the demand is higher in survey compared to mobile phone regardless of the hour of the day. On the other hand, the hourly travel demand during the morning peak period is higher in the survey compared to mobile phone regardless of the zone. These preliminary analyses let us identify a "systematic" bias present in the data. In the following, we propose a heuristic-based method to cope with these biases.

2.3.3 De-biasing Procedure

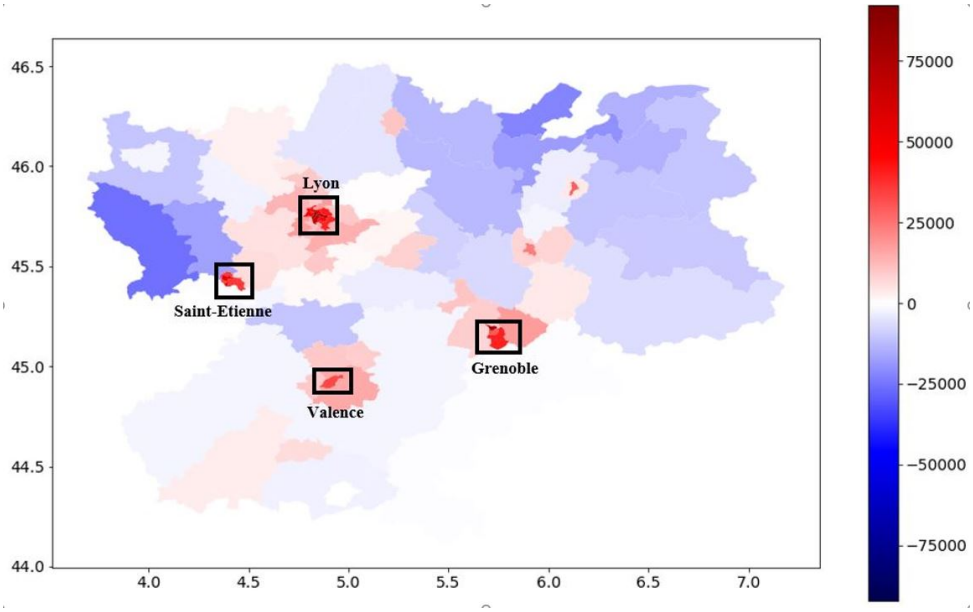
Based on the previous considerations on the biases present in signaling data-based demand estimation in both spatial and temporal dimensions, two different approaches have been proposed for mitigating such biases. Concerning the spatial correction, Figure 2.4b shows that the emitted demand difference between survey and mobile phone is abnormally highly-correlated to the urban density of the sectors. Specifically, the difference is much higher for denser urban areas compared to rural one. This aspect can be interpreted as an underestimation of the urban area travel demand in the case of the signaling data, due to the previously discussed difficulty of such data in capturing short-distance trips (more likely to happen between adjacent/smaller urban areas). In order to address this bias, we have thus applied a spatial correction factor estimated per zone and calculated using the regression



(a)

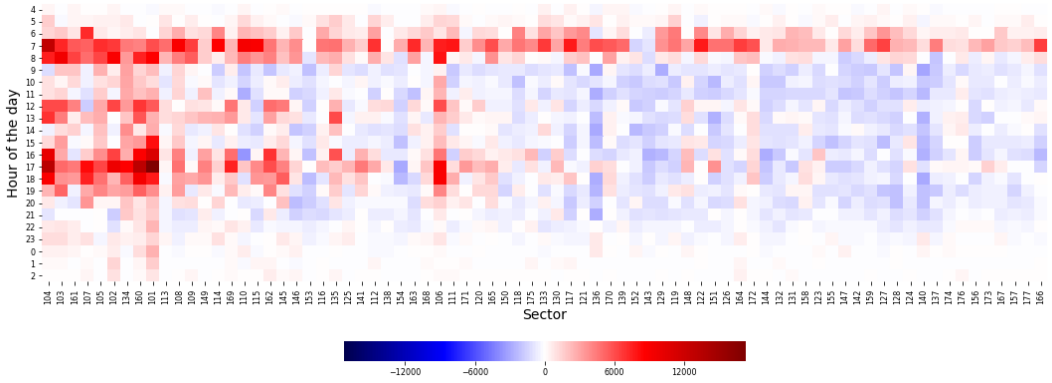


(b)

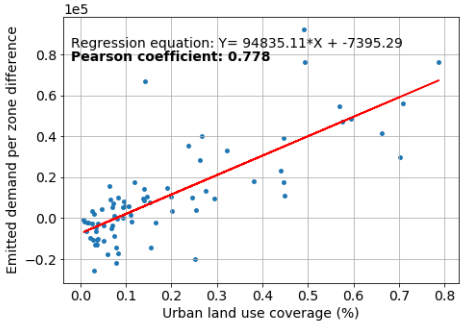


(c)

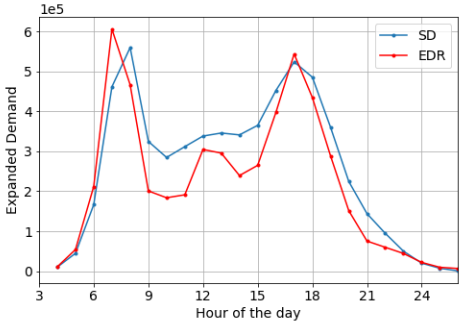
Figure 2.3: (a) Emitted demand per zone (each zone is described on the x-axis by its sector ID). The zones are sorted (from left to right) in descending order of urban land use percentage per zone. (b) Correlation between signaling data (SD) and survey (EDR) emitted demand per zone and (c) heatmap of the emitted demand difference per zone from SD and EDR



(a)



(b)



(c)

Figure 2.4: (a) Spatio-temporal distribution of the emitted demand difference between signaling data (SD) and survey (EDR) (b) correlation between emitted demand difference (EDR – SD) and urban land use percentage per zone and (c) hourly demand distribution of EDR and SD after application of the de-biasing procedure

equation shown in Figure 2.3b:

$$D_{Survey}(x) - D_{SD}(x) = 94835 \cdot U_{la}(x) - 7395 \tag{2.1}$$

The expression of this factor is the following:

$$S(x) = 1 + \frac{94835 \cdot U_{la}(x)}{D_{SD}(x)} \tag{2.2}$$

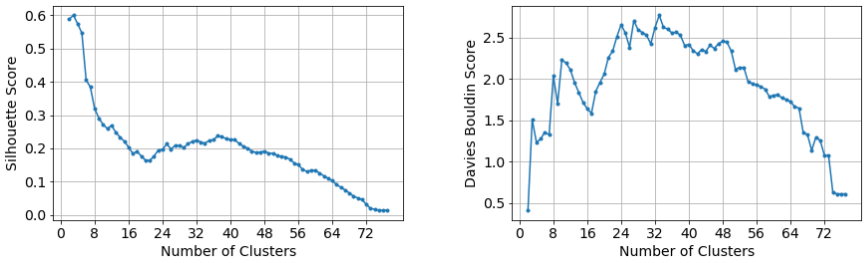
where $U_{la}(x)$, $D_{Survey}(x)$ and $D_{SD}(x)$ represent respectively, the urban density (as computed from land use data) and the emitted demand associated to each zone x (as estimated from the signaling data). By applying this correction factor, we can decorrelate this difference with respect to the urban density.

Concerning the temporal bias, it shall be noted that this phenomenon is often observed in mobile phone data (e.g., especially in CDRs) and has been raised by several researchers without being resolved [40, 56]. In our case, we have noticed that the observed underestimation (i.e., between 5 and 9am) of the travel demand also appears in the temporal distribution of LAU (events passively generated by the device and not by explicit users' communications, i.e., when changing the Location Area zone). Also, the distribution of periodic events (e.g., LAPU) which implicitly reflect the active/idle behaviour, has been analysed. It shows that during the early morning period [3-9am] the residents are notably less active, hence generate less cell phone logs, with high number of LAPU events. In addition to that, users usually tend to turn off their terminals during night period and until early hours of the day. Therefore, we can reasonably assume that the observed underestimation of trip flows is due to the low total volume of signaling traffic during morning peak hours. To address this bias, a uniform correction factor has been applied on all mobile phone-based trips with a start time estimated during and around the morning peak period [5-9am]. This factor has been calculated as the ratio of the afternoon and morning peaks in the LAU profile, allowing to consider the non-observed cell phone transactions and extract the hidden information from them. Based on these considerations, the applied temporal correction factor results equal to a value of 1.3. It is important to note that the temporal correction is a de-biasing procedure totally independent from the survey data (used for the comparative analysis), thus being easily reproducible by solely relying on information collected by the mobile phone operator (i.e., the LAU events distribution). After applying both the spatial and temporal correction factors to the trips reconstructed via network signaling data, the travel demand profile has been recomputed in Figure 2.4c, which clearly exhibits higher correlation with respect to the survey-based one.

2.3.4 Spatial Clustering and Travel Patterns Extraction

This study aims at extracting meaningful travel patterns of the residents' OD trips within the region and to analyse their dynamics during the day. By aggregating all individual trips generated from each zone on the same time window, the temporal demand profile emitted from each given area can be derived. Hence, in order to identify the different existing mobility patterns through the region, an unsupervised clustering method is proposed to group zones with similar travel behaviors. Thus, a hierarchical agglomerative clustering is performed on the normalized hourly temporal profiles of the 77 studied sectors. The normalization consists in dividing each hourly volume by the total daily volume of the profile. The correlation coefficient

has been used as similarity measure between profiles. Two indicators are computed, i.e., the Silhouette [97] and Davies–Bouldin indexes [29], to decide the number of clusters to select. These indicators are depicted in Figure 2.5. A good clustering is heuristically associated to lower values of the Davies Bouldin and higher values of the Silhouette index. Therefore, according to such rules of thumb, several cluster configurations have been analyzed (e.g., 3, 7 and 9), preferring in the end the clustering associated to a choice of 9 clusters, that allowed us to discriminate a larger number of finer-grained patterns (which corresponds to a local minimum of the Davies Bouldin index and relatively high values of the Silhouette index). The results were consistent among the different analyzed cluster configurations. Once the clustering is achieved, each zone is characterized by its own temporal profile, the cluster number and the average temporal profile of the cluster to which the zone belongs to.



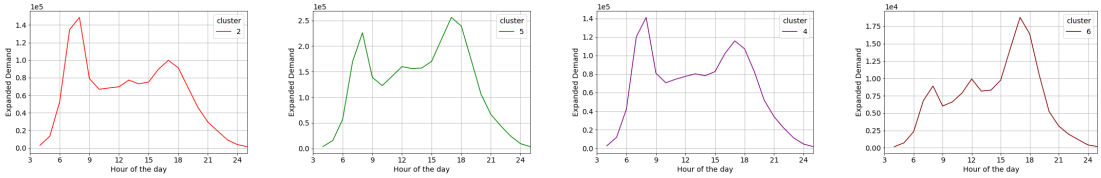
(a) Silhouette score (b) Davies Bouldin score

Figure 2.5: Distribution of clustering scores

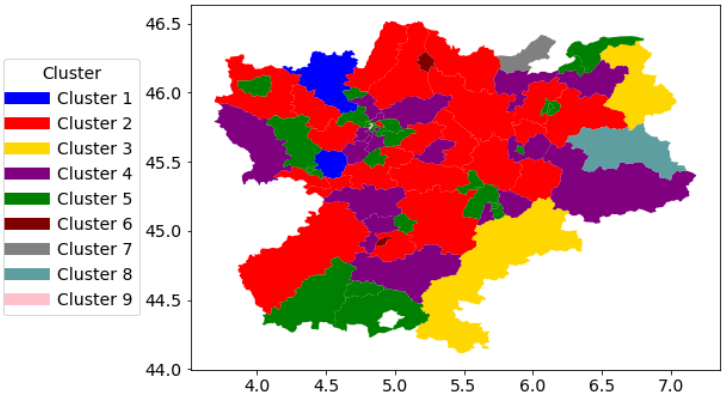
Among the 9 clusters, we could distinguish 3 main clusters, each including at least 18 zones, and 6 minor clusters with at most 2 zones each (zones in each cluster represent trip origins). The map representing all these clusters is shown in Figure 2.6e.

The average temporal profiles, noted as "ATP" in the following, of the 3 main clusters (2,4 and 5) as well as one minor cluster (6) are represented in Figure 2.6. Based on these demand profiles, the following interpretations can be given:

- Cluster 2 represents rural areas. The ATP emitted by the related areas (Figure 2.6a) is composed of two peaks with a morning peak much higher than the afternoon peak. Given that these rural areas are mostly residential and rather unattractive in terms of business or leisure activities, we can safely state that a large amount of people leave this cluster to reach working places at the morning peak and come back at the late-afternoon peak
- Cluster 5 represents urban areas. The ATP emitted by these areas (Figure 2.6b) is composed of two peaks with an afternoon peak higher than the morning one. These zones are both residential (high population density) and attractive in terms of jobs and leisure. During the morning peak, these areas generate a high number of home-work commuting trips within the cluster and to other areas, but, at the same time, attract a significant amount of demand from the surrounding areas that is supposed to leave the cluster (thus generating trips) later on, at the afternoon/evening peak.



(a) (b) (c) (d)



(e) Spatial distribution map of all sector clusters

Figure 2.6: Temporal demand profile of clusters 2 (a), 4 (b), 5 (c) and 6 (d). Spatial distribution map of all clusters (e)

- Cluster 4 can be described as a mixture of cluster 2 and 5 in the sense that these areas are mostly neither rural nor highly dense urban zones. In this case, the ATP (Figure 2.6c) is more balanced with a morning peak slightly higher than the afternoon peak (rather than the asymmetrical profile observed for cluster 2).

- The remaining clusters (1,3,6,7,8,9) have rather peculiar profiles compared to the major clusters previously discussed. For instance, this appears evident from the highly asymmetric ATP of cluster 6 (Figure 2.6d) with a major significant peak during late afternoon. This cluster depicts a particular travel behaviour of highly urban areas, which might be considered as specific cases of the observations reported for cluster 5, mostly composed of urban areas as well. Also, similar ATP has been identified for cluster 9 which is a single-sector cluster including the whole city centre of Lyon.

Our previous conclusions on the nature of each cluster (i.e., rural, urban and mixed), drawn from the interpretation of the emitted demand estimated via signaling data, have been further validated using land use data retrieved from the European CORINE land use dataset. Figure 2.7 shows the distribution of land use per cluster (i.e., percentages of cluster area covered by each specific land use). We considered 7 main categories of land use, i.e., urban, industrial, rural, sport/leisure, transport, water and other. First of all, we can observe that, for all the major clusters, rural areas cover the largest part of the cluster (even urban areas have a large rural land use proportion) due to the rather large spatial extension of the analysed sectors. However, the density of urban and industrial land use for clusters 6 and

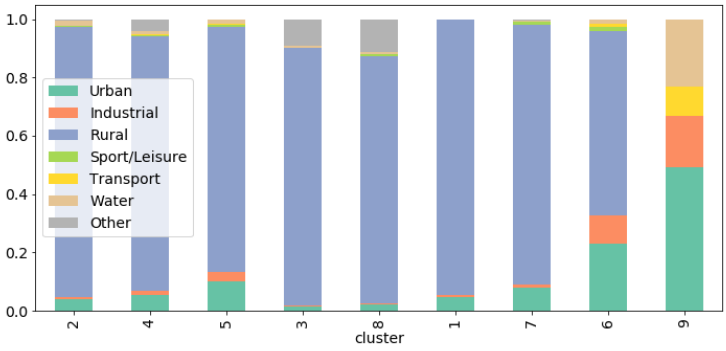


Figure 2.7: Land use coverage percentage per cluster. The clusters are sorted in ascending order of urban land use coverage: the 3 main clusters are sorted separately on the left

9 is significantly higher compared to the one of cluster 5, which is in turn significantly higher if compared to the one of cluster 4. The land use analysis appears to clearly corroborate the previously reported interpretations of the clusters and of the associated reconstructed demand.

In order to verify the resulting mobility patterns within the region, we have compared the ATPs estimated from signaling and survey data for the three main clusters 2, 5 and 4 (Figures 2.8a, 2.8c and 2.8e respectively) and the minor cluster 6 (Figure 2.8g). The patterns derived from the two sources of data well agree in terms of Pearson coefficients (between 0.89 and 0.96) shown in Figures 2.8b, 2.8d, 2.8f and 2.8h. For high urban areas (cluster 5 and 6), signaling-based estimations are slightly lower than those estimated from survey at afternoon peak, but they properly preserve the specificity of the distribution shape. For mixed (cluster 4) and rural (cluster 2) areas, cell phone-based estimations well match those from survey with slight differences at morning peak. These results confirm that signaling data can act as a good sensor and solve the sampling rate problem of surveys in large mixed and rural areas, if properly de-biased. Also, we notice that, for all clusters, signaling data provide less sharp curves in the proximity of the midday period, while survey data depicts considerable peaks centered at noon. Since the analysed flows consist of only inter-zone trips and the individual movements during the lunchtime period are more likely to be short (or very short) distance trips (i.e., intra-sector trips), the signaling data-based pattern (with rather a flat curve) surrounding this time window (11am-2pm) seems to be more relevant and realistic than the one observed via the survey. Therefore, the overall resulting observations show that the cellular signaling data can capture unknown and more reasonable flow patterns specifically for low density and large-scale areas where accurate travel data are often not available.

2.4 Discussion

For decades, traditional approaches such as travel surveys have been the major source of information for transportation planners to estimate trip flows, necessary for calibration and simulation of transport models. These travel surveys, although providing rich socio-demographic details about the respondent and his/her trips, suffer from several drawbacks such as limited sample size of involved individuals,

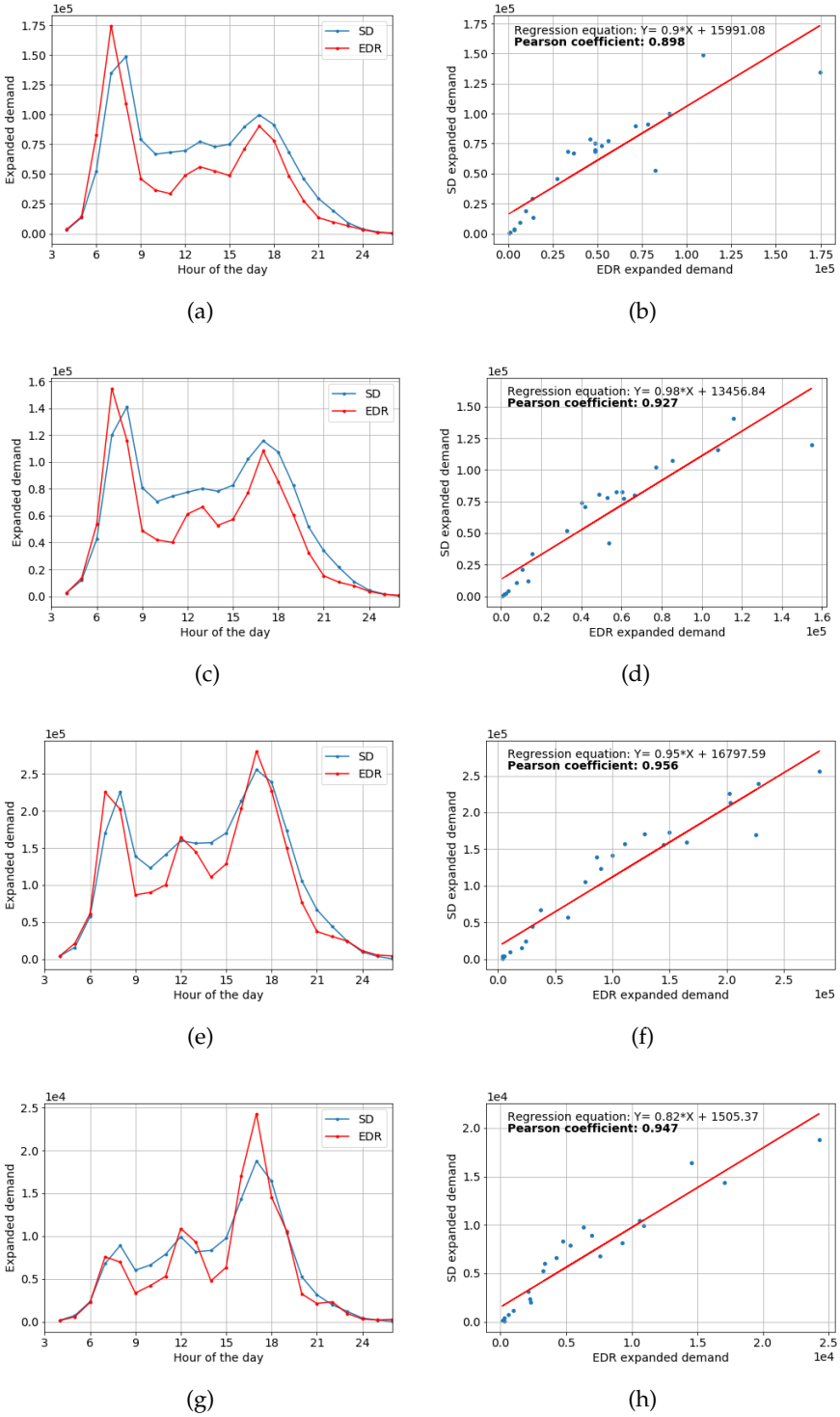


Figure 2.8: Comparison of average temporal demand profiles from signaling data (SD) and survey (EDR) and correlation between the hourly demand estimations after application of the correction for (a, b) cluster 2 (c, d) cluster 5 (e, f) cluster 4 and (g, h) cluster 6

the high deployment costs and, subsequently, the low frequency of the gathered information making them rapidly outdated and inappropriate for dynamic travel

behavior studies.

This study has demonstrated the feasibility and the potential of using cellular signaling data to track individual trips over a large-scale geographical area, to derive the travel demand by time-of-day and extracting relevant mobility spatial patterns. It thus represents an advanced step toward building a convenient framework to leverage rich mobile phone data for travel demand modeling purposes that could be used as an alternative to traditional surveys.

While promising, the obtained results have some limitations. First of all, due to the method applied to detect trips (Fekih *et al.* [33]), travel demand inference can not be done with fine spatial granularity. Indeed, small zone sectors would make the stationary activities detection even more difficult and as a result increase the spatial bias. This phenomenon has been previously discussed in Section 2.3.3. Moreover, the approach without debiasing has still significant error, the error could be reduced with more refined approach. Finally, the preprocessing approach studied in this chapter does not allow to retrieve important information about human mobility such as the destination of the trips, the trajectory of the user during his trip. The encouraging results related to the use of NSD for travel demand estimation as well as the above mentioned limitations, motivated the work done in Chapter 3.

2.5 Conclusion

This chapter introduces a framework to process cellular network signalling data for estimating accurately travel demand patterns. The framework relies on preliminary data pre-processing and filtering steps in order to only retain data that are useful for the extraction of pertinent mobility information. Then, by analysing signalling data of 2 million mobile phone users, we show that NSD is feasible to robustly extract residents' trips and estimate the hourly trip distribution throughout the studied region, on the condition that spatio-temporal biases of cell phone signalling transactions are properly detected and removed. Finally, by clustering the trip flows based on the temporal profile of the emitted demand of each zone and matching them with official land use data, we also unveil interesting and relevant heterogeneities in dynamic travel demand patterns related to trip production zones. In this study, results were obtained by exploring signalling data covering a large territory of about 44,000km2 including different socio-demographic and economic zone profiles. The evaluation performed on both the temporal and the spatial dimensions show that the resulting travel demand profiles strongly match to those obtained from travel survey data, with correlation coefficients higher than 0.9. This confirms that signalling data can be effectively exploited as a good proxy for population mobility estimations. Thus, such data should be acknowledged as a valuable cost-effective mobility data source, especially in the case of territories where accurate mobility data are not available or hard to collect via surveys or dedicated traffic probes and sensors. Moreover, we were able to identify significant correlations between mobile phone-based dynamic patterns and land use profiles. Very dense urban zones are characterized by a high afternoon peak, while low density areas depict a rather high morning peak. Besides, in the rural and mixed/suburban zones, cell network signalling data exhibit significantly higher trip flows and more reasonable patterns than survey data.

Chapter 3

Trajectory Inference at Scale

In this chapter, we present TRANSIT, TRAjectory inference from Network SIgnaling daTa, a new framework capable of processing NSD to accurately distinguish mobility phases from stationary activities for individual mobile devices, and reconstruct, at scale, fine-grained human mobility trajectories, by exploiting the inherent recurrence of human mobility and the relatively high sampling rate of NSD. The validation on a ground-truth dataset of GPS trajectories showcases the superior performance of TRANSIT (80% precision and 96% recall) with respect to state-of-the-art solutions in the identification of movement periods, as well as an average 190 m spatial accuracy in the estimation of the trajectories. We also leverage TRANSIT to process a unique large-scale NSD dataset of more than 10 millions of individuals and perform an exploratory analysis of city-wide transport mode shares, recurrent commuting paths and the construction of mobility tensor. TRANSIT aims at overcoming, at scale, the main limitations of NSD.

The chapter is structured as follows. Section 3.2 provides the problem that this chapter tackles. Section 3.3 and 3.4 present respectively the related works and the mobile phone dataset collected. Then, our framework TRANSIT as well as state of the art approaches are described in Section 3.5, followed by the validation of TRANSIT and discussion in Section 3.6. Section 3.7 studies the properties of the proposed framework and Section 3.8 presents some new human mobility applications that TRANSIT make possible. Finally, Section 3.9 and Section 3.10 present the conclusion with suggestions for future research directions.

This chapter contains parts of the article [16]:

Bonnetain L., Furno A, El Faouzi N.-E., Fiore M., Stanica R., Smoreda Z., Ziemlicki C., (2021), "TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data". In: *Transportation Research Part C: Emerging Technologies*.

Patent:

Bonnetain L., Furno A, El Faouzi N.-E., Fiore M, (2021), "Détermination de trajectoires à partir de données de téléphonie mobile". Patent number: FR2107437. France. Deposed July 8, 2021.

3.1 Notation for this chapter

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

Table 3.1: Chapter 3′ specific notations

3.2 Introduction

The preprocessing approach presented in the previous chapter as well as most of the works in the field take for granted the intrinsic limitations of mobile phone data: large spatial error, sparsity in time and oscillation effect. To overcome these limitations, they often rely on coarse aggregation along both temporal and spatial dimensions for deriving mobility information at macroscopic scale [33, 49]. These approaches seem not to leverage the full potential of mobile phone data for estimating human mobility, and, in particular the repetitive nature of human mobility with its associated consequences on a user's signalling activity.

Instead, an idea would be to reconstruct mobility information at individual level as fine as possible. Our rational is the following: if the mobility can be better estimated at individual level, the resulting aggregated mobility could be estimated at finer spatio-temporal scale. The aim of this chapter is twofold. On the one hand, we want to develop an approach able to tell apart movement intervals from stationary activity periods for each mobile device. This step is fundamental for estimating travel demand with mobile phone data and for instance, is also required before applying map-matching approach. On the other hand, for unlocking the potential of mobile phone data, we aim at developing an approach which overcomes the two main limitations of NSD: sparsity in time and large uncertainty error.

One original research direction that we have explored to solve the second problem is to exploit the repetitive nature of human mobility for improving in space and time the mobile phone trajectories. Indeed, in her mobility routine, the same individual who is performing many trips between two given locations over time, generally follows very similar paths. This creates redundancy in the mobility information that can be used to increase the spatio-temporal accuracy of the trajectories.

3.3 Literature Review

In the last two decades, CDR have been at the core of a large corpus of research related to reconstructing human mobility from large-scale passively collected data. These works have traditionally targeted the estimation of travel demand [7, 116, 18], [54], [26], the construction of signatures for automated identification of land use and urban fabrics [36, 115], the analysis of urban dynamics [83], the estimation of population density [62] and patterns discovery in human activities [38, 55]. However, despite their potential, CDR present inherent spatio-temporal biases and sparsity that have impeded their universal adoption for operational purposes related *e.g.*, to city planning and transportation. Conversely, research has flourished around the challenges aimed at improving the quality of CDR-based approaches [140] for human mobility reconstruction and modelling.

Concerning the temporal dimension, several approaches have been proposed to exploit the repetitive nature of human activities, which can be captured via a sufficiently long observation of the same user over time. The general idea is to recover information from multiple observations of the user's communication activity and thus increase the generally low frequency at which mobile phone traces are normally available. Such methods are traditionally based on machine learning techniques [24] and rely on custom spatio-temporal distances to detect trajectory similarity [71].

Regarding the spatial dimension, the geographical information associated to CDR usually comes only in the form of the coordinates of the base station to which the user is associated when a mobile phone event is issued and logged. Traditionally, the geographical area assigned to each base station is roughly determined via Voronoi or other regular (*e.g.*, grid-based) tessellations of the mobile network topology and elected as the user's position whenever an event is logged at that base station. As a result, in the most traditional case of a Voronoi tessellation, the spatial resolution of CDR only depends on the density of base stations, ranging from hundreds of meters at best in dense urban areas to several kilometers in rural ones. Another important issue affecting both the spatial and the temporal dimension of CDR is represented by the oscillation phenomenon that traditionally characterizes cellular communications [129]. Since user association in mobile networks follows operator-specific schemes based on dynamic metrics such as received signal power or base station load, oscillations can easily take place between two or more antennas, even in the absence of an actual mobility of the user. These characteristics add noise to the localization information that can be inferred from CDR data and make extremely hard the task of reliably discriminating between static and mobile sessions with CDR [60].

In the following, we focus on two approaches recently proposed in the literature to overcome location-related limitations, that represent the most related proposals to our problem. Wu *et al.* [129] propose a framework, called DECRE, to remove oscillations from CDR and reduce spatial uncertainty for enhanced human mobility modeling. To that purpose, the authors adopt a heuristic-based approach composed of three major steps, namely *detect*, *expand* and *remove*. The rational behind this approach is to remove oscillations assuming that the antennas causing oscillations are noise events that tend to reduce the spatial accuracy of the trajectory. The approach is presented in detail in Section 3.5.1.

A different strategy has been proposed in [27] and later improved by Bachir *et al.* [7], and applied to both a simple CDR dataset and a second one containing CDR enriched with location update events (a type of traffic control generated on the mobile network when a user moves over medium to long distances). Instead of filtering out the oscillations directly, the authors argue that these oscillations can be used to infer with increased accuracy user locations, by assuming that, if oscillations occur, the user should be, by triangulation, in the barycenter of these oscillation antennas. The approach, named *Cumulative Weighted Moving Average* (CWMA), consists in smoothing each mobile phone position by computing a weighted barycenter of all the consecutive antennas the user connects to within a given time-window. In particular, Bachir *et al.* [7] exploit the CWMA technique to segment the sequence of mobile phone events generated by a given user into a set of mobile and static sessions. This approach is presented in detail in Section 3.5.2.

Other approaches, tested on small samples of mobile phone data, aim at reducing the spatial inaccuracy by relying on map-matching methods [6, 15]. These methods match sequences of mobile phone events from the operator network to the nodes and edges of the transportation one, by relying on hidden Markov modeling. Despite promising results in terms of spatial accuracy, the computation time for processing a single mobile phone trace in urban environments with a dense transportation network is extremely high, thus making these approaches hard to scale to city-wide populations of mobile phone users. Some solutions, *e.g.*, [116], manage to assign trips extracted from CDR to the transportation network at scale, but require external information and assumptions, such as a route choice model. In addition, these

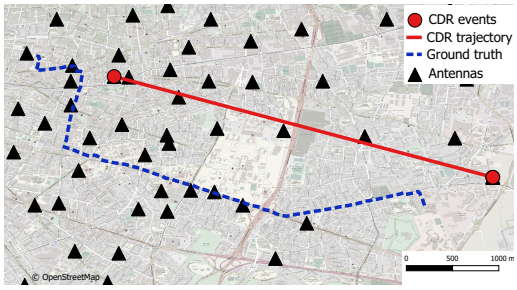
approaches can be used in combination with preprocessing approaches aiming at improving the spatiotemproal accuracy of the mobile phone trajectories.

Some types of network signaling data have also been used in the literature. Ahas *et al.* [1] use Positium, an active data collection tool, which allows them to control the temporal granularity in their dataset. However, such tools are not commonly deployed by network operators and they are more intrusive from a privacy point of view than passive approaches that simply log the user activity. Janecek *et al.* [53] use handover and location area update information for travel time estimation and map matching on the highway network. However, all these approaches tend to exhibit low performance in urban environments with a dense road network, due to the large set of similar alternative paths and low-resolution of the spatial information that is directly derived from the location of the antennas in the cellular network. Leontiadis *et al.* [70] achieve better results, but on a small mobile network signaling data, recorded by a smartphone application on a few tens of users. Recent studies included information regarding the user data connections [139] or even information regarding the increasingly popular machine-type communications [85], but without focusing explicitly on human-centric mobility. Zhao *et al.* [138] use large scale Internet access data and propose a machine learning approach to detect public/private transportation mode.

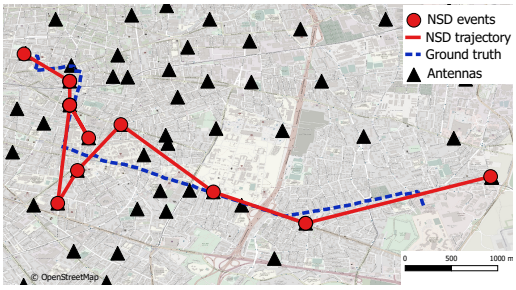
Very recently, some authors have started harnessing the potential of large-scale Network Signaling Data for different purposes. Qin *et al.* use NSD for sensing traffic conditions in urban networks [91] and making individual cellular usage prediction [90]. In [137], Zhao *et al.* compared different mathematical-based human mobility models from the literature by using NSD as ground truth. Such a study allows to improve the understanding of human mobility as well as providing tools for the simulation of mobility at both individual and population levels. In the literature, there are almost no works that aim at solving the spatiotemporal limitation of NSD. The only exception is the recent work by Song *et al.* [104]. The authors propose MIFF (Multi-Information Fusion Framework), a tool that leverages similar mobility patterns of individuals as a preliminary step before performing a map-matching of NSD to derive personal trajectories. There are some main limitations in this work. The approach has not been applied on large scale NSD. Thus, the scalability of MIFF is not demonstrated yet. They do not propose any trajectory segmentation approach as we aim to do in this chapter. Finally, as many works in the field they applied noise filtering as [129]. Instead, our idea which is original and counter intuitive is to use oscillation as triangulation for improving the spatial accuracy of the NSD trajectories. The rational behind our approach is presented in detail in Section 3.5.3.

3.4 Data Collection

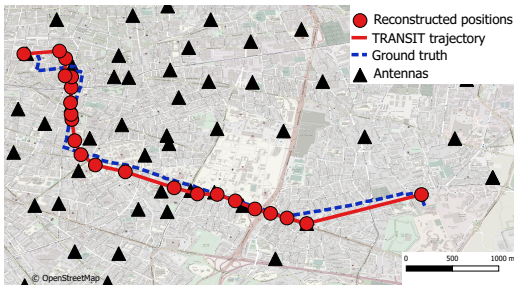
The Network Signaling Data (NSD) used in our study were collected in the production infrastructure of Orange, a leading mobile operator internationally and the largest telecommunications provider in France. The content of NSD have already been presented in Section 1.3 from Chapter 1. We next present the data collection process, in Section 3.4.1, and then investigate their statistical properties, in Section 3.4.2. We provide a comparison of NSD against other mobile network data sources in order to contextualize our framework, and broaden the understanding of NSD, whose adoption is still at early stages.



(a) Trajectory from Call Detail Record



(b) Trajectory from Network Signaling Data



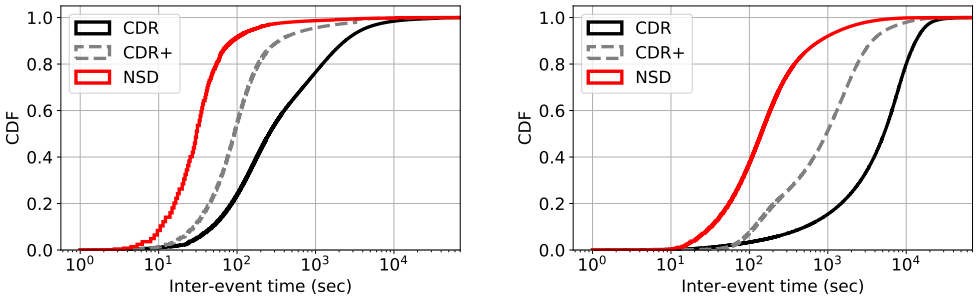
(c) Trajectory inferred with TRANSIT

Figure 3.1: Examples of inference of one trajectory of a volunteer from (a) CDR, (b) NSD, and (c) our NSD-based TRANSIT approach.

3.4.1 Large-Scale Data Collection and Ethical Considerations

The NSD used in our work cover all Orange subscribers observed in two major metropolitan areas of France, *i.e.*, Paris and Lyon; in the following, the NSD datasets in the two cities are denoted by $\mathcal{D}_P$ and $\mathcal{D}_L$, respectively. The resulting total user base tallies to over 10 millions of individual mobile subscribers identifiers (IMSI) and over 3 millions of estimated residents in the two considered cities. The data were gathered during three consecutive months in 2019, from March 15^{th} to June 15^{th}, including more than 150 billions of logged events overall, observed on a mobile phone network including more than 4,600 antennas. More details on the $\mathcal{D}_P$ and $\mathcal{D}_L$ NSD datasets are reported in Table 3.2.

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



(a) Median (b) Average

Figure 3.2: CDF of inter-event times recorded in NSD, CDR, and CDR+. The plots refer to (a) median, and (b) average times per user.

3.4.2 Comparison with Other Mobile Network Data Sources

The assortment of situations **(i)–(vi)** captured by NSD is much wider than the sole call- and text-related events in **(i)**; this naturally leads to a much higher sampling frequency of the locations of devices (hence, users) over time in NSD with respect to traditional CDR. Below, we investigate the added accuracy of NSD along the temporal and spatial dimensions. NSD is a much more recent data sources investigated by the research community compared to CDR. Thus, it is important to discuss the differences between these data along the spatial and the temporal dimensions.

Temporal Accuracy

A quantitative inspection of the increased temporal accuracy of NSD is provided in Figure 3.2. The two plots present Cumulative Distribution Functions (CDF) of the time between subsequent NSD events; specifically, the distributions are computed over the (a) median and (b) mean inter-event time recorded for each device, hence they provide a fair view of the statistics across the observed population. We also report equivalent CDF obtained using other kinds of mobile network data: (*i*) CDR, which, as already mentioned, only capture voice and texting communication events in **(i)**, and (*ii*) CDR augmented with LA (Location Area) and TA (Tracking area) update events in **(iii)**, which we term CDR+. The rationale is that CDR are the most widely adopted source of data from mobile networks, whereas CDR+ have been previously used for human mobility trajectory inference in the literature [9]. We directly extrapolated CDR and CDR+ from the available NSD database, by simply retaining only the spatiotemporal samples generated by the events that are captured by such data sources (*i.e.*, types **(i)**, and **(i)+(iii)**, respectively), while filtering out the information associated to all other network event types.

The distributions in Figure 3.2 yield a number of interesting observations. NSD grants a median inter-event time below 1 minute for 90% of the users, while that figure grows to 5 minutes for CDR+ and over 30 minutes for CDR. Per-user averages that are biased by long inactivity periods highlight even more the difference between the data sources: NSD keeps averages below 15 minutes for 90% of the users, whereas CDR+ and CDR record mean inter-arrivals of up to 1 hour and 3.5 hours

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

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

Table 3.2: Statistics on the large scale network signaling data

for the same user fraction. Similar considerations hold for users with very heterogeneous levels of network activity, as the CDF remain neatly separated across the whole domain in abscissa. The conclusion is that NSD ensure a sampling rate increase of more than one order of magnitude with respect to CDR and of a factor 5 over CDR+. Importantly, these results are fairly uniform over the considered population.

Spatial Accuracy

NSD do not bring any advantage over other classes of mobile network positioning data in terms of the absolute spatial accuracy *of each location sample*. As a matter of fact, NSD, CDR, CDR+, and any other network data types, are collected on the same radio access network infrastructure: therefore, the locations used to geo-reference the events are those of a matching set of base stations to which mobile devices associate over time. To prove our point, we run experiments with ground truth GPS data collected by a small set of volunteers, described in detail later in Section 3.6. For each volunteer, we compute the distance between the location of the antenna associated to all generated network events and the corresponding GPS position at the time. Repeating the process for all CDR, CDR+ and NSD events yields very similar average distances, between 0.26 and 0.28 km, in the three cases.

However, NSD provide a much more accurate spatial representation *of the trajectory as a whole*, as a direct consequence of the increased sampling rate. This is clearly shown in plots (a) and (b) of Figure 3.1 for a single trajectory, as well as in plots (a) and (b) of Figure 3.4 for multiple trips of a same user. These figures highlight the capability of NSD to capture individual mobility patterns in a much more exhaustive way compared to CDR. The unprecedented spatiotemporal resolution of NSD is at the basis of TRANSIT.

3.4.3 Impact of the Radio Technology

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

Table 3.2 breaks down the number of unique devices observed under each technology, as well as the number of events recorded, separately reported for the two large-scale datasets related to Paris and Lyon, $\mathcal{D}_P$ and $\mathcal{D}_L$, respectively. The figures evidence how the number of users that can be monitored by the three radio access technologies is comparable, and partially overlapping. However, the sets of geo-referenced NSD collected for the monitored devices is completely different: the number of events grows by more than one order of magnitude when moving from one cellular generation to the next.

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

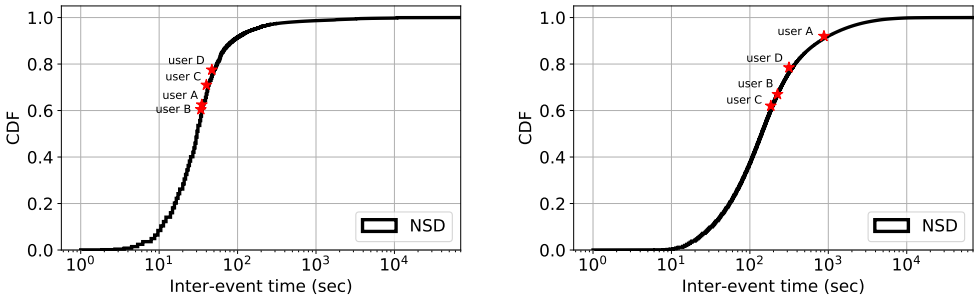
3.4.4 Small Scale Data Collection

Besides, the large scale NSD provided by Orange, we launch another experiment for overcoming the above mentioned limitations of the first ground truth dataset. The trajectory data used in our validation was collected by four Orange subscribers who voluntarily agreed to be monitored by a GPS tracking app installed on their smartphones, and who provided informed consent for their NSD to be extracted from the network operator database before pseudonymization and employed for the purpose of this research. For the sake of the experimentation, we developed a GPS tracking smartphone app called LicitGPSLogger downloadable on the play store [2]. Once gathered, all data were in any case pseudonymized, and accessed by authorized personnel of the research team only. The combined GPS and NSD data of the four users, denoted as A, B, C and D in the following, were collected during a continued period of three months, March 15 and June 15 2019, in the city of Lyon, France. We stress that size of the volunteer set, although limited, is aligned with that of state-of-the-art studies [104], with respect to which we collect a much larger number of human trajectory samples.

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

The NSD dataset, named $\mathcal{E}_{NSD}$ in the following, contains all network signaling events associated to the mobile devices of the four voluntaries, across 2G, 3G and 4G technologies. We highlight that (*i*) all volunteers were Orange subscribers at the time of the data collection campaign, and (*ii*) their were explicitly invited to maintain their regular mobile communication and service consumption habits during the measurement period. This limits biases, and we indeed observe that A, B, C and D have fairly heterogeneous profiles in the way they use mobile network services: Figure 3.3 shows that the median and average inter-event times in their NSD fall between the 60^{th} and 93^{th} percentiles of the distributions for all users in the $\mathcal{D}_P$ and $\mathcal{D}_L$ datasets that capture all subscribers in Paris and Lyon.

[2] https://play.google.com/store/apps/details?id=fr.licit.gpsloggerhl=frgl=US



(a) Median (b) Average

Figure 3.3: CDF of inter-event times recorded in NSD for the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$ (solid curve), and corresponding values for the voluntary users in the validation dataset$\mathcal{E}_{NSD}$. The plots refer to (a) median, and (b) average times per user.

Overall, the validation datasets $\mathcal{E}_{GPS}$ and $\mathcal{E}_{NSD}$ provide corresponding GPS and NSD data for over 900 hours, and encompass over 300 ground-truth trajectories of the four volunteer users. Such ground-truth trajectories were identified by first applying a recent segmentation approach for spatiotemporal GPS data [60] to $\mathcal{E}_{GPS}$, and then having the volunteers verify the resulting movement patterns via visual inspection.

3.5 Trajectory Identification, Augmentation Frameworks

3.5.1 DECRE

Wu *et al.* [129] have built an algorithm called DECRE: Detect, Expand, Check and REmove. This approach is a 4 steps heuristic based approach aiming at finding and removing oscillations in the mobile phone data. By removing outliers, the approach is able to improve to spatial accuracy of the NSD trajectories. The following concept have to be defined before describing the method.

Definition 1 (Stable period) *Given a sequence of events* $(E_1, E_2, \ldots, E_n)$ *emitted by a mobile phone ordered by datetime and* $(A_1, A_2, \ldots, A_n)$ *the corresponding sequence of antennas, the same-cell sequence are the continuous sequences of events where the antenna is unchanged, i.e,* $A_1 = A_2 = \cdots = A_n$*. The duration of the same cell sequence is the time duration from the time of the first event to the time of the last event of the same-cell sequence, i.e,* TD_{E1,E_n}*. If the time duration of a same cell sequence is long enough (e.g, longer than a threshold L), the sequence is labelled as stable period.*

First, for detecting the oscillation, 4 heuristics have been used. First, if the time difference between two stable periods is lower than a threshold T_1 (heuristic 1) then the intermediate events are labelled as *oscillations*. Then, if the time difference between an event and the last event of a stable period is lower than a threshold T_2 and the distance between these two events is higher than a threshold D_2 then the event is also labelled as *oscillations* (heuristic 2). The third heuristic, prevent the mobile

phone trace from having abnormal jumps with unrealistic speed, can be expressed as follows:

$$\begin{aligned}&(V_{A_{i-1},A_i} > V) \wedge (V_{A_i,A_{i+1}} > V)\\ \wedge &(D_{A_{i-1},A_i} > D_3) \wedge (D_{A_{i-1},A_i} > D_3) \wedge (D_{A_{i-1},A_{i+1}} > \frac{D_3}{2})\end{aligned} \tag{3.1}$$

Where V is a threshold for speed and D_3 a threshold for distance. If for the index i this condition is satisfied then the corresponding event is considered as an *oscillation*. The last heuristic is the following : if within a short period of time T_4, there are at least N_e events from at least N_a antennas then the corresponding events are labelled as *supsicious sequence*

The second step is called expand. Given a suspicious sequence, the authors expand the sequence by looking at most T_4 minutes before the suspicious sequence and at most T_4 minutes after the suspicious sequence. The look-back (or look-after) process stops when it encounters a log whose cellular tower did not appear in the suspicious sequence.

The third step aims at checking whether each suspicious sequence contains a cycle. If it does not, they remove the suspicious sequence label for the events belonging to this sequence.

The last step is remove. The events that have been labelled as *oscillation* from heuristic 1,2 and 3 are removed. Concerning the heuristic 4, for a given suspicious sequence, each cellular get a score which depends on its frequency in the sequence and its average distance to other antennas appeared in the sequence. The events corresponding to the antenna which get the highest score are kept, the others are removed.

More information about the approach can be found in Wu *et al.* [129]. We have implemented this method with the following parameters: $T_1 = 2min$, $T_2 = 1min$, $D_2 = 5km$, $D_3 = 2km$, $V = 200km/h$, $T_4 = 1min$, $N_e = 3$ and $N_a = 2$.

DECRE has several limitations. By removing events, the approach tends to improve spatial accuracy but reduce the temporal granularity of NSD trajectories. The approach does not leverage the repetitive nature of human mobility. In addition, DECRE does not contain any trajectory segmentation approach. Finally, removing oscillations is not the best strategy for improving the spatial accuracy of NSD as we show in Section 3.6.2.

3.5.2 CWMA

The approach CWMA used in [7, 27] is based on the following idea. Due to mobile network operator management, the position of the antenna does not always reflect the position of the user, the closest antenna is not automatically the one which the mobile phone user connect. The underlying assumption is that if user keeps connecting to a single antenna this is very likely that this is the closest antenna of the actual user position. Otherwise, if the device oscillates between several nearby antennas, this reveals that the real device position is probably between the oscillating cells. This filtering approach is computed as follows. Let's consider a sequence of coordinates: $(x_1, x_2, \ldots, x_n)$ of the antennas which a mobile phone has been connected

at times $(t_1, t_2, \ldots, t_n)$. The authors propose to calculate a smoothed positions of the user, denoted by $(y_1, y_2, \ldots, y_n)$ as following :

$$y_i = \sum_{j \in B_\delta(i)} w(j) \cdot x(j) \tag{3.2}$$

With $B_\delta(i) = \{j, |t_j - t_i| < \delta\}$, where $B_\delta(i)$ denotes the indices of the events emitted the mobile phone within a maximum interval δ from the current time t_i of the event. The weight w_i is computed as follows :

$$w_i = 1 - \frac{|t_j - t_i|}{\delta} \tag{3.3}$$

Where i is the index of the index which the coordinate is smoothed.

This method has been implemented by Bachir *et al.* [7] with $\delta = 8min$. It is this approach with parameter settings used by Bachir *et al.* that we have implemented in this study.

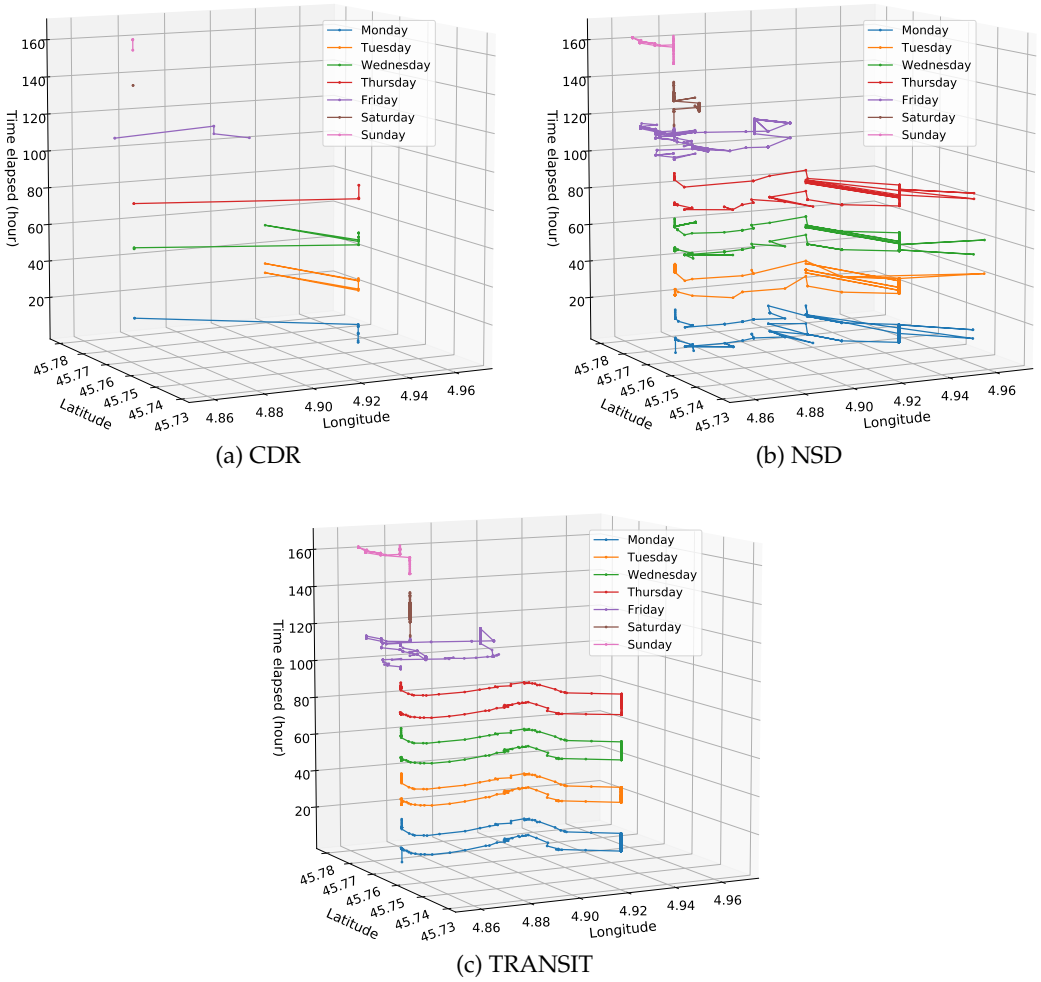
However, CWMA has two main limitations: firstly, the moving average smoothing tends to excessively distort the reconstructed trajectories; secondly, both of the approaches proposed in [7, 27] do not take into account the existence of high regularity in human movements, and consequently in mobile phone events.

3.5.3 TRANSIT

In this section we present the approach we develop, *i.e.*, TRANSIT. The performance of the approach will be compared to above mentioned frameworks in Section 3.6.1 and 3.6.2. The rationale behind TRANSIT is to leverage the inherent *regularity* of individual mobility, in combination with the high temporal resolution of NSD, to reconstruct the fine-grained mobility of individuals in urban areas. Previous works have already identified the high regularity that characterizes human movements [102, 98], and possibly used it to help coarse mobility inference at, *e.g.*, hourly resolution [23]. Indeed, regularity is already visible with the CDR employed in such earlier studies, as exemplified by Figure 3.4a. Yet, NSD provide a much more accurate perception of individual movement regularity, as illustrated in Figure 3.4b, which TRANSIT takes advantage of.

Our framework receives as input the set of NSD events of a mobile device i denoted by $\mathcal{T}^i = \{e_1^i, \ldots, e_n^i, \ldots, e_{N_i}^i\}$, where e_n^i is the n^{th} NSD event recorded for device i. Each NSD event is the result of a communication activity between a mobile device and a base station antenna of the telecommunication network, across all 2G, 3G and 4G technologies; it is defined as a tuple $e_n^i = (c_n^i, t_n^i)$, where c_n^i is the antenna at location l_n^i that handled the network event, and t_n^i is the timestamp of the instant at which the event was recorded. The NSD events in a mobile phone trace $\mathcal{T}^i$ are ordered by their timestamps t_n^i, and N_i denotes the number of events for device i. Then, TRANSIT processes $\mathcal{T}^i$ to produce two outputs in succession, as follows.

- *Trajectory identification*. The framework labels each NSD event $e_n^i \in \mathcal{T}^i$ as either static, if the user i is deemed to be engaged in an activity at a same location at the event time t_n^i, or mobile, if i is performing a movement at t_n^i. The labeling factually allows telling apart the continuous time intervals during which an individual is moving or not, and building a set $\mathcal{A}^i$ of *static activity sessions* and a



(a) CDR

(b) NSD

(c) TRANSIT

Figure 3.4: Sample weekly trajectories of one voluntary user inferred from CDR: (a), NSD: (b) and TRANSIT: (c).

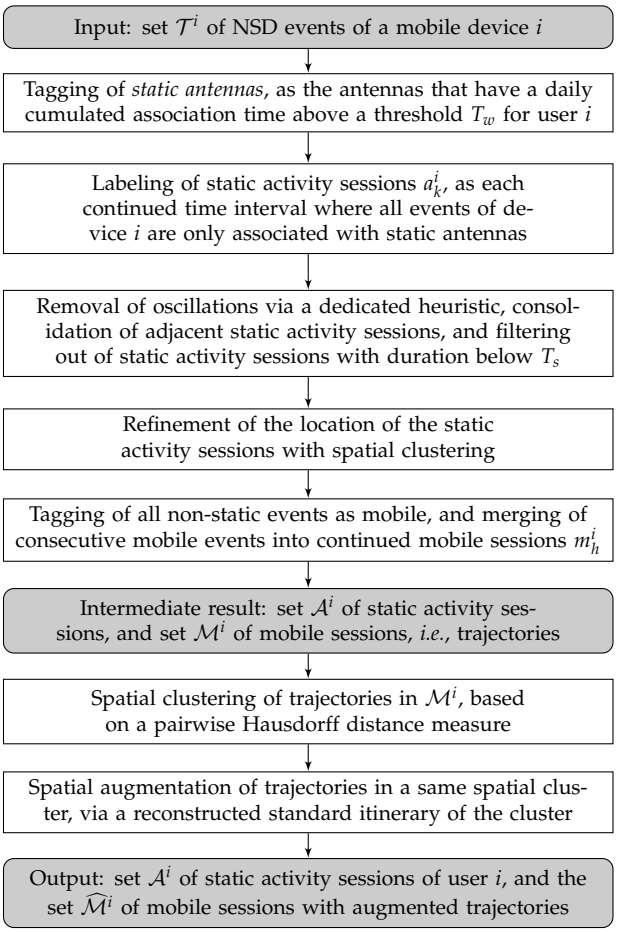


Figure 3.5: Flowchart of TRANSIT

set $\mathcal{M}^i$ of *mobile sessions*. As a result, the set $\mathcal{M}^i$ also identifies all the *trajectories*, *i.e.*, continued sequences of movement in time, of user i.

- *Trajectory augmentation*. The framework enhances the trajectories associated to mobile sessions in $\mathcal{M}^i$, by exploiting the fact that the same individual typically performs many trips between two given locations over time, generally following very similar paths. This creates redundancy in the mobility information that can be used to increase the spatiotemporal accuracy of the trajectories, as shown in Figure 3.4c. The resulting set of mobile sessions possibly augmented trajectories is denoted as $\widehat{\mathcal{M}}^i$.

Ultimately, the output of TRANSIT are the set $\mathcal{A}^i$ of static activity sessions of user i, and the set $\widehat{\mathcal{M}}^i$ of mobile sessions with augmented trajectories. Table 3.1 summarizes our notation, and Figure 3.5 presents a flowchart of the stages of TRANSIT.

Trajectory Identification

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

Figure 3.6 illustrates the process of trajectory identification using TRANSIT. The interpolation of NSD events is portrayed as the black solid line. Figure 3.6a refers

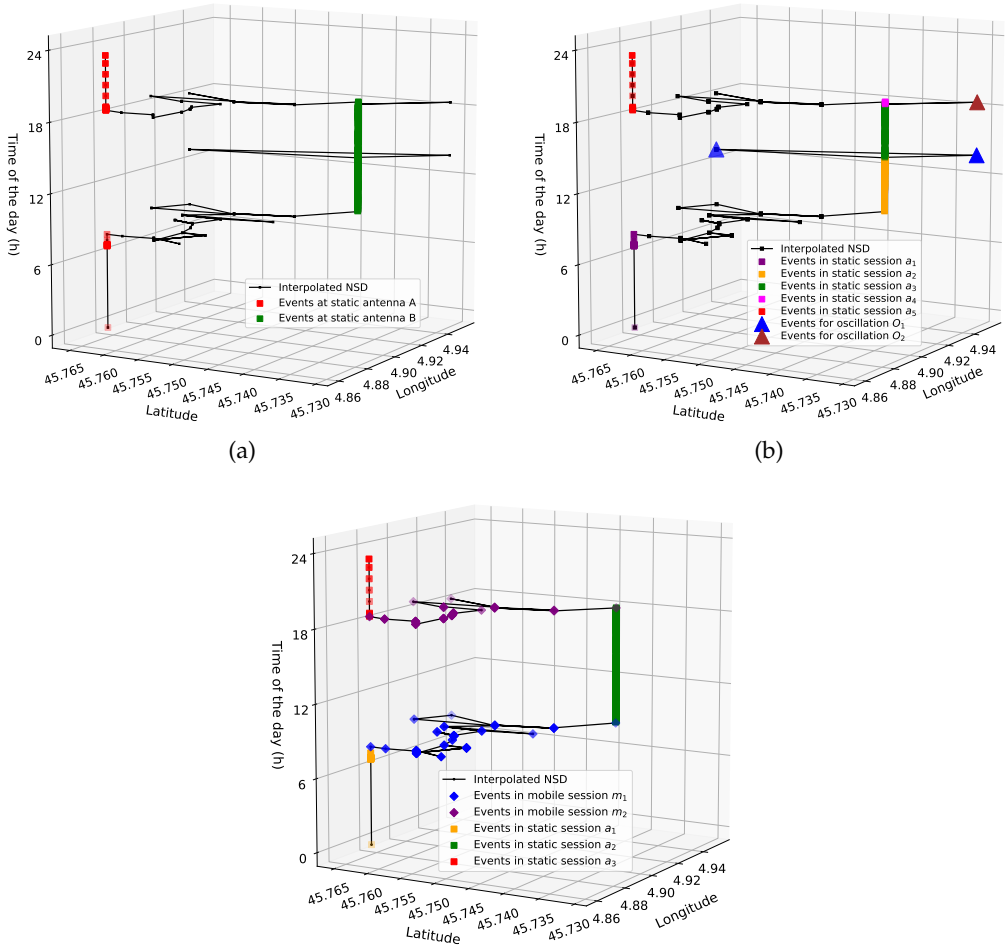


Figure 3.6: Main steps of the trajectory identification via TRANSIT.

to static antennas with daily accumulated association time above T_w. Figure 3.6b identifies static activity sessions as obtained from consecutive sequences of static antennas only, and detected oscillations. Figure 3.6c exhibits the final static activity sessions upon removal of oscillations, as well as the consequent detected mobile sessions.

We start by assuming that the time spent by user i at the antenna c_n^i associated to event e_n^i is $t_{n+1}^i - t_n^i$, *i.e.*, the temporal span to the subsequent event e_{n+1}^i. Given the high temporal resolution of NSD, this simple approach already provides a very good estimation of the time the user is associated to a given antenna, at a low computational cost. Then, a preliminary labeling is performed to trim down candidate static events. To this end, we calculate the cumulated time spent by user i at each antenna c_n^i, on a daily basis. As devices stay connected to a limited set of antennas while still, we expect such antennas to yield a non-negligible cumulated time during the target day. We thus tag as *static antennas* for user i those antennas with a daily cumulated time above a threshold T_w. In our experiments, we set T_w to 20 minutes, which falls within the range of commonly accepted values for the typical minimum duration of a significant activity carried out by an individual at a same location [33,

55], and is employed also with high-frequency longitudinal (*e.g.*, GPS) data [60]. An example is provided in Figure 3.6a.

A continued time interval where all events of device i are only associated with static antennas is then denoted as a static activity session a_k^i. The set of all such sessions across the whole observation period is $\mathcal{A}^i = \{a_1^i, \ldots, a_{K_i}^i\}$.

Typically, during one day, a user can have several static sessions, and each can be composed of one or multiple antennas.

After the stage above, only part of the antennas are labeled. Unlabeled antennas are either encountered during movements, or the result of oscillations that are known to characterize mobile device association to the radio access infrastructure [60]. Oscillations can in fact affect both static and mobile users. In the former case, they can cause the separation of continuous static activities into different static sessions in $\mathcal{A}^i$ interleaved by non-static antennas. In order to address the issue, and remove oscillations from $\mathcal{A}^i$, TRANSIT adopts the following heuristic. If (*i*) two consecutive static sessions a_k^i and a_{k+1}^i present at least one common (static) antenna, and (*ii*) the number of unique antennas associated to events observed after a_k^i and before a_{k+1}^i is below a threshold N_o, we merge all the events in a_k^i and a_{k+1}^i into a new, single static session. The new sessions replaces the former pair in $\mathcal{A}^i$. An example of oscillation detection and static sessions before the merging process is shown in Fig 3.6b.

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

TRANSIT also enforces consistency in the locations of events associated to static activity sessions, as follows. First, we compute the centroid of the locations l_n^i of all events in each session a_k^i; then, the well-known DBSCAN clustering algorithm[3] is run on the centroids of all $a_k^i \in \mathcal{A}^i$. This lets us group together all static sessions related to a same activity, and compute a consolidated location for the activity as the barycenter of all centroids in a same cluster. The locations l_n^i of all events in each session a_k^i are then replaced with the barycenter of the corresponding cluster. Note that the position of the static activity sessions that are labeled as outliers by the DBSCAN algorithm are left unchanged. An example of the resulting $\mathcal{A}^i$ is in Fig 3.6c.

Finally, all events that have not been labeled as static are labeled as mobile. This directly identifies the mobile sessions m_h^i of user i, as the time-continuous sequences of mobile events; an important remark is that the two static events immediately preceding and following the mobile session are also integrated into m_h^i. As a result, the set of mobile sessions is $\mathcal{M}^i = \{m_1^i, \ldots, m_{H_i}^i\}$. Each m_h^i corresponds to one trajectory of user i identified by TRANSIT. An example is also in Fig 3.6c.

[3]The parametrization of DBSCAN for static session clustering leverages is discussed later in Section 3.6.3.

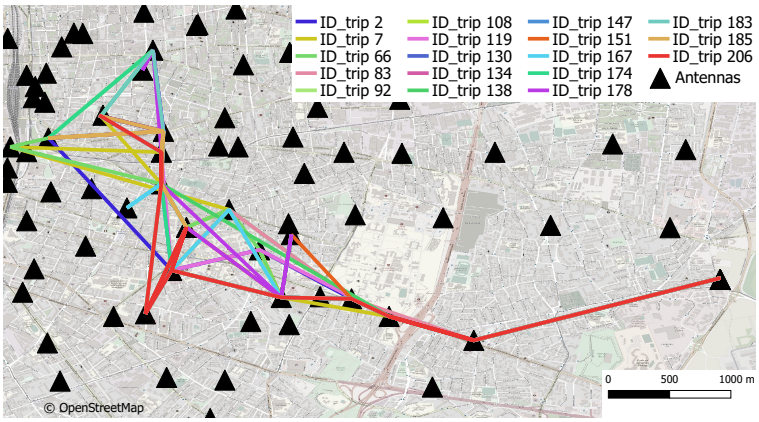


Figure 3.7: Set of trajectories of a voluntary user clustered by DBSCAN, for the Origin-Destination path in Figure 3.1.

Trajectory Augmentation

The sequences of NSD events in $\mathcal{T}^i$ that correspond to the single trajectories m_h^i of user i are still affected by the limited spatial accuracy of mobile network data, which affects NSD as explained in Section 3.4.2. In its second phase, TRANSIT thus aims at improving the geographical correctness of the movement information. As anticipated, the framework relies on the regularity of human mobility; more precisely, we use the the information from multiple similar trajectories identified for a same user to mutually improve their accuracy.

As a first step, a similarity measure is computed for all pairs of mobile session $m_h^i \in \mathcal{M}^i$. We employ the Hausdorff distance [109], which is defined as:

$$\begin{aligned} & d_H(m_{h_1}^i, m_{h_2}^i) = \max\{D(m_{h_1}^i, m_{h_2}^i), D(m_{h_2}^i, m_{h_1}^i)\}, \\ & \text{where} \quad D(m_{h_1}^i, m_{h_2}^i) = \sup_{l_{n_1}^i \in m_{h_1}^i} \inf_{l_{n_2}^i \in m_{h_2}^i} d(l_{n_1}^i, l_{n_2}^i), \end{aligned} \tag{3.4}$$

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

Then, DBSCAN is applied[4] to the distance matrix, in order to group trajectories that have similar spatial geometries, and correspond to diverse trips of the user between the same two static activity locations. Figure 3.7 shows an example of a set of mobile sessions, *i.e.*, trajectories, grouped together in the same cluster by DBSCAN, for the origin-destination activity locations in Figure 3.1. Based on the result of DBSCAN, we can tell apart the mobile sessions in $\mathcal{M}^i$ into two subsets: (*i*) trajectories that fall into a cluster, *i.e.*, which refer to a path that is recurrent in the mobility of user i, and which we denote as the set $\mathcal{M}_R^i$; and, (*ii*) outlier trajectories that represent unique movements of i, which are grouped in set $\mathcal{M}_O^i = \mathcal{M}^i \setminus \mathcal{M}_R^i$.

For trajectories in $\mathcal{M}_R^i$, TRANSIT operates a spatial augmentation, as follows. First, the average duration is computed for all trajectories assigned to a same spatial cluster by DBSCAN above; this corresponds to the expected time that user i takes to travel between the same origin-destination activity locations. The time information

[4]The parametrization of DBSCAN for mobile session clustering is discussed later in Section 3.6.3.

is used to filter out trajectories whose duration deviates from the median by 50% or more: these mobile sessions are considered not representative of the routine mobility patterns along the target path. The retained trajectories in a same cluster are then temporally scaled (*i.e.*, stretched or compressed) in time so as to match the average travel duration for the cluster. Finally, the scaled trajectories are temporarily binned according to a fixed time period of one minute, and the spatial coordinates of all different events that fall in a same time bin are averaged.

The previous steps lead to a set of positions, one per minute, which represent the reconstructed itinerary. If there is no event within a particular time slot, the resulting enhanced trajectory will have missing positions. All trajectories in the cluster are then matched to the reconstructed one, and become thus identical in the space dimension. However, they are re-conducted to their original duration (*i.e.*, via compression or stretching) so as to keep them faithful to their recorded travel time in the NSD.

As a result, each original mobile sessions in $\mathcal{M}_R^i$ is replaced by a set of reconstructed positions without any temporal deformation, and is enriched with information derived from multiple similar trajectories traveled by the same user. This set of enhanced mobile sessions is referred as $\widehat{\mathcal{M}_R^i}$. We recall that Figure 3.1c shows the final spatial trajectory inferred from the cluster in Figure 3.7. Trajectories in $\mathcal{M}_O^i$ stay instead unchanged, corresponding to those obtained from the simple interpolation of NSD data. The final set of mobile sessions is $\widehat{\mathcal{M}^i} = \widehat{\mathcal{M}_R^i} \cup \mathcal{M}_O^i$.

3.6 TRANSIT Validation

3.6.1 Trajectory Segmentation

We first assess the performance of TRANSIT in identifying trajectories, by separating the static activity sessions and mobile sessions of a user. To this end, we compare the sessions identified by our approach applied on $\mathcal{E}_{NSD}$ against the ground truth extracted from $\mathcal{E}_{GPS}$. We also include in our analysis one recent benchmark from the literature, *i.e.*, the CWMA approach [27, 7] respectively presented in Section 3.5.1 and in Section 3.5.2. We use classical *precision*, *recall* and *F1* metrics to evaluate the performance of the trajectory segmentation approaches. Formally:

$$\text{Precision} = \frac{TP}{(TP + FP)}, \quad \text{Recall} = \frac{TP}{(TP + FN)} \tag{3.5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{3.6}$$

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

Overall and per-user results are summarized in Table 3.3. Both TRANSIT and CWMA attain rather high values of precision and recall, typically in the 75–100%

Table 3.3: Performance evaluation results for the trajectory identification task. The second and third column report the temporal span of the combined GPS and NSD data, and the number of ground-truth trajectories, respectively. Best values are highlighted in bold

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

range. For users with enough trajectories, *i.e.*, B, C and D, this leads to F1 scores between 0.8 and 0.9. However, the session classification approach of TRANSIT performs consistently better, yielding a 5% relative improvement in the total F1 score with resect to CWMA.

A closer inspection reveals how CWMA tends to yield higher precision than recall, *i.e.*, to incorrectly label static events as mobile. We ascribe the problem to the oscillation phenomenon discussed in Section 3.5.3: CWMA lacks a tool to remove oscillations occurring during static activity phases, hence tags such events as movements and overestimates the incidence of mobile events. As a by-product, CWMA detects a large number of non-existent trajectories (504 against 310 in the ground truth), which are in fact network-driven changes of the antenna serving the static user.

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

3.6.2 Trajectory Enhancement

We now explore the capability of TRANSIT to improve the spatial representation of the individual trajectories identified above. We thus compare the augmented trajectories returned by our framework in $\widehat{\mathcal{M}}^i$ against the ground truth inferred from the GPS data in $\mathcal{E}_{GPS}$. We also consider a comprehensive set of benchmarks to contextualize the performance of our framework, as follows: (*i*) *DECRE/CDR* is the trajectory reconstruction method implemented by DECRE [129] as presented in Section 3.5.1 – in this case, we apply DECRE on CDR data extrapolated from NSD as explained in Section 3.4.2, as the method was originally conceived for this type of data; (*ii*) *CWMA/CDR+* is the trajectory reconstruction approach adopted by CWMA [27, 7] – here, it is applied to CDR+ data, also extracted from NSD as explained in Section 3.4.2, since these are the kind of data the approach was tested with by its authors; (*iii*) *Raw NSD* are the trajectories interpolated from the NSD directly, which is an important baseline for comparison; (*iv*) *DECRE* is the trajectory reconstruction method implemented by DECRE, run on NSD; (*v*) *CWMA* is the the trajectory reconstruction approach adopted by CWMA, run on NSD. Note that we are interested in comparing the different techniques in the specific task of trajectory augmentation: therefore, for the sake of fairness, we run TRANSIT and all benchmarks on the same set of trajectories $\mathcal{M}^i$, *i.e.*, those identified by our approach, as it provided the most accurate result in Section 3.6.1 above.

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

$$D_{GPS} = \frac{1}{|m_{GPS}|} \sum_{e_{n'} \in m_{GPS}} \min_{e_n \in m_{NSD}} d(l_{n'}, l_n) \tag{3.7}$$

$$D_{NSD} = \frac{1}{|m_{NSD}|} \sum_{e_n \in m_{NSD}} \min_{e_{n'} \in m_{GPS}} d(l_n, l_{n'}) \tag{3.8}$$

where m_{GPS} and m_{NSD} are, respectively, two trajectories inferred from GPS and mobile network data. The operator $|\cdot|$ denotes the cardinality of the argument set, *i.e.*, the number of samples in the case of a trajectory, and $d(\cdot,\cdot)$ the geodesic distance. We use both metrics as they are complementary: while D_{GPS} is representative of the error observed for continuously tracked user, D_{NSD} measures the error specific to events recorded by the mobile phone network.

The results are reported in Table 3.4, for each user and in total. Trends are clear and consistent across users: there is a neat increase of accuracy in the inferred trajectories when moving from the right to the left in the table. Clearly, using CDR and CDR+ data penalizes DECRE and CWMA in the two rightmost columns, where the average error in the trajectory locations is 680–1,000 meters for D_{GPS}, and 250–360 meters for D_{NSD}. A simple interpolation of the Raw NSD already improves the result substantially, with average errors at 380 and 260 meters, for D_{GPS} and D_{NSD}, respectively. Interestingly, DECRE cannot improve that performance, mainly because its oscillation removal process has alternating effects, and can also eliminate events that are in fact useful to reconstruct the correct itinerary. CWMA improves the average D_{NSD}, bringing it down to 160 meters, however does not affect D_{GPS}. TRANSIT achieves the best performance in nearly all situations, and attains average errors that are as low as 220 meters for D_{GPS} and 160 meters for D_{NSD}.

Overall, the relative performance in Table 3.4 prove that TRANSIT does not simply rely on the added temporal resolution of NSD to advance the current state of the art; instead, it also introduces original processing that can take full advantage of NSD. From an absolute performance viewpoint, TRANSIT sets a new bar for the quality of individual trajectories inferred from mobile network data: with errors in the order of 150 meters, it demonstrates that a tailored processing of NSD can result in positioning information that is sufficiently accurate to support mobility monitoring applications at scale. We will provide multiple examples later, in Section 3.8.

3.6.3 Parameter Setup and Implementation Settings

TRANSIT requires the setting of five tunable parameters (*i.e.*, T_w, T_s, N_o, D_s and D_m), reported in the notation table of Table 3.1.

Table 3.4: Performance evaluation results for the trajectory augmentation task. Numbers represent the mean plus/minus the standard deviation, expressed in kilometers. Best values are highlighted in bold.

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

Sensitivity Analysis on D_s, D_m and N_o

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

Finally, we report in Figure 3.8 the results of the sensitivity analysis performed to determine the two distance thresholds D_s and D_m, used as the maximum distance allowed in DBSCAN cluster for both location enhancement of static sessions and for the identification of similar mobile sessions. As performance criterion of the analysis, we have used the average of the two distance metrics D_{GPS} and D_{NSD} described in Equation 3.8. For each configuration of the parameters D_s and D_m in the ranges reported in Figure 3.8, we obtained a different set of $\widehat{\mathcal{M}}_R^i$ for all users in $\mathcal{E}_{NSD}$ and computed the corresponding value of our performance metric. The figure highlights that the selected performance criterion attains its minimum value (*i.e.*, better reconstruction of the real trace) when $D_s = 0.15Km$ and $D_m = 2.5Km$. These values have been thus selected as the default values for the two parameters.

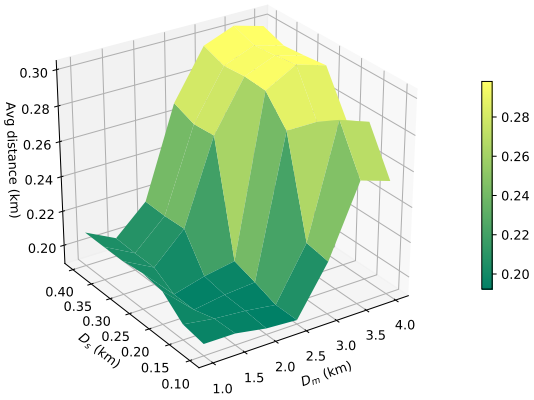


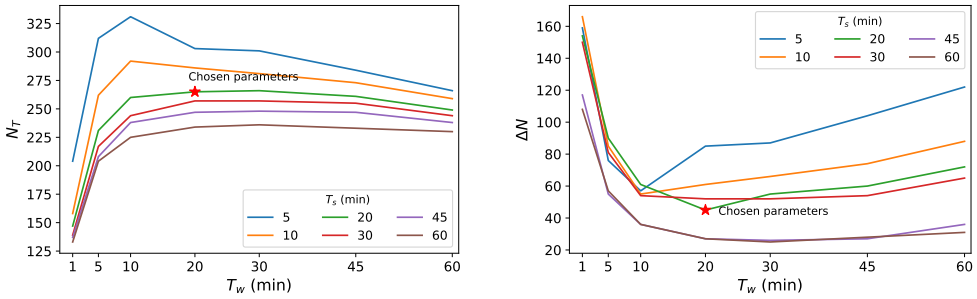
Figure 3.8: Parameter D_m and D_s sensitivity on trajectory enhancement performance.

Sensitivity analysis on T_s and T_w

The criterion that has been considered for the selection of the T_s and T_w thresholds is twofold. Firstly, to determine T_s we consider the number of mobile sessions (i.e., non-enhanced trips) detected by TRANSIT on our validation dataset $\mathcal{E}_{NSD}$, namely N_T. Secondly, to determine T_w we consider the number of missed trips by TRNSIT, namely ΔN, with respect to a benchmark segmentation method for GPS data [60], used as ground truth. We recall that T_s is the minimum duration of a static session, and it is a shared parameter of the benchmark segmentation method used with GPS data and the trajectory segmentation approach of TRANSIT used with NSD. T_w is

the minimum cumulated time for an antenna to be labeled as static, and it is only related to the segmentation approach of TRANSIT.

On the one hand, figure 3.9a shows the sensitivity of N_T on T_w and T_s. We can observe that for a given T_w, N_T increases when T_s decreases. In other words, when the minimum duration of a static session T_s decreases, TRANSIT captures more trips, i.e., trips taking place between shorter stationary activities, which can correspond to *e.g.*, leisure stops, public transport connections and modal shifts, taking children at school, stops at traffic lights, etc. Therefore, T_s has to be chosen accordingly to considerations that are specific to the kind of mobility analyses one might want to study. With that regard, the nature of the trips that we aim to reconstruct and enhance with TRANSIT via NSD is mainly related to recurrent itineraries linked to different kinds of transport motifs, which do not normally include trips between very short stationary activities. For this reason, we set T_s to 20 minutes, which allows detecting a fairly high number of trips (as from figure 3.9a) and is also a typical reference values from state-of-the-art literature on stationary activity detection via mobile phone and GPS data for recurrent mobility analyses [33], [55], [60]. On the other hand, figure3.9b shows the sensitivity of ΔN on T_w and T_s. In particular, the number of missed trips ΔN is minimized by larger values of T_w as larger values of T_s are considered. Specifically, the value of T_w that minimizes the ΔN error is 10 minutes for $T_s = 5$ minutes, 20 minutes for $T_s = 20$ minutes, 30 minutes for $T_s = 60$ minutes, etc. Thus, given our choice of T_s equal to 20 minutes, the value of T_w has been set to 20 minutes in order to minimize the number of trips missed by TRANSIT with respect to the adopted benchmark method.



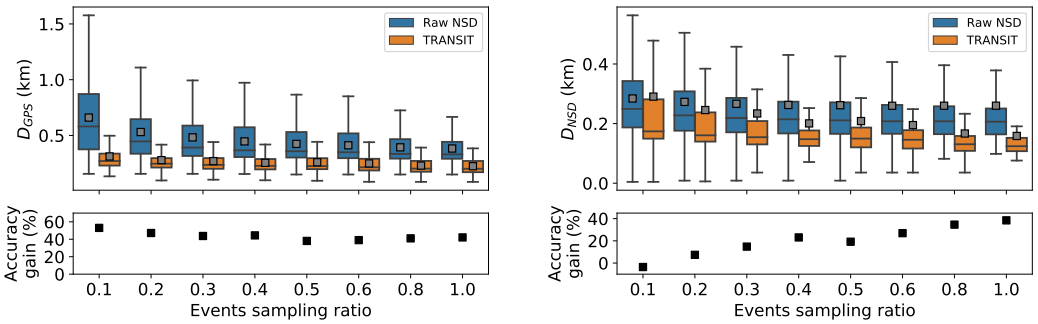
(a) Number of trips detected by TRANSIT

(b) Number of trips difference between GPS and TRANSIT

Figure 3.9: Sensitivity of TRANSIT on T_w and T_s in terms of: (a) volume of trips detected by TRANSIT; (b) error in the number of detected trips using [60] applied to $\mathcal{E}_{GPS}$ and TRANSIT applied to $\mathcal{E}_{NSD}$

Implementation Settings

TRANSIT has been implemented in PySpark and run on a Spark cluster deployed at the mobile network provider's facilities. The Spark execution environment consists in 50 executors, each configured with 4 cores and 28 Gigabytes of memory. All the main algorithmic components of TRANSIT from Figure 3.1 have been implemented via PySpark User-Defined Functions (UDF) and applied in a distributed manner to the whole sets of subscribers' network signaling traces considered in our analyses.



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

Figure 3.10: Analysis of the performance of TRANSIT versus the ratio of NSD events retained by subsampling, for the (a) D_{GPS} and (b) D_{NSD} distance metrics.

Specific optimizations have been required in order to process the three months of NSD from the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$.

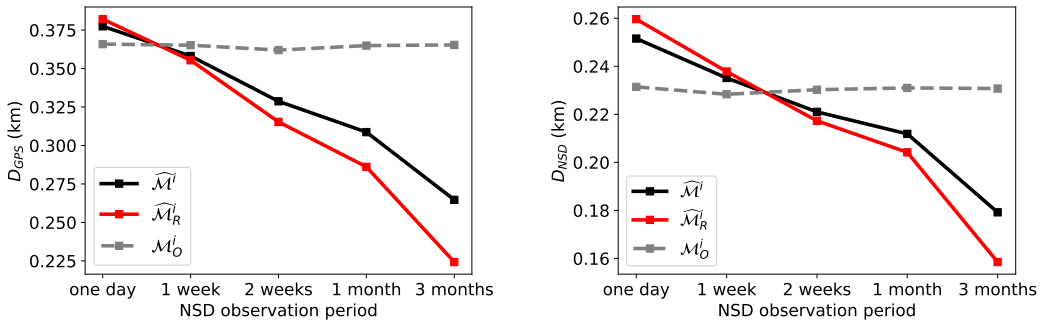
Among the different optimizations, a special attention was dedicated to the computation of the pair-wise Hausdorff distance matrix, which represents the most time-consuming step of our approach (taking approximately 70% of the total computation time). Specifically, we avoid computing the Hausdorff distance for all pairs of trajectories having different origin and/or destination. In such case, we set their distance to a value larger than the D_m parameter, thus making it impossible for DBSCAN to cluster them together. Similarly, the Hausdorff distance is immediately limited to D_m when a value larger than D_m is found during the iterative computation of the inner distances $D(.,.)$ from Eq. 3.4. It is worth to note that this simple optimization allows us saving significant computation time, as well as keeping the result of the clustering unchanged.

3.7 TRANSIT Properties

3.7.1 Impact of Sampling Rate

We investigate further the settings that help TRANSIT achieve such a remarkable result in terms of accuracy of the inferred trajectories. As a first step, we consider the impact of the spatiotemporal sparsity of the NSD that is fed to TRANSIT. We do this by randomly subsampling the NSD of each user $i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ down to a fraction of original mobile events in every sessions in $\widehat{\mathcal{M}}^i$; we then run the trajectory augmentation method of TRANSIT on the sparser trajectories. Due to the stochastic nature of the subsampling, we averaged the metrics D_{GPS} and D_{NSD} over 10 trials for each distinct sampling ratio.

Figure 3.10 shows the results. When looking at D_{GPS}, the impact of the sampling ratio on TRANSIT performance is marginal, even when retaining as little as 10% of the NSD events. The relative gain in term of spatial accuracy of TRANSIT compared to trajectories obtained from a naive interpolation of the Raw NSD grows from 40% to 60%, as the latter are obviously negatively impacted by a reduced NSD sampling frequency. Concerning D_{NSD}, the trend is different. Indeed, the average value of



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

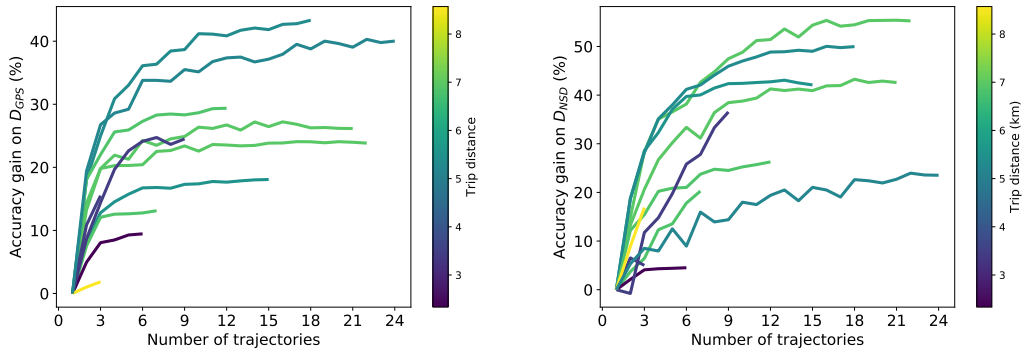
Figure 3.11: Analysis of the performance of TRANSIT versus the time span of the NSD data, for the (a) D_{GPS} and (b) D_{NSD} distance metrics. Different curves report the results for trajectories in $\widehat{\mathcal{M}}^i$, $\widehat{\mathcal{M}}^i_R$, and $\mathcal{M}^i_O$.

D_{NSD} remains constant for raw NSD, regardless the sampling frequency, whereas it decreases for TRANSIT in the case of higher sampling ratio. Indeed, there is no reason that an increased number of NSD events would improve the intrinsic spatial uncertainty of Raw NSD: as this error is linked to the geographical sparsity of the antennas, the distance between NSD and the closest GPS position stays constant at around 300 meters. However, TRANSIT decouples trajectory samples from base station locations, and can better approximate the actual position of the user by averaging over a higher number of NSD samples collected at different antennas. This lets TRANSIT increase its gain up to 40% as the sampling ratio grows.

3.7.2 Impact of Data History

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

The results are in figure 3.11. The average accuracy of all trajectories identified by TRANSIT in $\widehat{\mathcal{M}}^i$ substantially improves for longer observation periods. The errors decreases by 40% for both D_{GPS} and D_{NSD} when NSD are collected during three months rather than in a single day. Also, recall that $\widehat{\mathcal{M}}^i$ is in fact composed of trajectories that are actually augmented by TRANSIT, in the set $\widehat{\mathcal{M}}^i_R$, and trajectories that the framework could not improve due to the lack of similar movements in the user data, in the set $\mathcal{M}^i_O$. Thus, figure 3.11 also breaks down the results for these two categories. As expected, different NSD time spans do not affect the accuracy in $\mathcal{M}^i_O$. Instead, in the case of the recurrent trajectories in $\widehat{\mathcal{M}}^i_R$, a longer history of mobility helps clustering and averaging a larger number of similar mobility patterns of the user, hence reducing the natural spatial bias of the original NSD.



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

Figure 3.12: Analysis of the performance of TRANSIT versus the number of averaged trajectories per cluster in $\widehat{\mathcal{M}}_R^i$, for the (a) D_{GPS} and (b) D_{NSD} distance metrics. Different colors map to diverse geographical lengths.

3.7.3 Impact of Number of Clustered Trips

The results in figure 3.11 highlight that at least a few weeks of NSD data are needed in order for TRANSIT to be able to enrich recurrent trajectories. However, this is an artifact of the availability of additional comparable trajectories as we observe the user mobility in time. We thus decouple this phenomenon from the time dimension, and investigate how the number of clustered trajectories used to improve the spatial accuracy of NSD affects D_{GPS} and D_{NSD} directly. We conduct the following experiment: we select clusters of at least 3 trajectories, ending up with 11 clusters across all voluntary users. For these clusters, we test how the number N of trips within each cluster affects the spatial accuracy of the reconstructed itinerary. For instance, for one cluster, we select randomly N trips among all the trips within the cluster, we reconstruct the itinerary using these N trips and then compute the distance metrics. For each cluster, we are able to test N ranging from 1 to the number of trips within the cluster.

The results are shown in figure 3.12. On the ordinate, we represent the spatial accuracy gain compared to the scenario using 1 trajectory for doing the reconstruction. For D_{GPS} in figure 3.12a, we can observe that most of the curves have similar shapes, with a gain for a relatively low number of trajectories (between 2 and 6) and the emergence of a clear diminishing return effect afterwards. A similar phenomenon can be observed for D_{NSD} in figure 3.12b, albeit with less neat transition. This behavior is consistent across trajectories covering different spatial distances (colors), and achieving diverse accuracy gains (final value in the ordinate). We conclude that, at least in the set of trajectories we could study, a fairly small number of less than 10 instances of the same route is typically sufficient to achieve the maximum error reduction that TRANSIT can grant.

3.8 TRANSIT Validation and Applications

While the validation results are related to a reduced number of users, the interest of TRANSIT reveals at city-wide scales, where it can enable a number of mobility-related applications. This section analyzes three case studies related to urban mobility that leverage the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$ described in Section 3.3.

3.8.1 Comparison with Surveys

In order to validate TRANSIT, we compare the typical travel demand profile inferred by TRANSIT with those obtained with surveys as done in Chapter 2 for the area of Lyon. The studied area is a subset of the Rhônes-Alpes area described in Chapter 2 as $\mathcal{D}_L$ do not cover the whole Rhônes-Alpes area. This allows to compare TRANSIT and Fekih *et al.* approach described in Chapter 2 on the same region. The latter is composed of 16 zones and corresponds to the city of Lyon. The temporal travel demand profiles outputed by TRANSIT and those obtained with survey is depicted in Figure 3.13. For comparison purposes, we also computed the travel demand profile obtained with the methodology of Fekih *et al.* without the debiasing procedure which is also represented in the figure. The results show that the demand profile obtained by TRANSIT is better than the one obtained with Fekih *et al.* approach for inferring travel demand profile. We can observe that TRANSIT demand estimation is higher during non-peak hours demand compared to surveys. We can assume that this difference can be explained by an underestimation of the surveys which can accurately estimate the demand representing commuting trips but struggle to capture the other kind of trips. The latter results demonstrate the capability of TRANSIT to estimate accurately travel demand. Moreover, we also compare the result of TRANSIT with respect to Fekih *et al.* methodology (with and without debiasing procedure) at a finer spatial granularity. Indeed, for each zone, we computed the travel demand of the zone with TRANSIT, Fekih *et al.* approach and relying on the surveys. 2 metrics are used to compared the TRANSIT and Fekih *et al.* approach. The first one is the Pearson correlation between the temporal demand per zone inferred by the considered approach and the one obtained with surveys. This Pearson correlation is computed for each zone and then averaged for all zones. The average Pearson correlation is denoted as P_{corr}. The second one is the Root Mean Square Error between the temporal demand per zone inferred by the approaches and the one obtained with surveys. This RMSE is computed for each zone and then averaged for all zones. The average RMSE is denoted as $RMSE$. The results are given Table 3.5. TRANSIT has P_{corr} equal to 0.89 instead of 0.81 for Fekih *et al.* without debiasing. Besides, TRANSIT has $RMSE$ equal to 2080 instead of 2900 for Fekih *et al.* without debiasing. The results show that TRANSIT outperforms Fekih *et al.* methodology without debiasing with respect to both metrics. When compared with Fekih *et al.* approach with debiasing, TRANSIT obtained equivalent performance on P_{corr} and $RMSE$. However, TRANSIT do not need to rely on debiasing procedure to compute accurately travel demand profile. Finally, contrary to state of the art approaches, TRANSIT is able to infer travel demand at finer spatio-temporal scale. This point will be discussed in Chapter 5.

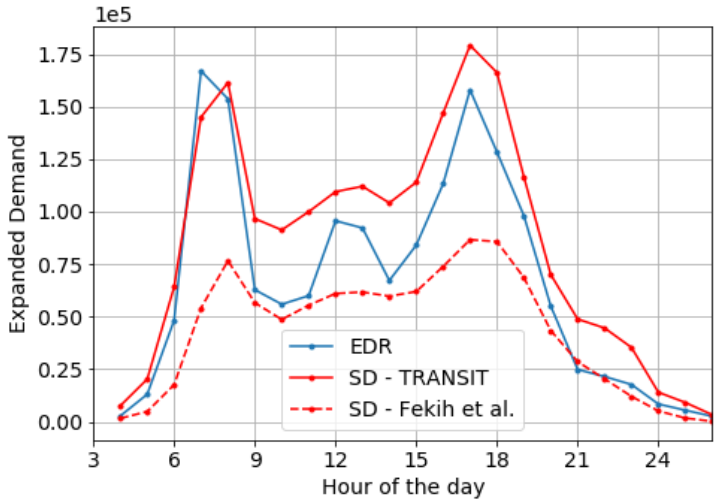


Figure 3.13: Temporal demand profile (number of hourly trips generated from all zones) from TRANSIT, Fekih *et al.* approach and survey data (EDR).

Measure	TRANSIT	Fekih et al. with debiasing	Fekih et al.
P_{corr}	0.89 ± 0.06	0.87 ± 0.07	0.81 ± 0.09
$RMSE$	2080 ± 556	2008 ± 715	2900 ± 1263

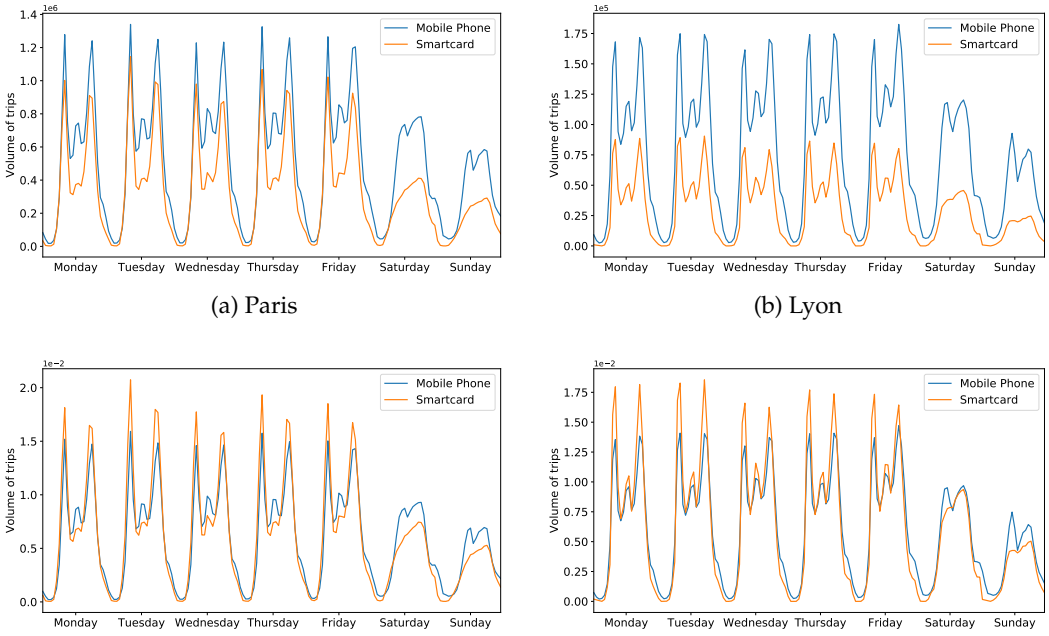
Table 3.5: Evaluation of travel demand profiles inference per zone

3.8.2 Urban Mobility and Public Transport

By counting the number of concurrent active trips inferred via TRANSIT over time, we are able to reconstruct accurate temporal profiles of the travel demand in urban regions. For such profiles to be dimensionally correct, a rescaling is needed to account for the penetration rate of the technology (close to 100% in developed countries like France) and the market share of Orange (at 37% over the French territory). The resulting average weekly demand profiles computed in Paris and Lyon are depicted in blue in Figure 3.14a and Figure 3.14b, respectively. Our estimates are that around 1,300,000 individual trips occur at the same time in Paris during commuting peaks, while the figure is at 180,000 for Lyon.

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

Also in this case, a rescaling is required: while the TRANSIT trajectories refer to the resident population, the smart card data include both residents and non-residents. In order to make the numbers comparable, we apply a scaling factor of 0.81 to the smart card temporal profile; the factor has been calculated from the raw network signaling data, by computing the average instantaneous fraction of resident subscribers present in the target cities, over the total number of observed users. The



(a) Paris

(b) Lyon

(c) Paris, normalized

(d) Lyon, normalized

Figure 3.14: Average weekly profiles of the number of concurrent trips in (a) Paris and (b) Lyon, as inferred from TRANSIT and smart card data. Normalized versions with integral one of the same profiles are in (c) and (d).

weekly profiles from smart cards are superposed to the TRANSIT-inferred ones, as the orange curves in Figure 3.14a and Figure 3.14b.

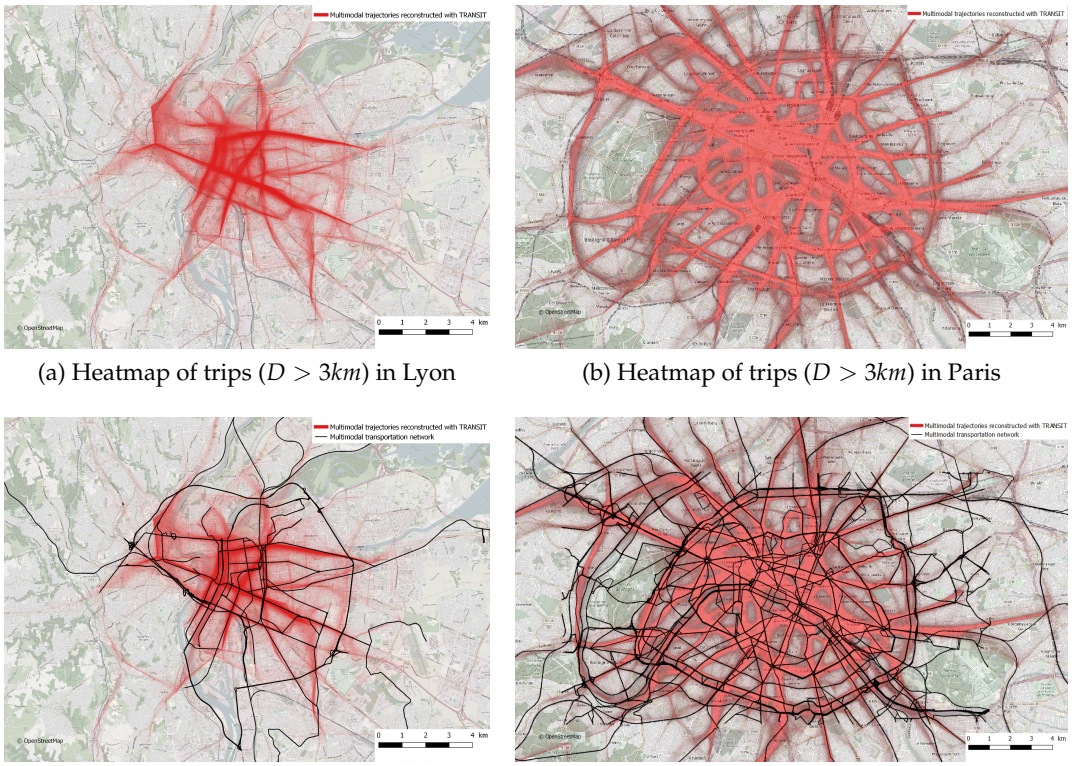
The comparison of the profiles reveals interesting facets of mobility in Paris and Lyon. Clearly, the volume of trips identified by TRANSIT is higher than that reconstructed with smart card data: NSD allows monitoring virtually all transport modes, including those beyond public means, *e.g.*, private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

In addition, we can investigate the temporal incidence of public transports by looking at versions of the same profiles that are normalized so that the integral of all curves is one. Figure 3.14c and Figure 3.14d show the result. This perspective lets us appreciate how in Paris public transports are especially important during commuting hours, but relatively less used during the lunch break or weekends. A slightly different pattern emerges in Lyon, where public transports are also very much used around midday, but have a lower incidence on total mobility during evenings and weekend mornings. We highlight that obtaining this type of insights is hardly achievable by solely relying on surveying, which demonstrates the value of NSD and a method like TRANSIT that can exploit them.

As a final remark, we highlight that the results in Figure 3.14 can also be considered as a partial validation of the trajectories inferred by TRANSIT in large-scale settings. Indeed, the near-perfect match of the timing of commuting peaks or overnight low mobility among curves proves the capability of our trajectory segmentation approach to identify trips that are very consistent with data collected in the field over time.

3.8.3 Popular Paths of Commuting Trips

The previous section indicates that TRANSIT accurately extracts urban mobility patterns. Therefore, as a second application, we focus on inferring popular commuting trips within a city. The knowledge of such trips is an extremely precious source of information for transport authorities and city planners as: *(i)* they represent the largest share of the the daily urban traffic demand of a city; *(ii)* they identify the typical commuting behaviors of travelers which regularly stress the transport network infrastructure, especially during peak hours; *(iii)* they are hard-to-quantify and characterize at city-scale because of the absence of dedicated sensors or probes that can precisely capture the multi-modal, diverse and time-varying nature of such trips.



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

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

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

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

Figure 3.15: Heatmap of commuting trips in Lyon and Paris.

By applying our framework to the large-scale datasets $\mathcal{D}_P$ and $\mathcal{D}_L$, we associate to each user i of the two analyzed cities a set of trips $\widehat{\mathcal{M}}^i$ for the whole period of 3 months. As explained above, $\widehat{\mathcal{M}}^i$ can be divided in two subsets: $\widehat{\mathcal{M}}^i_r$, a subset of recurrent trips enhanced by TRANSIT, and $\mathcal{M}^i_o$, a set of non-recurrent trips of user

u. Considering that commuting trips are, by definition, recurrent, in the remainder of this section we focus our analysis only on subset $\widehat{\mathcal{M}}_r^i$.

Furthermore, to extract commuting trips from $\widehat{\mathcal{M}}_r^i$, we filter only those trips associated to the two most popular locations of each user, under the constraint that at least 10 trips are present between these two locations. The underlying assumption is that the remaining set should mostly contain the two most popular trips performed by users in their daily routine, *i.e.*, home-to-work and work-to-home trips (commuting trips).

The spatial density (heatmap) of the reconstructed trips is represented in Figure 3.15a for Lyon and Figure 3.15b for Paris. As a first consideration, the recurrent trips appear to have overall a good match with the multi-modal urban transportation network, graphically overlapped to the heatmap in figure 3.15c for Lyon and in figure 3.15d for Paris. A more in-depth inspection of the figures highlights that, for both cities, the subway network, the tramway lines and most important urban roads clear show up among the commuting trips reconstructed via TRANSIT. In the case of Paris, NSD trips appear to have a near perfect match to the underlying multi-modal transport network. The less evident match for the case of Lyon, especially characterizing some peripheral roads (however present in the heatmap), can be explained by the lower number of available trips and the lower density of the cellular network of Lyon in these areas, compared to those from the capital city.

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

3.9 Discussion

Our developed framework TRANSIT overcomes two main limitations of the mobile phone data. First of all, TRANSIT solves the trajectory identification task with high accuracy with respect to state of the art approaches. In addition, by leveraging the repetitive nature of human mobility and the oscillation effect, TRANSIT outperforms other approaches like CWMA and DECRE on the trajectory augmentation task. Moreover, our approach is scalable so that it enables new human mobility related applications as presented in this chapter. We also demonstrated that TRANSIT is able to estimate accurately travel demand in urban environment with comparison analysis with surveys. Based on these strong results, we explored other applications allowed by TRANSIT in Chapter 6.

Despite its already satisfying results, TRANSIT can be further improved over several dimensions. The reconstructed trajectories could be easily map matched to the different lines and modes of the underlying transportation network to further reduce their spatial error. This aspect is investigated in Chapter 4. Other kinds of data on human mobility, such as smart card logs or GPS floating car data, could be jointly leveraged with NSD, *e.g.*, to better estimate the typical duration of similar trips and support a more informed filtering during the trajectory clustering process.

Further technical optimizations could be helpful towards a stream-based online implementation of TRANSIT that could support a large variety of real-time applications, such as urban anomaly detection, data-driven dynamic control of transport infrastructures, as well as advanced location-aware caching schemes and scheduling policies for telecommunication networks. Finally, TRANSIT exploits the regularity of the mobility of a particular user to increase spatio-temporal accuracy of NSD, inter-individual mobility regularity by leveraging clustering algorithm could be explored in future works.

3.10 Conclusion

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

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

We applied TRANSIT at scale, to the whole subscriber base of a major network operator in two major cities in France, Paris and Lyon. This allows to validate TRANSIT at large scale showing that it achieves 0.89 Pearson correlation with surveys for inferring temporal demand profiles. Such a result allows TRANSIT to outperform state of the art approaches including the one studied in Chapter 2. Moreover, on the large scale dataset, TRANSIT is able to identify 480 million trajectories of over 10 millions of individuals during a period of three months in 2019 – and improve substantially the accuracy of 100 millions of those. We leverage such a unique information to carry out preliminary explorations of: (*i*) the fraction of trips using public transport versus other modes, (*ii*) the metropolitan-scale commuting paths. To the best of our knowledge, we are the first to employ NSD to conduct mobility analysis at such a large scale.

Chapter 4

Multi-Modal Path Inference in Urban Environment

In this chapter, in line with the previous chapter, we study the problem of map-matching network signaling trajectories to the urban transportation network to further refine the information that we can retrieve from network signalling trajectories. Thus, we developed an Hidden Markov Model based map-matching approach that we apply on the set of mobile sessions outputed by TRANSIT. At microscopic scale, our map-matching approach is 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 result is 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. At macroscopic scale, we propose a method for overcoming the latter assumption and show that our approach is promising for inferring popular paths per transportation mode.

The chapter is structured as follows. Section 4.2 provides the problem that this chapter tackles. Section 4.3 presents the related works. Section 4.4 is dedicated to develop the theoretical background of our approach based on Hidden Markov Model. Then, we present, our Hidden Markov Model based map-matching in Section 4.5. Results at microscopic and macroscopic scales are presented respectively in Section 4.6 and Section 4.7. In Section 4.8 the results are discussed and future research directions are proposed. Finally, Section 4.8 present the conclusion of this chapter.

This chapter contains parts of the following article:

Bonnetain L., Furno A., Krug J., El Faouzi N.-E. (2019), "Can we map-match individual cellular network signaling trajectories in urban environments? Data-driven study". In: *Transportation Research Record: Journal of the Transportation Research Board*.

Bonnetain L., Furno A., El Faouzi N.-E. (2021), "Multi-modal fine-grained map-matching of mobile phone network signaling data in urban areas". In: *Transportation Research Board*.

4.1 Notation for this chapter

Symbol	Description
v_i	Hidden state i.
o_j	Observation j.
a_{ij}	Transition probability from hidden state v_i to hidden state v_j.
A	Transition probability matrix.
π_i	Initial probability of hidden state v_i.
π	Initial probability matrix.
B_{ij}	Emission probability that hidden state v_j emits observation o_i.
B	Emission probability matrix.
G_{road}	Graph representing road network.
G_{tc}	Graph representing public transport network.
G_{train}	Graph representing train network.
V	Set of vertices.
E	Set of edges.
k	Node degree.
l	Edge length.
t_i	Timestamp of i^{th} event of a trajectory T.
O	Set of observation $(o_1, \ldots, o_M)$.
o_i	i^{th} signaling event.
T	Cell phone trajectory defined as a set of observation $(o_1, \ldots, o_n)$.
M	Number of observations.
N	Number of hidden states.
$\mathcal{M}^i$	Set of all mobile sessions across the whole observation period $[t_0^i, t_{N-1}^i]$ of device i.
$\mathcal{M}_R^i$	Set of trajectories from $\mathcal{M}^i$ that are classified in a cluster by DBSCAN and identified as recurrent by TRANSIT.
$\widehat{\mathcal{M}_R^i}$	Set of recurrent trajectories from $\mathcal{M}^i$ that are spatially augmented by TRANSIT.
$\mathcal{M}_O^i$	Set of unique trajectories from $\mathcal{M}^i$ that are classified as outliers by DBSCAN and left spatially unmodified by TRANSIT.
$\widehat{\mathcal{M}^i}$	Final set of trajectories retrieved by TRANSIT corresponding to $\widehat{\mathcal{M}_R^i} \cup \mathcal{M}_O^i$.

Table 4.1: Chapter 4′ specific notations

4.2 Introduction

Surveys, despite the limitations described in Chapter 1, are able to retrieve mobility knowledge on all transportation modes. This knowledge, although coarse at the spatiotemporal level allows to capture the city scale dynamics of all transportation modes. The situation changes when it comes to data-driven human mobility analysis. Among all these data-sources, none of them bring reliable information on transportation mode at large scale. For instance, smartcard data capture only public transport dynamics. For the GPS data, the transport mode inference is possible but the small samples available with this kind of data do not allow large scale mobility studies. LBSN trajectory are too sparse to apply transportation mode inference. In this context, the mobile phone data remain the only candidate susceptible to capture multiple transportation mode dynamics at large scale.

In the previous chapter, we presented our approach TRANSIT able to increase temporal granularity and reduce the spatial error of raw signalling trajectories. In this chapter, we will investigate whether it is possible to accurately reconstruct mobility information based on the output of TRANSIT. In particular, we studied the problem of map-matching network signalling trajectories in order to reconstruct the exact path followed by the user in a multi-modal transportation network. The value of map-matching approach in multi-modal settings is twofold. First, such an approach would allow to further reduce the spatial error of the signalling trajectories of TRANSIT. Second, it could allow to infer very valuable mobility information such as the transportation mode and especially the path followed by the user on the transportation network.

However, given the nature of signalling trajectories, multimodal map-matching is a very challenging task. On the one hand, state of the art map-matching approaches on mobile phone data consider only the road network. On the other hand, the transportation mode knowledge extractable from approaches of the literature is very limited: coarse transportation detection or transport mode detection in inter-urban environment. The next section 4.3 will discuss the latter points in detail. This chapter aims at developing a map-matching approach that could be applied to the mobile sessions that TRANSIT outputs (Chapter 3). The latter is an attempt to solve the challenging problem of multimodal path inference in urban environment.

4.3 Literature Review

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 [74].

Several approaches exist in the literature to solve the problem of map-matching GPS traces to a transportation network. The map-matching of GPS trajectories is a widely studied topic and the approaches to solve such a problem are necessary for navigation systems. Quddus *et al.* [94] 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 [126].

- *Topological approaches* use geometric information as well as topological information like the existence of connectivity between nodes of the network [135]. Very sensitive to noise and outliers, these approaches are not appropriate to solve map-matching problems in presence of highly noisy and sparse data.
- *Probabilistic methods* exploit 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.
- *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 [82], Dempster–Shafer theory [135], fuzzy logic models [93], or the application of Bayesian inference [58].

The latter category corresponds to state-of-the-art algorithms which may achieve a very high accuracy (location error lower than 10 meters) with high sampling rate GPS data. Newson *et al.* [81] 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 trajectories 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. Schulze *et al.* [99] 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.* [52], which reaches 75% of median accuracy.

In addition, HMM-based map-matching has become the state-of-the-art approach for noisy and sparse location data and, *a fortiori*, mobile phone trajectory. Thiagarajan *et al.* [113] and, more recently, Algizawy *et al.* [4] 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 is the case in [113]. Jagadeesh et al. [52] proposed an online map-matching algorithm combining HMM-based map-matching and route choice model.

The main limitation of the above mentioned works is to study the map-matching of cell phone trajectories only on 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.* [6]. The authors have developed an approach, namely CT-Mapper, which is an unsupervised HMM model which aims at mapping sparse multimodal cellular trajectories by using a multi-layer transportation network. Yet, CT-Mapper has some limitations: the multi-layer network

does not cover all transportation modes such as bus and tramways. In addition, CT-Mapper requires already preprocessed 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.* filtered out trajectories whose lengths are shorter than 5 kilometers, only keeping longer 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 as the choice of covered modes seems to confirm.

Other works tackle the transportation mode detection problem without relying on map-matching. These approaches from the literature usually deal with simplified settings. For instance, some of them aim at differentiating between road and train inter-city trips [101]. Some works focus on retrieving easy to detect transportation mode like train, metro or plane [130, 45] Other works, simplified the transportation mode inference by grouping several transportation modes such as public transport versus private transport [88, 92], air versus ground, moving versus stationary [20, 48] and rail versus road [6].Only two works develop approaches aiming at differentiate finely all transportation mode. Danafar *et al.* [28] develop an approach for retrieving 6 different modes: car, bus, tram, train, cycling and walking. However, there is no quantitative result that would allow to evaluate the performance of the approach. Chin *et al.* [25] propose an hybrid model based rule based heuristics combined with random forest. Their classification approach achieves accuracy of 73% when differentiating 5 modes. However, the approach has several limitations. The ground truth dataset is unbalanced in terms of transportation mode trips distribution with an overepresentation of subway trips which are easier to infer compared to other transportation mode. They applied supervised approach on a small ground truth dataset which can lead to overfitting. Finally, they did not apply and validate their approach at large scale.

Based on such an analysis of the related work, we will investigate the problem of map-matching cellular trajectories using Hidden Markov Model. By considering the problem in a multimodal context and for an urban use case, we aim at investigating and overcoming the limitations of state-of-the-art approaches. Considering such settings makes the map-matching problem even more challenging.

4.4 Theoretical Background

Hidden Markov Model is the core model on which our map-matching approach is built. Thus, we present the theoretical background of this model in the following sections.

4.4.1 Markov Chain

A Markov chain is a stochastic process in which the system follows the Markov property. A Markov chain is based on states and transitions, a transition represents the passage from one state to another (that can be the original state itself). The Markov property states that the probability of being in a given state depends only on the previous state.

The figure 4.1 shows an example of Markov chain. The states are denoted as v_1 and v_2. For instance, if the system is in state v_1, it can transit to the state v_2 with

probability a_{12} or stay in the same state v_1 with the probability a_{11}. In general, the stochastic processes studied depend on time. Thus, at each *unit of time*, a transition is made, which ultimately generates a sequence of states.

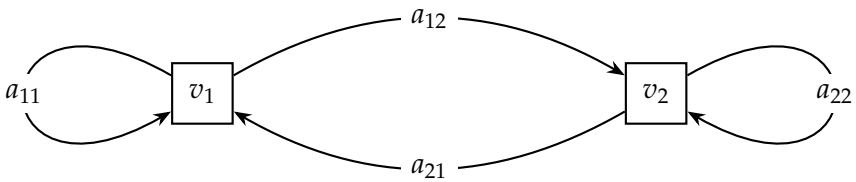


Figure 4.1: Markov chain

4.4.2 Hidden Markov Model

The Hidden Markov Model is a statistical mathematical model derived from Markov chains. In this model, we cannot directly observe the sequence of states: the states are *hidden*. Nevertheless, each state emits *observations* which are observable. The system generates a sequence of observations that are emitted by hidden states. Information about these hidden states can be retrieved from the observations. Indeed, the probability that the system generates a given observation depends on the state of the system.

The figure 4.2 represents a Hidden Markov Model with two hidden states v_1 and v_2 and two observations o_1 and o_2. It is based on the Markov chain with additional emission probabilities b_{ij} and initial probabilities π_i. For instance, the emission probability b_{12} represents the probability of the state v_1 to emit observation o_2 *i.e.*, the probability of observing o_2 given that the hidden state is v_1. Finally, the initial probabilities π_1 and π_2 corresponds to the probability that the user is respectively in the hidden state v_1 and v_2 at the first unit of time.

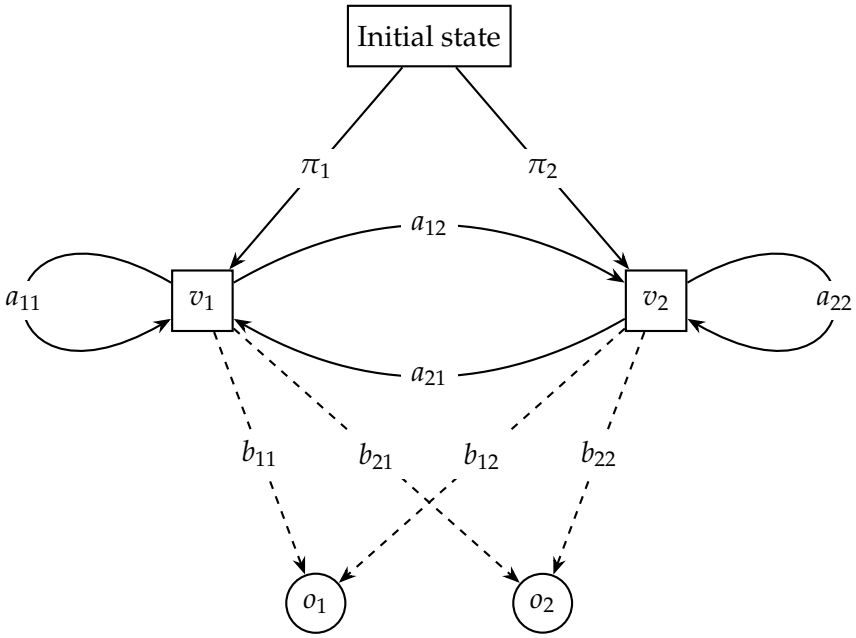


Figure 4.2: Hidden Markov Model

In a more formal way, the Hidden Markov Model with N hidden states and M observations is defined by a five-fold $\langle V, O, \pi, A, B \rangle$:

- $V = \{v_1, \ldots, v_N\}$ is a set of N hidden states.
- $O = \{o_1, \ldots, o_M\}$ is a set of M observations.
- π is a $N \times 1$ vector of initial probabilities, it defines the probability of being on each of the hidden states at the initial moment. In addition, π is probability distribution: $\sum_{i=1}^{N} \pi(i) = 1$.
- A is a $N \times N$ transition matrix, it defines all transition probabilities between all the states. The transition probability from v_i to v_j is $P(v_j|v_i)$ denoted as a_{ij}. Besides, $\forall v_i \in V$, $\sum_{v_j \in V} P(v_j|v_i) = 1$.
- B is a $M \times N$ emission matrix, it defines all emission probabilities between the observations and the hidden states. The probability that the hidden state v_j emits the observation o_i is $P(o_i|v_j)$ denoted as b_{ij}. Besides, $\forall v_i \in V$, $\sum_{o_i \in O} P(o_i|v_j) = 1$.

Given this model, given a sequence of observations, the application of the hidden Markov model allows to solve three typical problems :

- determines the probability that the system emits a given sequence.
- determines the most likely sequence of hidden states resulting from a given sequence.
- determines the optimal parameters of the Hidden Markov Model so that the probability that the system emits a given sequence is maximum.

For solving the map-matching problem using a hidden Markov model, we are interested in solving the second category of problem. The algorithm used to find the solution of this problem is the Viterbi algorithm presented in Section 4.4.3

4.4.3 Viterbi Algorithm

This section presents the basis of the Viterbi algorithm. The algorithm is based on formulas demonstrated in this section and will be further used in the implementation of our map-matching approach presented in the Section 4.5.

The Viterbi algorithm [119] is a recursive algorithm which aims at finding the most likely sequence of hidden states $(v_1 \ldots v_t \ldots v_l)$ given a sequence of observations $(o_1 \ldots o_t \ldots o_l)$, with l the length of the sequence.

Let us define $o_t \in O$, the observation at the unit of time t, $t \in [\![1, l]\!]$.

Let us define $v_t \in V$, the current state of the system at the unit of time t, $t \in [\![1, l]\!]$.

Thus, the problem that the Viterbi algorithm solves is the following :

$$max(P(v_1 \ldots v_t \ldots v_l | o_1 \ldots o_t \ldots o_l)) \tag{4.1}$$

According to the standard definition of the conditional probability, we get :

$$P(v_1 \dots v_t \dots v_l | o_1 \dots o_t \dots o_l) = \frac{P(v_1 \dots v_t \dots v_l \cap o_1 \dots o_t \dots o_l)}{P(o_1 \dots o_t \dots o_l)} \tag{4.2}$$

As the probability $P(o_1 \dots o_t \dots o_l)$ in equation 4.2 is constant the maximisation problem becomes :

$$max(P(v_1 \dots v_t \dots v_l \cap o_1 \dots o_t \dots o_l)) \tag{4.3}$$

By assuming that the hidden states $v_i, v_j, (1 \leq i, j \leq N)$ are independent as well as the observations $o_i, o_j, (1 \leq i, j \leq N)$, we get :

$$\begin{aligned} P(v_1 \dots v_l, o_1 \dots o_l) = P(v_1 \dots v_{l-1}, o_1 \dots o_{l-1}) \cdot P(v_l | v_1 \dots v_{l-1}, o_1 \dots o_{l-1}) \\ \cdot P(o_{l-1} | v_1 \dots v_{l-1}, o_1 \dots o_{l-1}) \end{aligned} \tag{4.4}$$

Since the Hidden Markov Model is without memory, the state of the system only depends on the previous state of the system and does not depend on any other anterior state or observation. Thus, we obtain:

$$P(v_t | v_1 \dots v_{t-1}) = P(v_t | v_{t-1}) \tag{4.5}$$

Similarly, the observations only depend on the previous state of the system but do not depend on all anterior states or observations. We also get :

$$P(o_t | v_1 \dots v_t, o_1 \dots o_{t-1}) = P(o_t | v_t) \tag{4.6}$$

By substituting equation 4.5 and equation 4.6 in equation 4.4, the new formulation of the problem is the following :

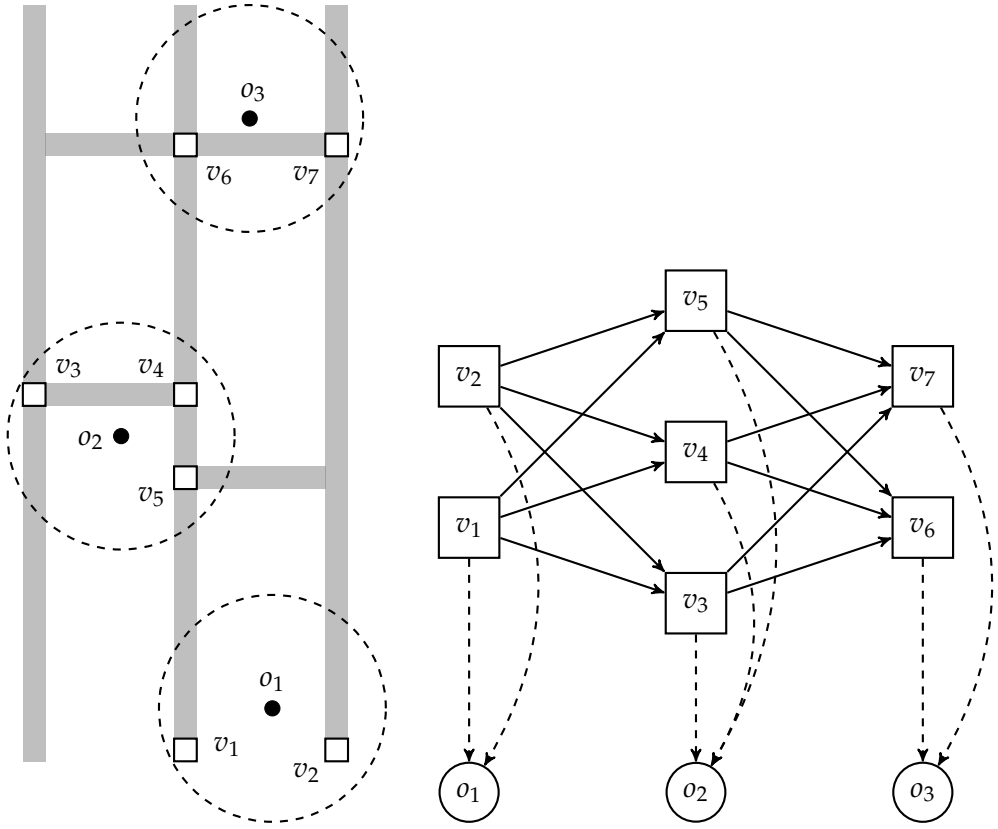
$$max(P(v_1 \dots v_t \dots v_l | o_1 \dots o_t \dots o_l)) = max(P(v_0) * \prod_{t=1}^{l} (P(v_t | v_{t-1}) * P(o_t | v_t)) \tag{4.7}$$

$P(v_0)$ is given by π, $P(v_t | v_{t-1})$ is given by A and $P(o_t | v_t)$ is given by B. Thus, given the HMM parameters, the solution of equation 4.7 can be found.

4.5 Map-matching Signalling Trajectories

4.5.1 HMM based Map-Matching

Our map-matching problem can be modelled using a Hidden Markov model. The hidden states are modelled as the set of vertices (nodes) of the graph representing the transportation network. The observations are modelled as the unique set of x-y coordinates of the cell phone trajectories considered. The Hidden Markov Model allows to solve the following problem: from the sequence of positions, the mobile phone trace (sequence of observations), find the most probable sequence of nodes on the graph (sequence of hidden states). This modelling of the map-matching problem is illustrated in Figure 4.3. The observations o_1, o_2 and o_3 correspond to a sequence of positions representing the mobile phone trace. Each position has an area which correspond to the spatial error of the mobile phone position. The nodes of the transport



(a) Example of signalling trajectories (o_1, o_2 and o_3) and candidates nodes ($v_1, v_2, \cdots v_7$)

(b) Hidden Markov Model

Figure 4.3: Illustration of Hidden Markov Model based map-matching

network within it are candidate nodes. For example, the nodes v_1 and v_2 are candidate nodes for having transmitted the observation o_1. Then, the solid arrows correspond to the different possible transitions between the candidate nodes between two consecutive instants. The dotted arrows represent the possible emissions of observations by the candidate nodes at each instant.

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

4.5.2 Network Modeling

In line with Chapter 2 and Chapter 3, we consider the multi-modal transportation network of the city of Lyon 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). The assumption of considering the three sub-graphs G_{road}, G_{tc} and G_{train} separately and not connected at each other is discussed and stressed in Section 4.6. For the rest of the chapter, 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 [13], 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.* [89], 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

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

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.* [89]. Some statistics of the final multi-modal transportation network G is given in Table 4.2.

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 4.2: 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 $\langle l \rangle$.

4.5.3 HMM Parameters

In the following, the parameters $\langle V, O, \pi, A, B \rangle$ denotes the HMM parameters for map-matching a set of signalling trajectories (set of mobile sessions $\widehat{\mathcal{M}}^i$ inferred by TRANSIT) to any transportation graph (G_{road}, G_{tc}, G_{train} or G).

Initial Probability

In our scenario, no particular *a priori* exists, as a user could start his trip from any initial location. Therefore, 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:

$$\pi(v_m) = \frac{1}{N} \tag{4.8}$$

Transition Probability

The transition probability corresponds to the probability that a mobile phone user moves on the underlying transportation network from hidden state v_m at time $t-1$ to hidden state v_n at time t. In the following, we choose the definition proposed by Putra et al. [89], *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

[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.

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 $a(v_m, v_n)$ between the generic pair of nodes v_m and v_n is defined to be exponentially decreasing according to the travel-time weighted shortest path between the two nodes v_m and v_n. Formally:

$$a(v_m, v_n) = \exp^{-\beta \cdot tt_{v_m,v_n}}, \quad tt_{v_m,v_n} = \sum_{\forall (s_u,s_v) \in SP_{mn}} tt_{v_u,v_v} \tag{4.9}$$

where (s_u, s_v) is the generic edge on the travel-time weighted shortest path SP_{mn} connecting the two nodes v_m and v_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_{v_u,v_v} denotes the travel time between each two nodes v_u and v_v. β is a damping factor to control the effect of the travel time.

Emission Probability

When the trajectory is a sequence of reconstructed positions with a given spatial error, the map-matching problem can be viewed as a map-matching problem with noisy GPS points. Similarly to Newson *et al.* [81], we model the emission probability as a Gaussian noise centered on the hidden state v_m and an empirically estimated standard deviation of the distance error between hidden states and observations:

$$b(v_m, o_k) = \frac{1}{\sqrt{2\pi}\alpha} e^{-0.5\left(m \frac{d_{v_m,o_k}}{\alpha}\right)^2} \tag{4.10}$$

where d_{v_m,o_k} is the geodesic distance between the generic observation o_k and the generic node v_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 Section 4.6.1.

4.5.4 Map-Matching Algorithm

Procedure 1 Map-Matching Algorithm

Input:
Transportation Network, $G_j(V, E)$
States (Network Nodes), $V = \{v_0, \ldots, v_{N-1}\}$
Cell phone trajectory, $T = (o_0, \ldots, o_{l-1})$, l is the length of the sequence
Initial probabilities, π_i such that $i \in V$
Transition probabilities, a_{ij} such that $i, j \in V$
Emission probabilities, b_{ik} such that $i \in V$ and $k \in C$

Output:
Maximum probability, $OutputProb$
Most likely expanded node sequence, $FinalPath =< v_{o_0}, \ldots, v_{o_{l-1}} >$

First step: Optimized Viterbi Algorithm
1: $StateProb[t][y] \leftarrow 0, \forall t, y$
2: $Path \leftarrow \{\}$
3: **for** all y in V **do**
4: $\quad StateProb[0][y] = \pi_y \cdot b_{y,o_0}$
5: $\quad Path[y] \leftarrow y$
6: **end for**
7: **for** $t \leftarrow 0$ to $l - 1$ **do**
8: $\quad$ **for** all y in $V | b_{y,o_t} \neq 0$ **do**
9: $\quad\quad \langle Prob, Pred \rangle \leftarrow \langle \max_{z \in V | a_{z,y} \neq 0} (StateProb[t-1][z] \cdot a_{z,y} \cdot b_{y,o_t}), z \rangle$
10: $\quad\quad StateProb[t][y] \leftarrow Prob$
11: $\quad\quad NewPath[y] \leftarrow Path[Pred] + y$
12: $\quad$ **end for**
13: $\quad Path \leftarrow NewPath$
14: **end for**
15: $\langle Prob, Pred \rangle \leftarrow \langle \max_{y \in V}(StateProb[l-1][y]), y \rangle$
16: $OutputProb \leftarrow Prob$
17: $OutputPath \leftarrow Path[Pred]$

Second step: Final itinerary reconstruction
18: $FinalPath \leftarrow OutputPath[0]$
19: **for** $k \leftarrow 1$ to $OutputPath.size() - 1$ **do**
20: $\quad FromNode \leftarrow OutputPath[k-1]$
21: $\quad ToNode \leftarrow OutputPath[k]$
22: $\quad IntPath \leftarrow ShortestPath(FromNode,\ ToNode,\ G_j)$
23: $\quad FinalPath \leftarrow FinalPath + IntPath$
24: **end for**
25: **return** $\langle FinalPath, OutputProb \rangle$

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, reported in Pseudo-code 1. The first phase consists in an optimized Viterbi algorithm [119]. 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 [4]. To improve performance, we implemented an optimized version of the Viterbi algorithm as done by Algizawy *et al.* [4] 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 v_m and observation o_k is larger than $2km$, the emission probability $e(v_m, o_k)$ is rounded to 0; (*ii*) similarly, if the distance between state v_m and state v_n is larger than $5km$, transition probability $a(v_m, v_n)$ 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.

4.6 Microscopic Validation

4.6.1 Datasets

The microscopic dataset used in this section is the same dataset used for validating TRANSIT in Chapter 2. As a reminder, 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 [60] 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 Section 4.5.1 makes our map-matching approach suitable for GPS trajectories according to recent works on GPS map-matching [81, 89]. 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 [81, 89], 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.

Parameter Choice

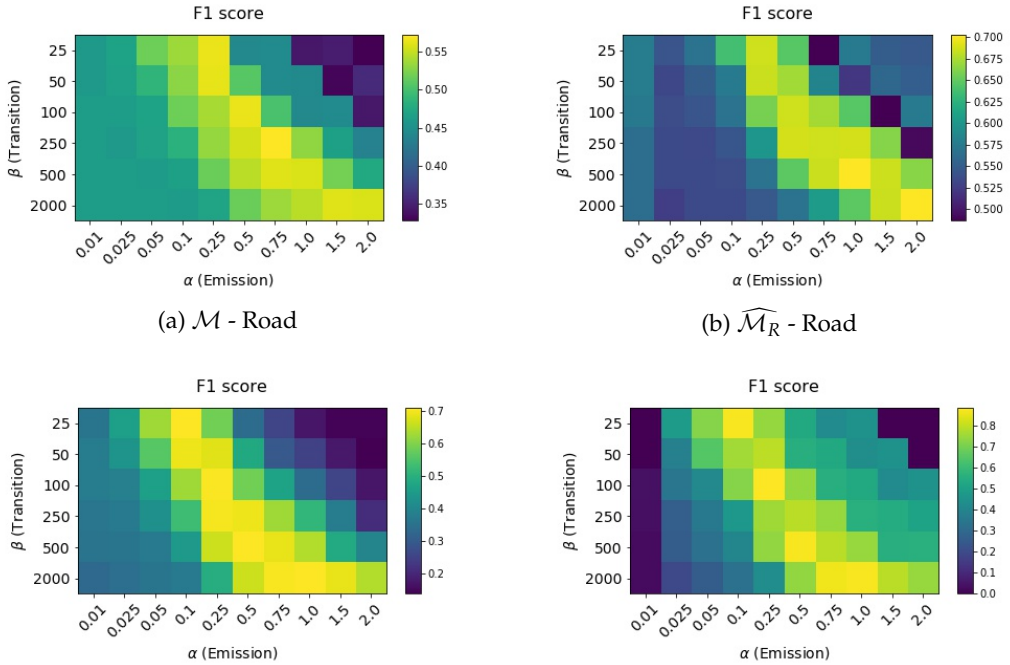
Our map-matching approach depends on two parameters, namely α and β, respectively associated to the emission and transition probabilities of the Hidden Markov Model. In order to choose the best values for such parameters, we apply our approach on M and $\widehat{\mathcal{M}}_R$ and then compute the average 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:

$$\text{Precision} = \frac{TP}{(TP + FP)}, \quad \text{Recall} = \frac{TP}{(TP + FN)} \tag{4.11}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{4.12}$$

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.

Figure 4.4 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 Figure 4.4 (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.



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

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

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

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

Figure 4.4: 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)

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 4.3: 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.

4.6.2 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:

$$MR = \frac{TP}{TP + FN + FP} \quad \text{and,} \quad G_e = \frac{1}{|m_{NSD}|} \sum_{e_n \in m_{NSD}} \min_{e_{n'} \in m_{GPS}} d(l_n, l_{n'}) \tag{4.13}$$

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

4.6.3 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.

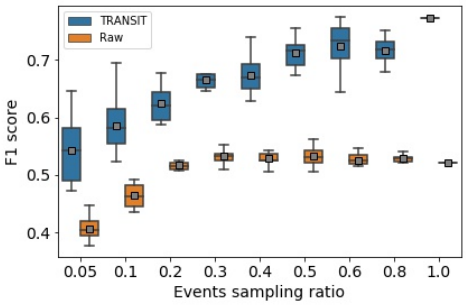


Figure 4.5: Performance of the map-matching with and without TRANSIT with varying events sampling rate.

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 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.

4.7 Macroscopic Validation

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 Section 4.6.2, 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 Section 7. 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[3]. 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 Figure 4.6, 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 Figure 4.6a and Figure 4.6b, 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 Figure 4.6c

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

and Figure 4.6d, 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 Figure 4.6b).

For the OD pair C_2, trips are associated either to the road or to the train subnetworks. For train trips, in Figure 4.6e and Figure 4.6f the only popular itinerary is correctly retrieved by our approach. Concerning car trips, in Figure 4.6g and Figure 4.6h, 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.

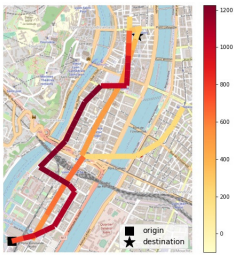
Finally, for the C_3 OD pair, as reported in Figure 4.6i, Figure 4.6j, Figure 4.6k and Figure 4.6l, trips are associated either to the road or to the public transport subnetworks. For both car and public transport trips, a good match for popular paths can be observed: our approach retrieves the main popular paths detected via route planners. We also highlight that popular itinerary 1 for public transport is a multimodal one.

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

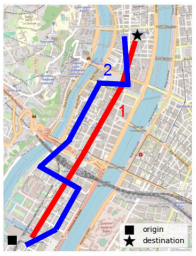
4.8 Conclusion

In this chapter, we investigated the potential of network signalling data to provide fine-grained spatio-temporal information to reconstruct users' mobility. Thus, 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 applied on the output of TRANSIT (the framework has been previously presented in detail in Chapter 3).

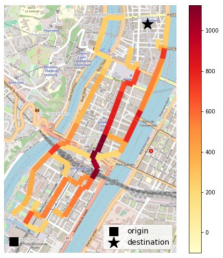
To validate our approach, we have analyzed an original case study, related to French city of Lyon, by leveraging both real cellular traces collected by a major network operator and GPS data collected via a mobile phone application. This data has been leveraged to perform a microscopic validation, aimed at both fine-tuning the parameters of the HMM-based map-matching step and at showing the capability



(a) C_1 - Reconstructed popular paths - PT

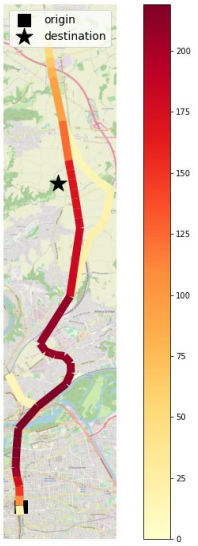


(b) C_1 - Ground-truth popular paths - PT



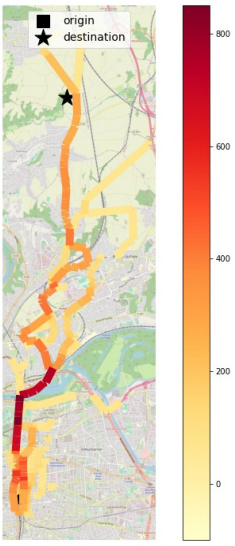
(c) C_1 - Reconstructed popular paths - Road

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



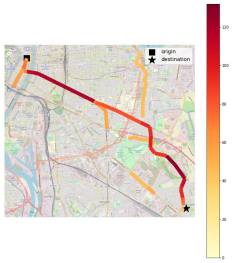
(e) C_2 - Reconstructed popular paths - Train

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



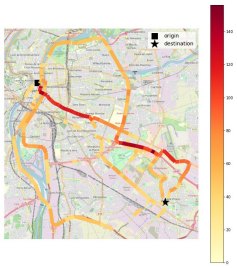
(g) C_2 - Reconstructed popular paths - Road

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

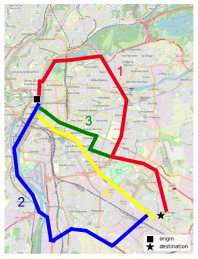


(i) C_3 - Reconstructed popular paths - PT

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



(k) C_3 - Reconstructed popular paths - Road



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

Figure 4.6: 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.

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.

Then, using simple transportation mode inference, we have demonstrated the possibility 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 works directions that are not tackled in this thesis include improvements on the HMM parameters. For the initial probability, instead of having all states equally probable, the probability could be weighted by the level of confidence that a user is in a particular transportation mode. This level of confidence could be estimated based on trip information (travel time, speed ...). Besides, transition matrix could be dynamic and estimated based on travel time depending on real traffic conditions.

Other work directions include improvement on the transportation mode inference technique. Indeed, instead of using only the speed, multiple features could be used for the inference such as: the probability that the Viterbi 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.

Chapter 5

Urban Mobility Dynamics Extraction

In this chapter, we study the problem of urban dynamics extraction from mobile phone data in the city of Lyon. Based on TRANSIT approach developed in Chapter 2, we can represent the urban mobility of Lyon as a mobility tensor with fine spatio-temporal granularity. We leverage Tucker decomposition approach to extract urban dynamics from this mobility tensor. The results show that the approach is able to retrieve typically known temporal patterns. In space, the approach is able to find relevant origin and destination communities. The *size* of these communities is consistent with the density of trips of these zones. Thus, the approach can retrieve interesting spatial and temporal patterns as well as complex spatio-temporal dependence. Finally, we demonstrate that the approach is still able to retrieve the main urban dynamics with incomplete mobility tensor (mobility sample sampled with 50% rate).

The chapter is structured as follows. Section 5.2 provides the problem that this chapter tackles. Section 5.3 presents the related works. Section 5.4 is dedicated to develop the theoretical background of our approach based on Non-negative Tucker Factorization. The parameter choices, the extraction of urban dynamics using our decomposition approach as well as a comparative analysis of state of the art approaches are provided in Section 5.5. Finally, Section 5.6 present the conclusion of this chapter and future directions that could be explored.

5.1 Notation for this chapter

Symbol	Description
M	Number of zones
N	Number of time slices
$\boldsymbol{\mathcal{E}}$	Random error tensor
$\boldsymbol{\mathcal{M}}$	The mobility tensor
m_{xyz}	The (x, y, z) elment of $\boldsymbol{\mathcal{M}}$
$\widehat{\boldsymbol{\mathcal{M}}}$	The reconstructed mobility tensor
$\mathbf{W}$	the urban context matrix
w_{pq}	The (p, q) element of $\mathbf{W}$
N_{trips}	Total number of trips
$\boldsymbol{\mathcal{C}}$	The core tensor
c_{ijk}	the (i, j, k) element of $\boldsymbol{\mathcal{C}}$
$\mathbf{O}$	Origin factor matrix
I	Number of origin spatial patterns
$\mathbf{D}$	Destination factor matrix
J	Number of destination spatial patterns
$\mathbf{T}$	Temporal factor matrix
K	Number of temporal patterns
$\mathbf{o}_x, \mathbf{d}_x, \mathbf{t}_x$	the x-th row vectors of $\mathbf{O}$, $\mathbf{D}$, $\mathbf{T}$
$\mathbf{o}_{:j}, \mathbf{d}_{:j}, \mathbf{t}_{:j}$	the i-th column vectors of $\mathbf{O}$, $\mathbf{D}$, $\mathbf{T}$
$\mathbf{o}_{xj}, \mathbf{d}_{xj}, \mathbf{t}_{xj}$	the (x, i) elements of $\mathbf{O}$, $\mathbf{D}$, $\mathbf{T}$
$\gamma, \delta, \epsilon, \zeta$	Regularization parameters
α, β	urban context parameters

Table 5.1: Chapter 5′ specific notations

5.2 Introduction

In Chapter 1, we presented the various sources of data that have emerged for studying human mobility. We showed the potential and the limitations of one of the most promising source: mobile phone data. The latter were leveraged in Chapter 2 to infer travel demand patterns. In the same chapter, we also demonstrated, with a comparison with surveys, that mobile phone data were able to infer travel demand accurately even in urban environment. Then, in Chapter 3, we developed TRANSIT, a framework able to transform noisy and sparse mobile phone trajectories into fine-grained trajectories with contextual information (mobility/staticity of the mobile phone user over time). This chapter aims at developing approaches able to extract urban dynamics, meaningful knowledge on human mobility in urban context. Our approach is based on TRANSIT which allows it to infer these dynamics at unprecedented spatio-temporal scale.

Understanding urban dynamics is a very important step for transportation planners. Such a knowledge is necessary to design and improve urban transportation systems. These urban dynamics of mobility are characterized by multiple dimensions. For instance, at individual level the mobility is characterized by a lot of attributes such as trip origin, trip destination, time, transportation mode, socio-economic status of the persona and so on. When it comes to analyze human mobility at aggregated level, this mobility can be analyzed along different dimensions, each dimension being an above mentioned attribute. Given the multivariate nature of human mobility and the complex travel behaviours of individuals, extracting interesting urban dynamics along multiple dimension is very challenging. This is especially true when the number of dimensions analyzed increases.

Moreover, representing the mobility in a form where it can be further analyzed is not trivial. If the mobility (a set of trips) have to be analyzed along one dimension, for instance the origin, a natural way to represent this mobility is to use a vector representation where each element denotes the number of trips emitted from each origin. Similarly, if the mobility has to be analyzed in two dimensions, a natural representation is a matrix. More generally, if the mobility has to be analyzed in N dimensions, the natural representation is a N-way tensor. However, analyzing this tensor is very challenging given the number of its elements. Developing approaches able to extract meaningful knowledge related to urban dynamics is the goal of this chapter.

5.3 Literature Review

The research community has largely studied the different tensor factorization approaches [64]. Thus, thanks to these theoretical advances, tensor factorization has recently fed many applications in different domains [64]: in medicine for the analysis of electroencephalogram (EEG) signals [77], face recognition [118], psychology analysis [63] or movies recommendation [111]. In this context, the tensor factorization has also became popular in the transportation domain. In the field, tensor factorization have been leveraged for two main purposes. The first category is to reconstruct tensors for predicting unknown values in multi-variate data sets. This includes a lot of applications such as completing missing traffic data [110], inferring urban gas consumption [143], predicting travel time [125] or recommending social

tags [108]. In the second category, the works aims at discovering latent structure from multivariate data. Our work falls in this second category.

In the mobility field, tensor factorization has been widely used to understand urban dynamics. The common approach is to represent the the aggregated mobility of a region as dynamic origin destination matrices which is a 3-way tensor (with time, origin and destination as dimensions of this tensor). Then, the tensor factorization can be applied on the 3-way tensor. Among the possible approaches for estimating urban mobility patterns, Wang *et al.* [121] have shown that that Tucker decomposition performs better than CANDECOM/PARAFAC. Sun *at al.* explore tensorial probabilistic latent semantic analysis (PLSA) which is equivalent of tensor decomposition. These decomposition has been applied for modelling urban dynamics with different sources of data: smartcard data [106], mobile phone data [120] and taxi data [112, 87]. Recently, another model has been proposed which combine Tucker decomposition with external urban context knowledge from point of interest data and neighboring regularization [122]. For this chapter, we implemented this approach. Thus, the latter will be presented in details in the next section.

However, these works have limitations that we aim at overcoming. First, the works often rely on coarse space segmentation for creating their dynamic OD matrices (mobility tensor). When the data source is mobile phone data [120], the reason of such a choice is the lack of methods able to estimate accurately the origins and the destinations of the mobile phone trips. Thus, to mitigate the bias, they have to build OD matrices with poor spatial granularity. This problem has already been discussed in Chapter 2 and Chapter 3. The above mentioned problem is one the reason we have developed TRANSIT. Thus, with TRANSIT, as we know accurately the origins and the destinations of the trips, the tensor can be build based on a fine space segmentation which improves the granularity of the inferred urban dynamics. In addition, there is a lack of work in the literature which compare the different tensor factorization approaches.

5.4 Methodological Background

5.4.1 Problem Formulation

In this section we will present the theoretical background of the core method of this chapter: sparse non negative Tucker decomposition. The approach is based on the one proposed by Wang *et al.* [122] .

The following will present the main notations used in this chapter. Let us assume that an urban region can be divided into M zones with N_{trips} trips occurring in the region during a day. The day can also be divided into N time slices. Let us denote m_{xyz} the number of trips from an origin zone $x \in 1, \ldots, M$ to a destination zone $y \in 1, \ldots, M$ with the starting time of the trip taking place within a time slice $z \in 1, \ldots, N$. A third order tensor $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{M \times M \times N}$ is then defined by having m_{xyz} as the generic (x, y, z) element of the tensor. Thus, we have $\sum_{x,y,z} m_{xyz} = N_{trips}$. The tensor $\boldsymbol{\mathcal{M}}$ is called mobility tensor.

Tucker decomposition decomposes a tensor into a smaller tensor with predetermined dimensions $\boldsymbol{\mathcal{C}}$ (called the core tensor), multiplied by factor matrices $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$ along each dimension, given by:

$$\mathcal{M} = \mathcal{C} \times_o \mathbf{O} \times_d \mathbf{D} \times_t \mathbf{T} + \mathcal{E} \tag{5.1}$$

where $\mathcal{E} \in \mathbb{R}^{M \times M \times N}$ is a random error tensor, and $\times_n$ denotes the tensor n-mode product. $\mathcal{M}$ is the mobility tensor as described above. $\mathbf{O} \in \mathbb{R}^{M \times I}$, $\mathbf{D} \in \mathbb{R}^{M \times J}$ and $\mathbf{T} \in \mathbb{R}^{N \times K}$ are the factor matrices and represent the latent urban dynamics of each dimension: origin, destination and time. $\mathcal{C} \in \mathbb{R}^{I \times J \times K}$ is the core tensor and capture the multimodal dynamics of the factor matrices. I, J and K are the number of latent structures associated to each factor matrices. An illustration of Tucker decomposition is represented Figure 5.1. $o_i \in \mathbb{R}^M$ for $i \in 1, \ldots, I$ is a vector representing a hidden structure associated to the dimension origin also denoted as origin pattern. Similarly, $d_j \in \mathbb{R}^M$ for $j \in 1, \ldots, J$ is a vector representing a hidden latent structure associated to the dimension destination also denoted as destination pattern. Finally, $t_k \in \mathbb{R}^N$ for $k \in 1, \ldots, K$ is a vector representing a hidden latent structure associated to the dimension time also denoted as temporal pattern.

The reconstructed mobility tensor $\widehat{\mathcal{M}}$ using the core tensor and the factor matrices is given by the following equation:

$$\widehat{\mathcal{M}} = \mathcal{C} \times_o \mathbf{O} \times_d \mathbf{D} \times_t \mathbf{T} \tag{5.2}$$

$\mathcal{M}$ is estimated with the data. The core tensor $\mathcal{C}$ as well as the projection matrices $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$ are unknown variables that have to be estimated. Hence, our task is twofold:

- Infer $\mathcal{C}$, $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$ from $\mathcal{M}$, i.e., reduce the error between $\mathcal{M}$ and $\widehat{\mathcal{M}}$;
- Understand urban dynamics from the analysis of $\mathcal{C}$, $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$.

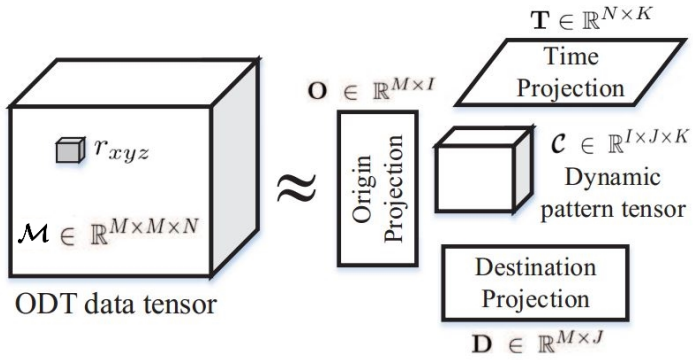


Figure 5.1: Tucker factorization

In the approach proposed by Wang *et al.*, the authors add additional urban-context similarity matrix W to the above mentioned modeling. The idea is to use external information such as point of interest data to help the finding of the latent patterns. Indeed, the point of interest of a city can be classified in different categories (for instance food service, transportation facilities, residence ...). Moreover, we can characterize each zone by its POI distribution in each category. Thus, similar zones will exhibit similar POI distribution. We can build a similarity matrix W where element w_{ij} denotes the similarity between origin i and destination j and computed as the vector product of the two POI distribution of i and j. Based on the reasonable

assumption that similar urban zones should exhibit similar spatial patterns, this similarity matrix can also be estimated with the origin and destination patterns. Let us take the origin case, this matrix will be denoted as W^O and the element w_{ij}^O is computed as the vector product of the latent structure of origin i (the score of all origin patterns associated to origin i) and the latent structure of origin j. With such a definition $W_{ij}^O = \mathbf{O}\mathbf{O}^\top$. For the destination dimension we get $W_{ij}^D = \mathbf{D}\mathbf{D}^\top$. Finally, we have the following relationships between $\mathbf{W}$ and projection matrices $\mathbf{O}$ and $\mathbf{D}$ that can be added to our model :

$$\mathbf{W} = \mathbf{O}\mathbf{O}^\top + \mathbf{E}_O, \; and \; \mathbf{W} = \mathbf{D}\mathbf{D}^\top + \mathbf{E}_D \tag{5.3}$$

where $\mathbf{E}_O$ and $\mathbf{E}_D$ are random error matrices.

The idea is to minimize the error between $\mathbf{W}$ and $\mathbf{O}\mathbf{O}^\top$ as well as the error between $\mathbf{W}$ and $\mathbf{D}\mathbf{D}^\top$

5.4.2 Non-Negative Tucker Decomposition

For building our Tucker factorization model, we made multiple assumptions.

1. The residual tensor $\boldsymbol{\mathcal{E}}$ follows a normal distribution: $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, \sigma_{\mathcal{M}}^2)$
2. we assume that the factor matrices and the core tensor follows zero mean Laplace distribution
3. the observed data cells are mutually independent

Given that hypothesis *1*-st and *3*, the conditional distribution over the observed data in $\boldsymbol{\mathcal{M}}$ is defined as:

$$\begin{aligned} P(\boldsymbol{\mathcal{M}}|\boldsymbol{\mathcal{C}}, \mathbf{O}, \mathbf{D}, \mathbf{T}, \sigma_{\mathbf{M}}^2) &= \prod_{x=1}^{M}\prod_{y=1}^{M}\prod_{z=1}^{N} \mathcal{N}(m_{xyz}|\boldsymbol{\mathcal{C}} \times_o \mathbf{o}_x \times_d \mathbf{d}_y \times_t \mathbf{t}_z, \sigma_{\mathbf{M}}^2) \\ &= \prod_{x=1}^{M}\prod_{y=1}^{M}\prod_{z=1}^{N} \frac{1}{\sigma_M\sqrt{2\pi}} e^{-\frac{(m_{xyz} - \boldsymbol{\mathcal{C}}\times_o \mathbf{o}_x \times_d \mathbf{d}_y \times_t \mathbf{t}_z)^2}{2\sigma_{\mathbf{M}}^2}} \end{aligned} \tag{5.4}$$

Given that hypothesis *2*-st and *3*, the distribution over the core tensor in $\boldsymbol{\mathcal{C}}$ and the factor matrices $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$ are defined as:

$$P(\mathbf{O}|\sigma_O) = \prod_{x=1}^{M} \mathcal{L}(\mathbf{o}_x|\mathbf{0}, \sigma_O\mathbf{I}_I) = \prod_{x=1}^{M}\prod_{i=1}^{I} \mathcal{L}(o_{xi}|0, \sigma_O) = \prod_{x=1}^{M}\prod_{i=1}^{I} \frac{1}{2\sigma_O} e^{\frac{-|o_{xi}|}{\sigma_O}} \tag{5.5}$$

$$P(\mathbf{D}|\sigma_D) = \prod_{y=1}^{M} \mathcal{L}(\mathbf{d}_y|\mathbf{0}, \sigma_D\mathbf{I}_J) = \prod_{y=1}^{M}\prod_{j=1}^{J} \mathcal{L}(d_{yj}|0, \sigma_D) = \prod_{y=1}^{M}\prod_{j=1}^{J} \frac{1}{2\sigma_D} e^{\frac{-|d_{yj}|}{\sigma_D}} \tag{5.6}$$

$$P(\mathbf{T}|\sigma_T) = \prod_{z=1}^{N} \mathcal{L}(\mathbf{t}_z|\mathbf{0}, \sigma_T\mathbf{I}_K) = \prod_{z=1}^{N}\prod_{k=1}^{K} \mathcal{L}(o_{zk}|0, \sigma_T) = \prod_{z=1}^{M}\prod_{k=1}^{K} \frac{1}{2\sigma_T} e^{\frac{-|t_{zk}|}{\sigma_T}} \tag{5.7}$$

$$P(\boldsymbol{\mathcal{C}}|\sigma_C) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \mathcal{L}(c_{ijk}|0, \sigma_C) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{1}{2\sigma_C} e^{\frac{-|c_{ijk}|}{\sigma_C}} \tag{5.8}$$

The posterior distribution of the pattern variables is according to the Bayes theorem:

$$\begin{aligned} & P(\mathcal{C}, \mathbf{O}, \mathbf{D}, \mathbf{T}|\mathcal{M}) \\ &= \frac{P(\mathcal{M}|\mathcal{C}, \mathbf{O}, \mathbf{D}, \mathbf{T})P(\mathcal{C})P(\mathbf{T})P(\mathbf{D})P(\mathbf{O})}{P(\mathcal{M})} \\ &= K \prod_{x,y,z} e^{-\frac{(m_{xyz} - \mathcal{C} \times_o \mathbf{o}_x \times_d \mathbf{d}_y \times_t \mathbf{t}_z)^2}{2\sigma_{\mathbf{M}}^2}} \prod_{i,j,k} e^{\frac{-|c_{ijk}|}{\sigma_C}} \prod_{x,i} e^{\frac{-|o_{xi}|}{\sigma_O}} \prod_{y,j} e^{\frac{-|d_{yj}|}{\sigma_D}} \prod_{z,k} e^{\frac{-|t_{zk}|}{\sigma_T}} \end{aligned} \tag{5.9}$$

where $K = \frac{1}{(\sigma_M\sqrt{2\pi})^{M\times M\times N}(2\sigma_O)^{M\times I}(2\sigma_D)^{M\times J}(2\sigma_T)^{N\times K}(2\sigma_C)^{I\times J\times K}P(\mathcal{M})}$

The log posterior distribution is then calculated by:

$$\begin{aligned} \ln P(\mathcal{C}, \mathbf{O}, \mathbf{D}, \mathbf{T}|\mathcal{M}) &= \ln K + \sum_{x,y,z} \ln e^{-\frac{(m_{xyz} - \mathcal{C} \times_o \mathbf{o}_x \times_d \mathbf{d}_y \times_t \mathbf{t}_z)^2}{2\sigma_{\mathbf{M}}^2}} + \\ & \sum_{i,j,k} \ln e^{\frac{-|c_{ijk}|}{\sigma_C}} + \sum_{x,i} \ln e^{\frac{-|o_{xi}|}{\sigma_O}} + \sum_{y,j} \ln e^{\frac{-|d_{yj}|}{\sigma_D}} + \sum_{z,k} \ln e^{\frac{-|t_{zk}|}{\sigma_T}} \\ &= \ln K - \sum_{x,y,z} \frac{(m_{xyz} - \mathcal{C} \times_o \mathbf{o}_x \times_d \mathbf{d}_y \times_t \mathbf{t}_z)^2}{2\sigma_{\mathbf{M}}^2} - \\ & \sum_{i,j,k} \frac{|c_{ijk}|}{\sigma_C} - \sum_{x,i} \frac{|o_{xi}|}{\sigma_O} - \sum_{y,j} \frac{|d_{yj}|}{\sigma_D} - \sum_{z,k} \frac{|t_{zk}|}{\sigma_T} \end{aligned} \tag{5.10}$$

Thus:

$$\begin{aligned} \ln P(\mathcal{C}, \mathbf{O}, \mathbf{D}, \mathbf{T}|\mathcal{M}) \propto & -\frac{1}{2\sigma_{\mathbf{M}}^2} \sum_{x,y,z} (m_{xyz} - \mathcal{C} \times_o \mathbf{o}_x \times_d \mathbf{d}_y \times_t \mathbf{t}_z)^2 - \\ & \frac{1}{\sigma_C} \sum_{i,j,k} |c_{ijk}| - \frac{1}{\sigma_O} \sum_{x,i} |o_{xi}| - \frac{1}{\sigma_D} \sum_{y,j} |d_{yj}| - \frac{1}{\sigma_T} \sum_{z,k} |t_{zk}| \end{aligned} \tag{5.11}$$

Therefore, to obtain the *Maximum A Posteriori (MAP)* estimation of $\mathcal{C}$, $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$ is equivalent to minimizing the object function:

$$\begin{aligned} \tilde{\mathcal{J}} = & \frac{1}{2\sigma_{\mathbf{M}}^2} \|\mathcal{M} - \mathcal{C} \times_o \mathbf{O} \times_d \mathbf{D} \times_t \mathbf{T}\|_F^2 + \\ & \frac{1}{\sigma_C} \|\mathcal{C}\|_1 + \frac{1}{\sigma_O} \|\mathbf{O}\|_1 + \frac{1}{\sigma_D} \|\mathbf{D}\|_1 + \frac{1}{\sigma_T} \|\mathbf{T}\|_1 \end{aligned} \tag{5.12}$$

where $\|\cdot\|_F^2$ is the Frobenius norm and $\|\cdot\|_1$ the L-1 norm.

We introduce urban contextual factors as context-aware regularization using a maximum *a posteriori* method. Assume the elements of $\mathbf{E}_O$ and $\mathbf{E}_D$ in Eq. [Ref eq] follow zero-mean Gaussian distributions, then we have:

$$P(\mathbf{W}|\mathbf{O},\sigma_{WO}^2)=\prod_{p=1}^{M}\prod_{q=1}^{M}\mathcal{N}(w_{pq}|\mathbf{o}_p\mathbf{o}_q^\top,\sigma_{WO}^2)=\prod_{p=1}^{M}\prod_{q=1}^{M}\frac{1}{\sigma_{WO}\sqrt{2\pi}}e^{\frac{-(w_{pq}-\mathbf{o}_p\mathbf{o}_q^\top)^2}{2\sigma_{WO}^2}} \tag{5.13}$$

$$P(\mathbf{W}|\mathbf{D},\sigma_{WD}^2)=\prod_{p=1}^{M}\prod_{q=1}^{M}\mathcal{N}(w_{pq}|\mathbf{d}_p\mathbf{d}_q^\top,\sigma_{WD}^2)=\prod_{p=1}^{M}\prod_{q=1}^{M}\frac{1}{\sigma_{WD}\sqrt{2\pi}}e^{\frac{-(w_{pq}-\mathbf{d}_p\mathbf{d}_q^\top)^2}{2\sigma_{WD}^2}} \tag{5.14}$$

Let $\Omega=\{\sigma_{\mathcal{M}}^2,\sigma_{WO}^2,\sigma_{WD}^2,\sigma_O,\sigma_D,\sigma_T,\sigma_C\}$. Given the data tensor $\boldsymbol{\mathcal{M}}$ and urban context matrix $\mathbf{W}$, the posterior distribution of $\mathbf{O}$, $\mathbf{D}$, $\mathbf{T}$ and $\boldsymbol{\mathcal{C}}$ is given by:

$$\begin{aligned}
&\ln P(\boldsymbol{\mathcal{C}},\mathbf{O},\mathbf{D},\mathbf{T}|\boldsymbol{\mathcal{M}},\mathbf{W},\Omega)\\
&\propto \ln P(\boldsymbol{\mathcal{M}}|\boldsymbol{\mathcal{C}},\mathbf{O},\mathbf{D},\mathbf{T})P(\boldsymbol{\mathcal{C}},\Omega)P(\mathbf{W}|\mathbf{O},\Omega)P(\mathbf{W}|\mathbf{D},\Omega)P(\mathbf{O}|0,\Omega)P(\mathbf{D}|0,\Omega)\\
&P(\mathbf{T}|0,\Omega)P(\boldsymbol{\mathcal{C}}|0,\Omega)\\
&\propto -\frac{1}{2\sigma_{\mathbf{M}}^2}\sum_{x,y,z}(m_{xyz}-\boldsymbol{\mathcal{C}}\times_o\mathbf{o}_x\times_d\mathbf{d}_y\times_t\mathbf{t}_z)^2-\\
&\frac{1}{2\sigma_{WO}^2}\sum_{p,q}(w_{pq}-\mathbf{o}_p\mathbf{o}_q^\top)^2-\frac{1}{2\sigma_{WD}^2}\sum_{p,q}(w_{pq}-\mathbf{d}_p\mathbf{d}_q^\top)^2-\\
&\frac{1}{\sigma_C}\sum_{i,j,k}|c_{ijk}|-\frac{1}{\sigma_O}\sum_{x,i}|o_{xi}|-\frac{1}{\sigma_D}\sum_{y,j}|d_{yj}|-\frac{1}{\sigma_T}\sum_{z,k}|t_{zk}|
\end{aligned} \tag{5.15}$$

To maximize the posterior distribution is equivalent to minimizing the sum-of-squared errors function with hybrid quadratic regularization terms, *i.e.*,

$$\begin{aligned}
\tilde{\mathcal{J}}=&\frac{1}{2\sigma_{\mathbf{M}}^2}\left\|\boldsymbol{\mathcal{M}}-\boldsymbol{\mathcal{C}}\times_o\mathbf{O}\times_d\mathbf{D}\times_t\mathbf{T}\right\|_F^2+\\
&\frac{1}{2\sigma_{WO}^2}\left\|\mathbf{W}-\mathbf{O}\mathbf{O}^\top\right\|_F^2+\frac{1}{2\sigma_{WD}^2}\left\|\mathbf{W}-\mathbf{D}\mathbf{D}^\top\right\|_F^2+\\
&\frac{1}{\sigma_C}\left\|\boldsymbol{\mathcal{C}}\right\|_1+\frac{1}{\sigma_O}\left\|\mathbf{O}\right\|_1+\frac{1}{\sigma_D}\left\|\mathbf{D}\right\|_1+\frac{1}{\sigma_T}\left\|\mathbf{T}\right\|_1
\end{aligned} \tag{5.16}$$

Which is equivalent to minimize the following equation :

$$\begin{aligned}
\tilde{\mathcal{J}}=&\left\|\boldsymbol{\mathcal{M}}-\boldsymbol{\mathcal{C}}\times_o\mathbf{O}\times_d\mathbf{D}\times_t\mathbf{T}\right\|_F^2+\\
&\alpha\left\|\mathbf{W}-\mathbf{O}\mathbf{O}^\top\right\|_F^2+\beta\left\|\mathbf{W}-\mathbf{D}\mathbf{D}^\top\right\|_F^2+\\
&\gamma\left\|\boldsymbol{\mathcal{C}}\right\|_1+\delta\left\|\mathbf{O}\right\|_1+\epsilon\left\|\mathbf{D}\right\|_1+\zeta\left\|\mathbf{T}\right\|_1\\
&\text{and }\boldsymbol{\mathcal{C}}\geq 0,\mathbf{O}\geq 0,\mathbf{D}\geq 0,\mathbf{T}\geq 0
\end{aligned} \tag{5.17}$$

where $\alpha=\frac{\sigma_{\mathbf{M}}^2}{\sigma_{WO}^2}$, $\beta=\frac{\sigma_{\mathbf{M}}^2}{\sigma_{WD}^2}$, $\gamma=\frac{2\sigma_{\mathbf{M}}^2}{\sigma_C}$, $\delta=\frac{2\sigma_{\mathbf{M}}^2}{\sigma_O}$, $\epsilon=\frac{2\sigma_{\mathbf{M}}^2}{\sigma_D}$ and $\zeta=\frac{2\sigma_{\mathbf{M}}^2}{\sigma_T}$

We adopt the Block Coordinate Descent-Proximal Gradient (BCD-PG) algorithm developped by Xu *et al.* [131] to solve the equation 5.17. While this function is not jointly convex with respect to $\mathbf{O}$, $\mathbf{D}$, $\mathbf{T}$ and $\boldsymbol{\mathcal{C}}$, it is block multiconvex with each one

when the other three are fixed. Therefore, we adopt a Block Coordinate Descent (BCD) procedure [131], which starts from an initialization on $\mathcal{C}^{(0)}, \mathbf{O}^{(0)}, \mathbf{D}^{(0)} and \mathbf{T}^{(0)}$, and then iteratively updates $\mathcal{C}^{(s)}, \mathbf{O}^{(s)}, \mathbf{D}^{(s)} and \mathbf{T}^{(s)}$ with $s \in \mathbb{N}$ by:

$$\begin{aligned}
\mathcal{C}^{(s)} &= \underset{\mathcal{C}}{\operatorname{argmin}}\, \mathcal{J}\left(\mathcal{C}, \mathbf{O}^{(s-1)}, \mathbf{D}^{(s-1)}, \mathbf{T}^{(s-1)}\right) + \gamma \|\mathcal{C}\|_1 \\
\mathbf{O}^{(s)} &= \underset{\mathcal{C}}{\operatorname{argmin}}\, \mathcal{J}\left(\mathcal{C}^{(s)}, \mathbf{O}, \mathbf{D}^{(s-1)}, \mathbf{T}^{(s-1)}\right) + \delta \|\mathbf{O}\|_1 \\
\mathbf{D}^{(s)} &= \underset{\mathcal{C}}{\operatorname{argmin}}\, \mathcal{J}\left(\mathcal{C}^{(s)}, \mathbf{O}^{(s)}, \mathbf{D}, \mathbf{T}^{(s-1)}\right) + \epsilon \|\mathbf{D}\|_1 \\
\mathbf{T}^{(s)} &= \underset{\mathcal{C}}{\operatorname{argmin}}\, \mathcal{J}\left(\mathcal{C}^{(s)}, \mathbf{O}^{(s)}, \mathbf{D}^{(s)}, \mathbf{T}\right) + \zeta \|\mathbf{T}\|_1
\end{aligned} \tag{5.18}$$

The Block Coordinate Descent approach implemented to solve the optimization problem defined in equation 5.17 is detailed in Appendix A.

5.5 Case Study

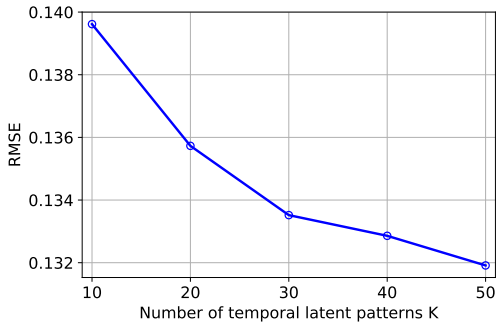
We leverage the results of TRANSIT to build daily mobility tensors. With TRANSIT, we are able to estimate accurately the origin and the destination of the trips. Thus, a fine grained spatial segmentation can be used. Indeed, in the studied Lyon area, we divided the area into $N = 625$ zones $(1, \cdots, M)$, each zone being a 800m$\times$800m square. In time, we can be divided into $M = 20$ slots, $(1, \cdots, N)$ (one slot per hour). The four missing slots correspond to slots before 4 a.m when the demand is very low. For each trip resulting from TRANSIT, we can map the time origin on $(1, \cdots, N)$, the origin and the destination to $(1, \cdots, M)$. By aggregating the results on all trips, we end up with the daily mobility tensor $\mathbf{M}$.

5.5.1 Parameter Choice

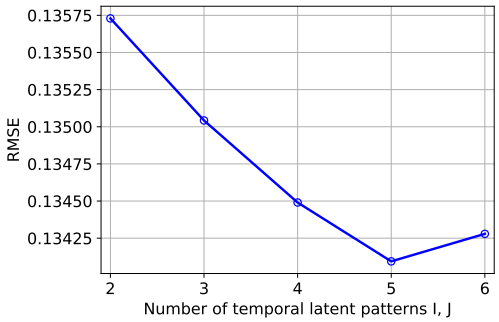
First, we have to determine the number of spatial patterns (I origin patterns and J destination patterns) and the number K of temporal patterns. This choice has to be a trade-off between the error of the approach and the complexity/interpretability of the urban dynamics. The error is measured as measures as the Root Mean Square Error (RMSE) between $\mathcal{M}$ and $\widehat{\mathcal{M}}$. If the number of patterns is high, the error model will be low but the number of latent patterns will be high and interpreting the model to extract knowledge on urban dynamics will be difficult. On the contrary, if the number of patterns is too low, the error of the model will be high and the latent structure inferred by the model will be trivial. Thus, we apply our Tucker decomposition approach for different numbers of spatial and temporal patterns (I,J and K) and compute the RMSE on each decomposition. We made 5 trials of our decomposition for I,J varying from 10 to 50 and K varying from 2 to 6. We kept the best model (in terms of loss function $\mathcal{J}$) among the 5 trials and consolidated the results by averaging the RMSE obtained by applying our approach on 3 different days of data (18^{th}, 19^{th} and 20^{th} of March, 2019). For simplification we considered the same number of origin and destination patterns, *i.e.*, $I = J$. The sensitivity analysis of $I = J$ and K are given in Figure 5.2a and Figure 5.2b. We can observe that the RMSE have a

decreasing trend when the number of patterns either spatial or temporal, increases. The results are consistent with what has been obtained in the literature [122]. Based on the latter results and a manual investigation of the decomposition we set $I = J$ to 20 and K to 4. Selecting a number of patterns higher than the chosen one would make the interpretation of the decomposition too difficult.

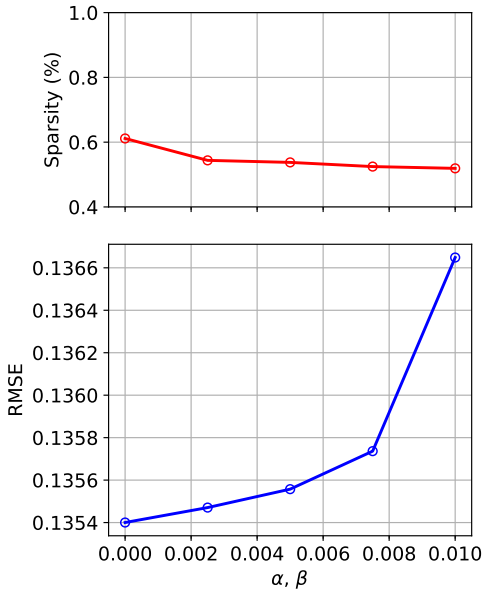
Then, we have to choose the best parameters for our approach and in particular our optimization problem. We have two types of parameters: the urban context parameters α and β which aims at obtaining both origin and destination similarity matrices close to the ones obtained with external data (POI), $\mathbf{W}$. The second type of parameters refers to the regularization terms: γ, δ, ϵ and ζ whose purpose is that our decomposition: $\mathcal{C}$, $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$ are as sparse as possible. We made 5 trials of our decomposition for α and β varying from 0 to 0.01 and γ, δ, ϵ and ζ varying from 0 to 10. We kept the best model (in terms loss function $\mathcal{J}$) among the 5 trials and consolidated the results by averaging the RMSE obtained on 3 days of data. For simplification matters we considered $\alpha = \beta$ and $\gamma = \delta = \epsilon = \zeta$.



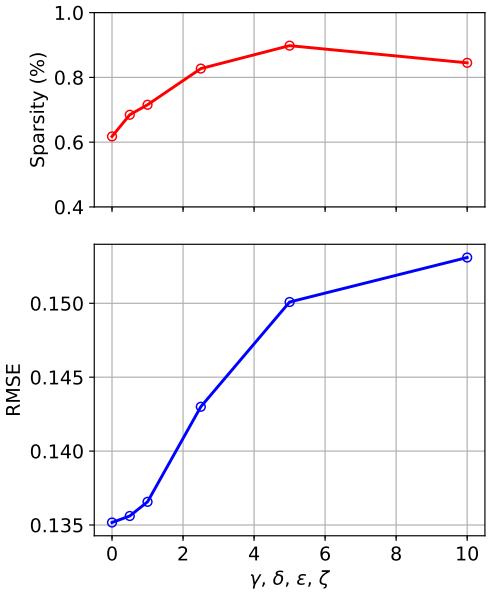
(a) Sensitivity analysis of K



(b) Sensitivity analysis of I, J



(c) Sensitivity analysis of α, β



(d) Sensitivity analysis of γ, δ, , ϵ and ζ

Figure 5.2: Sensitivity analysis on the dimension of the Tucker decomposition (α and β) and on the parameters of the approach (γ, δ, , ϵ and ζ).

The sensitivity analysis of $\alpha = \beta$ and $\gamma = \delta = \epsilon = \zeta$ are given in Figure 5.2c and Figure 5.2d. We can observe that the RMSE have an increasing trend for both context and regularization parameters. This result is not in line with those obtained by Xu *et al.* [122]. For both curves they obtained a RMSE decreasing and then increasing. Thus, the authors pick the minimum of the curve for choosing their parameters. However, such a result is surprising given the nature of the optimization problem. Indeed, the urban context and the regularization terms in the optimization of the problem can be seen as a constraint of the problem which forces the optimization to deviate from the only minimization error between $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ to also consider the urban context and regularization term. The latter allows to have a better interpretability of the results: the regularization allows to minimize the number of non-zero elements for the tensor and factor matrices of the decomposition and the urban context term ensures that the obtained origin and destination patterns respect properties of trusted ground truth data. Thus, it seems logical for us to have an increasing RMSE when the weight in the optimization problem given to the urban context and the regularization terms increases. In addition to study the RMSE, we also study the sparsity of our model (percentage of zero elements from all the elements of $\boldsymbol{\mathcal{C}}$, $\mathbf{O}$, $\mathbf{D}$ and $\mathbf{T}$). We aim at minimizing the RMSE while maximizing the sparsity. We reasonnably assume that decomposition with fewer number of elements will be easier to interpret. The urban context parameters do not have a strong effect on the sparsity whereas the regularization parameters allow to significantly increase the sparsity of the model. After an investigation on the results obtained by varying urban context terms we do not observe significant changes in the patterns obtained. Given the latter observation and the above mentioned sensitivity analysis, we set $\alpha = \beta$ to 0. Here, we nuance the result obtained by Xu *et al.* on the added value of urban context term for the Tucker decomposition. Concerning the regularization term, an investigation of the result reveals that a slight increase of the regularization term (0, 0.5 or 1) allow to improve the interpretability of the result with a RMSE remaining low. However, if we keep increasing this term, the obtained patterns are too sparse, the RMSE increase significantly and the obtained patterns become inconsistent. Given the latter observation and the sensitivity analysis we set $\gamma = \delta = \epsilon = \zeta$ to 1.

5.5.2 Daily Decomposition and Analysis

After choosing the parameters of our Tucker decomposition approach, in this section we aims at showing and analyzing the results of this decomposition for one day of data (the 20^{th} of March, 2019). The results of the decomposition is the following: the factor matrix $\mathbf{T} \in \mathbb{R}^{20\times 4}$ which represents the 4 hidden temporal patterns, the factor matrix $\mathbf{O} \in \mathbb{R}^{625\times 20}$ which represents the 20 hidden origin patterns, the factor matrix $\mathbf{D} \in \mathbb{R}^{625\times 20}$ which represents the 20 hidden destination patterns and the core tensor $\mathbf{C} \in \mathbb{R}^{4\times 20\times 20}$ which represents the hidden interactions between the above mentioned patterns.

The four hidden temporal patterns are represented in Figure 5.3. The results show that our approach is able to retrieve latent temporal patterns such as commuting trips which are are well known in the transportation community. The morning peak is centered at 8a.m whereas the afternoon peak is centered at 6p.m. We can also observe that the morning peak is sharper than the afternoon peak. This reveals that the morning peak of the commuting demand is sharper and of shorter duration compared to the afternoon peak. Such a phenomenon is also well known by the transportation experts community. The third temporal pattern is largely spread

over time and with lower intensity compared to the commuting temporal patterns. This corresponds to a demand between morning and afternoon peaks with relatively high values during lunchtime. The last temporal pattern has high values after 6 p.m and seems to correspond to leisure activities after working.

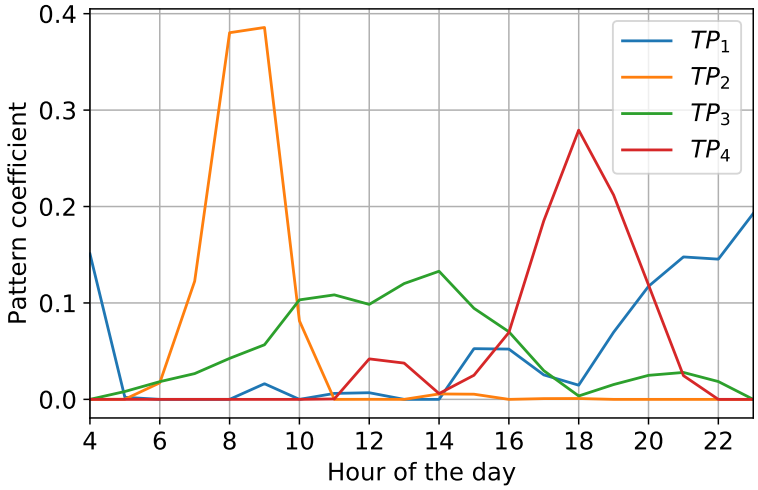
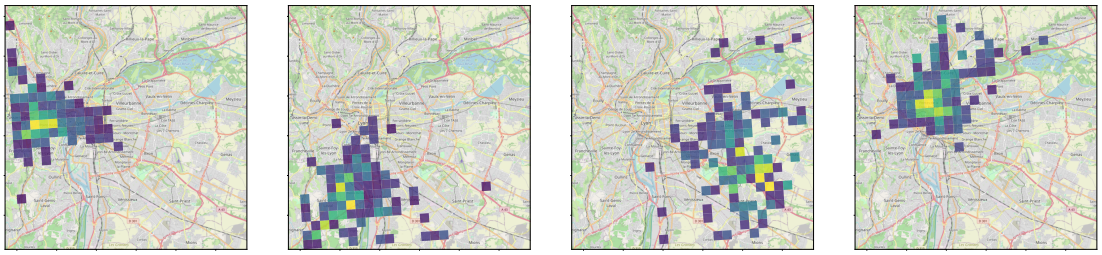


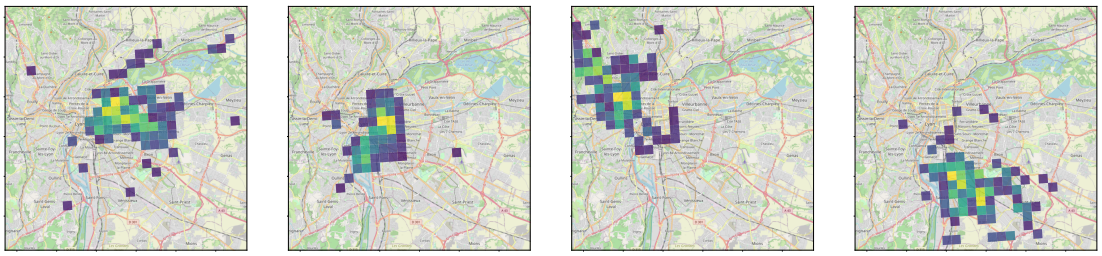
Figure 5.3: Hidden temporal patterns

The twenty hidden origin patterns are represented in Figure 5.4. The cells with yellow color corresponds to high values in the factor matrices and blue ones corresponds to low values. The first very interesting result is that our approach is able to infer consistent origin patterns i.e high values for cells origins which are close to each other. Usually, and especially for the surveys, based on expert knowledge, a spatial segmentation is defined to analyze the mobility. Here the approach is different, we retrieve the origin communities (pattern) that are the most relevant given the mobility tensor. The use of fine spatial segmentation (800m$\times$800m) that TRANSIT made possible, allow to improve the spatial granularity of the inferred origin patterns. This shows the capability of our approach to optimize the granularity of the origin patterns and adapt it to the density of the trips in space. Areas with a high density of trips will result with smaller origin patterns. Indeed, our approach infer small communities for origin trips within the city-center (origin patterns 6, 11, 13, 15, 17 for instance) and large communities for trips in the suburb of the city (origin patterns 1, 2, 4, 8, 12). Moreover, our approach is able to retrieve communities centered on the main stations of the city: origin pattern 6 corresponds to Part-Dieu station (the biggest station in Lyon) and origin pattern 13 corresponds to Perrache station (the second biggest station in Lyon). Another interesting result is the distribution of the values for each origin pattern. We can observe a small area with high values in the barycenter of the origin pattern and then decreasing values when the distance between cells and the barycenter increases. Such a representation allows to have a better knowledge of trips patterns compared to traditional approach which consist on having a coarse spatial segmentation with the assumption that the trips from an area are uniformly distributed within the area. Finally, there are some origin patterns that do not form a continuous cells in space such as origin patterns 16, 18 and 20. The origin patterns in Figure 5.4 are sorted by *order of importance* in the decomposition (by summing all coefficients for each origin pattern we get a score that indicates the importance of the origin pattern in the decomposition). The rank 16,

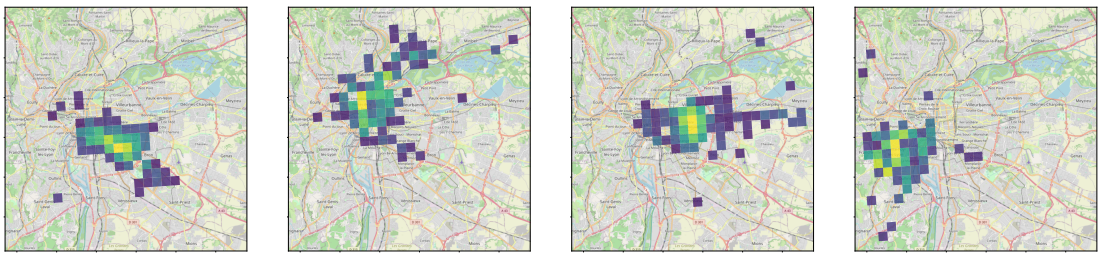
18 and 20 of the above mentioned origin patterns suggests that these origin patterns are not essential for reconstructing the mobility tensor, *i.e.*, they have a minor role in the urban dynamics of the city.



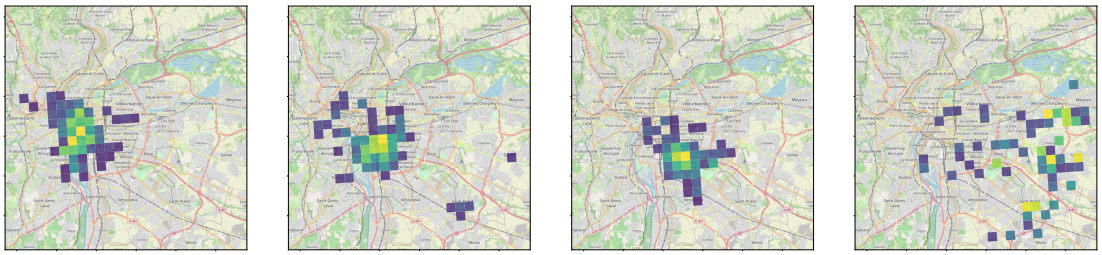
(a) Origin pattern 1 (b) Origin pattern 2 (c) Origin pattern 3 (d) Origin pattern 4



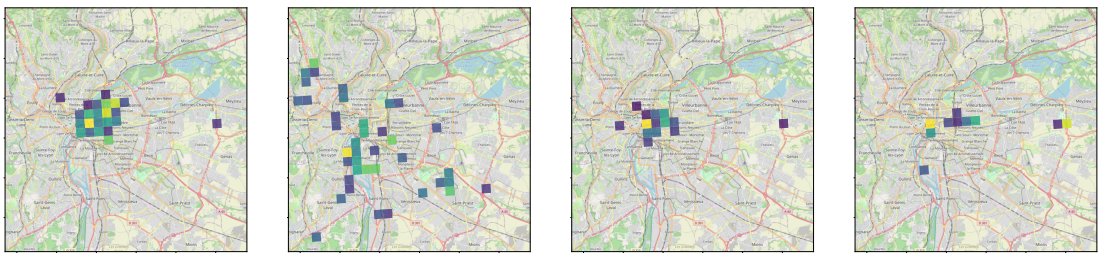
(e) Origin pattern 5 (f) Origin pattern 6 (g) Origin pattern 7 (h) Origin pattern 8



(i) Origin pattern 9 (j) Origin pattern 10 (k) Origin pattern 11 (l) Origin pattern 12

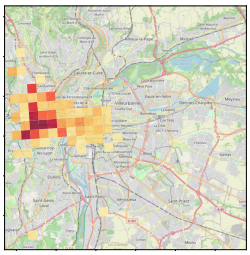


(m) Origin pattern 13 (n) Origin pattern 14 (o) Origin pattern 15 (p) Origin pattern 16

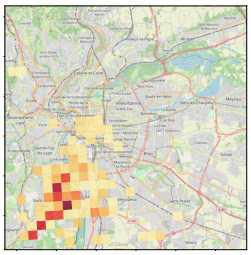


(q) Origin pattern 17 (r) Origin pattern 18 (s) Origin pattern 19 (t) Origin pattern 20

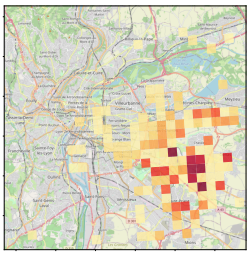
Figure 5.4: Hidden origin patterns - Lyon



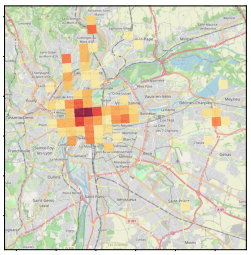
(a) Destination pattern 1



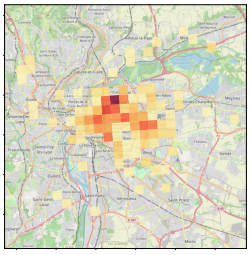
(b) Destination pattern 2



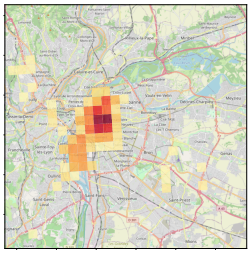
(c) Destination pattern 3



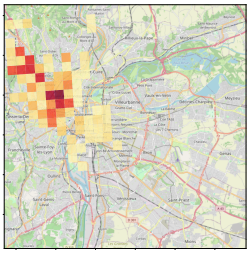
(d) Destination pattern 4



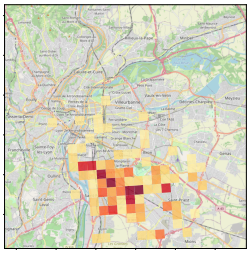
(e) Destination pattern 5



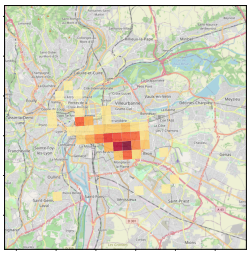
(f) Destination pattern 6



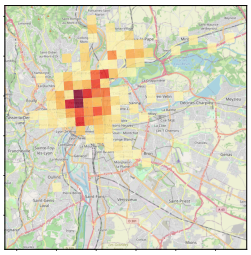
(g) Destination pattern 7



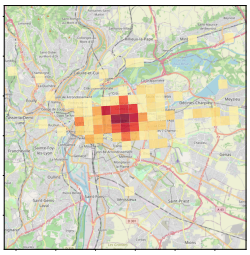
(h) Destination pattern 8



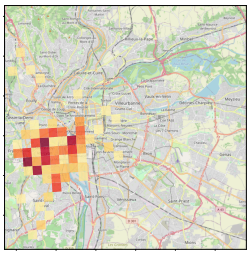
(i) Destination pattern 9



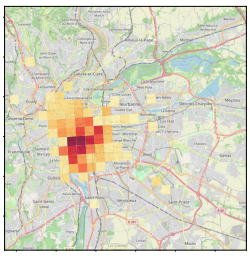
(j) Destination pattern 10



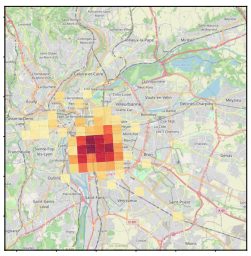
(k) Destination pattern 11



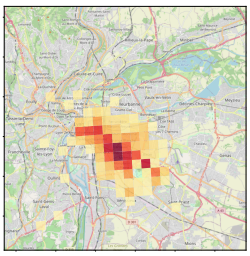
(l) Destination pattern 12



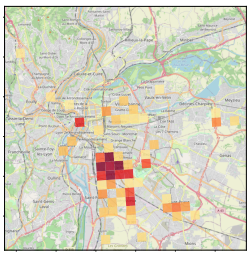
(m) Destination pattern 13



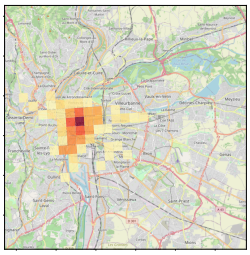
(n) Destination pattern 14



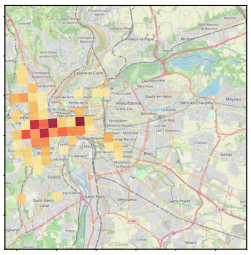
(o) Destination pattern 15



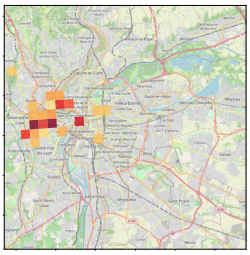
(p) Destination pattern 16



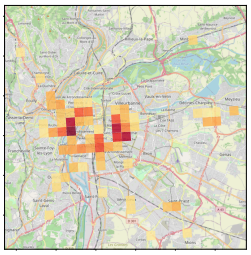
(q) Destination pattern 17



(r) Destination pattern 18



(s) Destination pattern 19

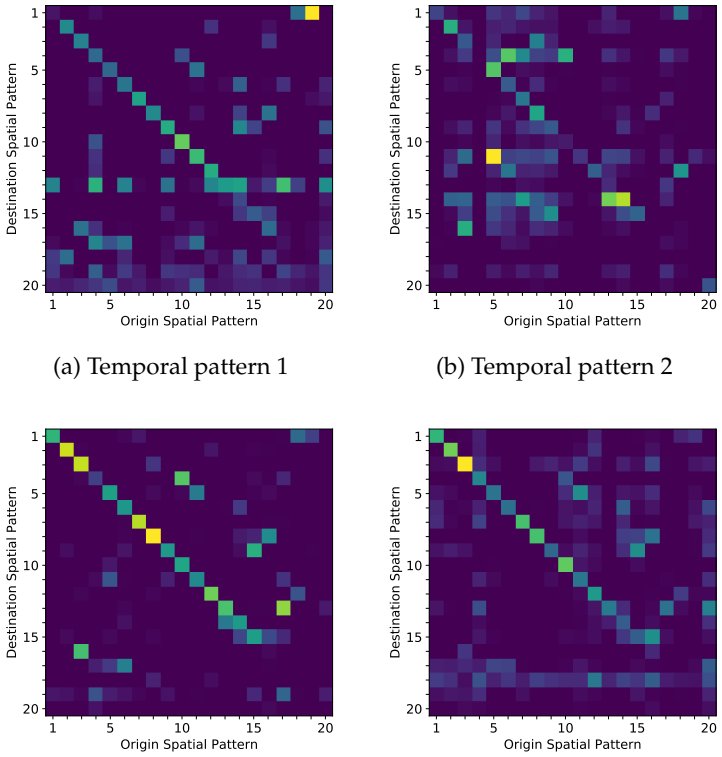


(t) Destination pattern 20

Figure 5.5: Hidden destination patterns - Lyon

The twenty hidden destination patterns are represented in Figure 5.5. The cells with red color corresponds to high values in the factor matrices and light orange

ones corresponds to low values. We can observe that there are strong similarities between the hidden origin patterns and the hidden destination patterns. This can be explained by the commuting pattern observed in the temporal patterns: the origin of the home to work trip is also the destination of the work to home trip and the destination of the home to work trip is also the origin of the work to home trip. In that context, it seems consistent to obtain very similar spatial communities for origin and destination patterns. Thus, the analysis of origin patterns done in the previous paragraph is also valid for the destination patterns.



(a) Temporal pattern 1

(b) Temporal pattern 2

(c) Temporal pattern 3

(d) Temporal pattern 4

Figure 5.6: Core tensor coefficient

After analyzing the factor matrices: temporal, origin and destination patterns, we will analyze the coefficient of the core tensor. This core tensor captures the interactions, dynamics between these patterns. The number of elements of the core tensor is $4 * 20 * 20 = 1600$ elements. The elements of the core tensor are represented in Figure 5.6 using 4 slices along the temporal dimension. The cells with yellow color corresponds to high values in the core tensor whereas blue ones corresponds to low values. The interactions between origin and destination patterns for the temporal pattern 1 is represented Figure 5.6a. As analyzed before, the temporal pattern 1 correspond to high demand of trips during the evening which seems to relate to leisure activities. Indeed, we can observe strong interactions between residential areas (original patterns 13, 19, 17) and leisure areas with a lot of bars (destination patterns 4, 14, 11). The interactions between origin and destination patterns for the temporal pattern 3 is represented Figure 5.6c.The temporal pattern 3 which correspond of trips between morning and afternoon peak is characterized with many origin/destination interaction in the diagonal which means that the trips stays within the same community. This indicates that for temporal pattern 3, the

trips are mainly short distance trips. The interactions between origin and destination patterns for the temporal pattern 2 and 4 are represented respectively Figure 5.6b and Figure 5.6d.The phenomena observed for temporal pattern 3, is also present for temporal pattern 2 but with lower intensity. For temporal pattern 2, we can observe longer distance origin/destination interaction. Indeed, strong interactions between the city center (origin pattern 4, 11, 14) and working area (destination pattern 8) school area (destination pattern 5) or area with a station which is also a working area (destination pattern 6). Most of the above interactions are also observed for temporal pattern 4. Different from temporal pattern 2, the destination pattern 4 is a strong destination which attracts surrounding origin patterns (origin pattern 6, 7, 13).

5.5.3 Comparative Analysis

In this section, we benchmark the performance of multiple tensor decomposition approaches: Regularized Non-negative Tucker Decomposition (R-NTF) which is the approach proposed by Xu *et al.* [122] and that we implemented in this chapter, Non-negative Tucker Decomposition which is the approach developed by Sun *et al.* [106] (NTF), Candecom-Parafac decomposition with a rank equal to 20 (CP-20) and Candecom-Parafac decomposition with a rank equal to 4 (CP-4). The difference between R-NTF and NTF, is that R-NTF leverages regularization and use alternating proximal gradient approach for solving the optimization problem. Instead, NTF do not rely on regularization and multiplicative algorithm is used for solving the optimization problem. Alternating proximal gradient algorithm that we implemented in this chapter can be found in Appendix A and details about the algorithm can be found in the work by Xu *et al.* [131]. Details about the Candecom-Parafac decomposition as well as multiplicative approach can be found in the article of Kolda *et al.* [64] which reviews the different tensor decomposition approaches in the literature.

Approach	Metric	50%	75%	100%
R-NTF [122]	RMSE / Sparsity (%)	0.101 / 78	0.118 / 75	0.131 / 72
NTF [106]	RMSE	0.100	0.116	0.129
CP-20	RMSE	0.100	0.116	0.128
CP-4	RMSE	0.107	0.125	0.140

Table 5.2: Comparative analysis of multiple tensor decomposition approaches R-NTF, NTF, CP-20 and CP-4 under different sampling ratio of the daily mobility tensor 50%, 75% and 100%

We also stressed the performance of the latter approaches under different sampling ratio of the daily mobility tensor 50%, 75% and 100% (which consist on removing randomly a certain percentage of the total trips before building the daily mobility tensor). The result of this benchmark is given Table 5.2. We can observe that the RMSE of CP-20 and NTF is almost the same whatever the sampling ratio used. We can also notice that the RMSE of the R-NTF is slightly higher compared to the above mentioned methods. However, whereas the sparsity of NTF and CP-20 is 0%, the sparsity of R-NTF is around 75%. This means that R-NTF needs a quarter of the number of parameters of NTF and CP-20 for a small additional error. As seen in Section 5.5.1, we accept to slightly increase the RMSE to reduce the number

of parameters and have a better explainability of the model. The last model CP-4 performs significantly worse than the other decomposition approaches. In addition, we can notice that the sparsity of the model decreases for R-NFT as the sampling ratio increases. It seems that the model is able to adapt the number of parameters of the model to the complexity of the daily mobility tensor in input. When the sampling ratio is equal to 50%, there is less information in the mobility tensor and the model has a greater sparsity compared to the 100% sampling ratio case where the full information is available.

Finally, we also explore the capability of R-NFT to retrieve correct urban dynamics with different sampling ratio. The temporal, origin and destination patterns obtained in the worst case, *i.e.*, 50% sampling ratio, for one day (18^th March, 2019) is given in Appendix B. The results show that even with 50% sampling ratio of the initial daily mobility tensor, R-NFT is able to retrieve the main urban dynamics compared to 100% sampling ratio case analyzed in Section 5.5.2. Thus, R-NFT is robust to sampling effect and the results advocate for the use of mobile phone data to capture the urban dynamics of the city even if the penetration rate of the mobile phone dataset do not allow to cover the whole population but a significant fraction of it. In our case, the market share of Orange is at 37% over the French territory.

5.6 Conclusion

In this chapter, we investigated the problem of urban dynamics discovery and use as a case study the daily mobility in the city of Lyon. We model the mobility within the city by an agnostic mathematical representation: a tensor. This tensor allow to represent the daily mobility at aggregated scale along multiple dimensions. Here, we selected three dimensions: time, origin and destination. Thanks to TRANSIT approach from Chapter 3 we are able to characterize this tensor at fine spatio-temporal granularity. In particular, the spatial segmentation used is a spatial grid composed of 625 squares, each square having the dimension of 800m$\times$800m. Such a square based segmentation is significantly finer compared to usual administrative spatial segmentation (as for instance the one used in Chapter 2).

To extract the urban dynamics of the city based on the daily mobility tensor, we rely on non-negative Tucker decomposition approach able to decompose the mobility tensor into a smaller tensor with predetermined dimensions (the core tensor), multiplied by factor matrices. The factor matrices are able to capture the latent urban dynamics of each dimension (origin, destination and time) whereas the core tensor grab the interactions between the factor matrices.

The approach implemented consist on an Non-Negative sparse Tucker Decomposition which is coupled with regularization. The optimization method used is the alternating proximal gradient. The results contained in this chapter allow to mitigate the role of regularization and the integration of urban factors for improving the decomposition compared to the result of litterature [122]. Besides, we demonstrate the capability of the approach in conjunction with fine-grained tensor computed from TRANSIT outputs to capture the latent urban dynamics of the city. Finally, we perform a comparative analysis of the state of the art tensor factorization approaches. While a sightly higher RMSE error compared to other methods, the studied approach have much less parameters. We also showed that the approach is still able to retrieve urban dynamics with a 50% sampled mobility tensor.

Futures works could explore the possibility to add more dimension in the mobility tensor such as the transportation mode to improve the quality of the extracted urban dynamics. Moreover, the mobility tensor could be built using multiple data sources using appropriate data fusion approach. This would allow to have a more accurate representation of the urban mobility.

Chapter 6

Mobile Phone Data: Opportunities and Challenges

In Chapter 3 and Chapter 4, we developed approaches able to overcome NSD limitations first and then infer rich mobility information. Thus, our approaches pave the way of the use of NSD for many new applications that were not possible using methods from the literature. As a proof that we were able to unlock the potential of NSD, we develop several applications including trajectory analysis of users in the ring road in Paris, human mobility analysis during abnormal events or mobility based epidemiological model for modeling COVID-19 propagation. Besides developing new applications in the transportation domain, we also discuss the main threats associated to the use of mobile phone data based for human mobility analysis. The first one is the spatio-temporal bias in the results derived from mobile phone data. The second one is the presence of personal sensitive information which make users' privacy at risk. Thus, we develop preliminary yet effective solution to tackle the second issue.

The chapter is structured as follows. Section 6.3 presents several applications that are made possible by the methods developed in this thesis. Then, Section 6.4 discusses the threats of NSD and preliminary solutions to tackle them. Finally, Section 6.5 presents the conclusion with suggestions for future research directions.

This chapter contains parts of the following articles [37, 78]:

Goel R., **Bonnetain L.**, Sharma R., Furno A., (2021), "Mobility Based SIR Model For Complex Networks –With Case Study Of COVID-19". In: *Social Network Analysis and Mining*.

Matet B., Côme E., Furno A., **Bonnetain L.**, Oukhellou L., El Faouzi N.-E., (2021), "A Lightweight Approach for Origin-Destination Matrix Anonymization". In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

6.1 Notation for this chapter

Symbol	Description
$s(t)$	Fraction of susceptible population to be infected at time t.
$i(t)$	Fraction of infected population at time t.
$r(t)$	Fraction of recovered population at time t.
$n(t)$	Population size at time t.
β	Incidence rate.
μ	Cure rate.
c_{ij}	Flow of individuals from location j to location i.
$S_i(t)$	Fraction of susceptible population to be infected from location i at time t.
$I_i(t)$	Fraction of infected population from location i at time t.
$R_i(t)$	Fraction of recovered population from location i at time t.
$N_i(t)$	Population from location i at time t.
k	Anonymity threshold.
U	Original grid containing $N \times N$ tiles.
n	Single tile.
$\mathcal{Q}$	Set of quadtrees whose root represents the complete study area and whose vertices correspond to a non-empty set of tiles in U.
q	Element of $\mathcal{Q}$.
$\mathcal{L}(q)$	Set of the leaves of q .
$G(\cdot,\cdot)$	General information loss.
S	Suppression threshold.
δ	Cost of suppression coefficient.
MAO	Mean area of origins.
MAD	Mean area of destinations.

Table 6.1: Chapter 6′ specific notations

6.2 Introduction

Our work advances the state of the art of mobile phone data based human mobility analysis. Our work studies a recent kind of mobile phone data: Network Signaling Data. By proposing approaches for overcoming the intrinsic limitation of NSD, we can unlock the potential of the latter. On the one hand, multiple already existing applications can be significantly improved. For instance, for the works related to the validation of the laws that govern human mobility, the studies work either on GPS or CDR. However, each data source bring strong limitations. GPS data allow human mobility analysis at very fine grained but with very limited number of trajectories. The CDR, which are the main source analyzed in the literature allow to model city-scale population but at very coarse spatio-temporal resolution. By applying TRANSIT on raw NSD, we end up with large scale and fine-grained trajectories that allow human mobility modeling at unprecedented spatio-temporal scale. Thus, our approach could improve the models studied in this area of research. Moreover, concerning the works focusing on travel demand estimation. On the one hand, we propose a segmentation approach exhibiting higher accuracy than approaches from the literature. On the other hand, the trajectory enhancement of TRANSIT enables unprecedented spatiotemporal resolution for travel demand inference. Besides, we strengthened the confidence on TRANSIT by validating the travel demand inferred with surveys. Thus, the whole framework allow to improve significantly the existing approaches for travel demand inference. In addition, our work paves the way of new studies along two main directions. The first one, is the analysis of large scale trajectory based dataset. Until now, the mobile phone data were mostly used for estimating the origin and the destination of users' trips in space and time. Besides improving the spatial inference of the origins and the destinations, our work allows to estimate accurately the trajectory as well as the path between the origin and the destination. This knowledge brings a lot of potential for new studies including route choice modelling, trajectory mining and so on. Until now, the latter studies were only conducted with GPS data which was the only source of data to provide fine-grained spatio-temporal trajectories. The second one is the analysis of the mobility from a multimodal perspective. While this subject was almost not covered by the literature, we demonstrate by combining TRANSIT with map-matching approaches that studying multimodality with mobile phone data was possible. We hope that our work will pave the way of new studies for dealing with remaining problems that were not tackle in thesis (transportation mode classification between road, public transport and train) and advances mobile phone data based approaches for studying human mobility from a multimodal perspective.

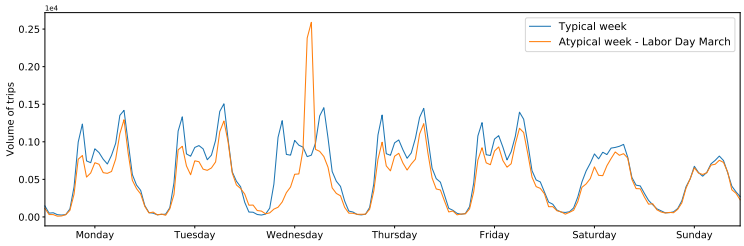
6.3 Opportunities

In the following we present different new cases studies based on NSD that have not been covered yet by the literature and/or enabled by the approaches developed in this thesis and presented in the previous chapters.

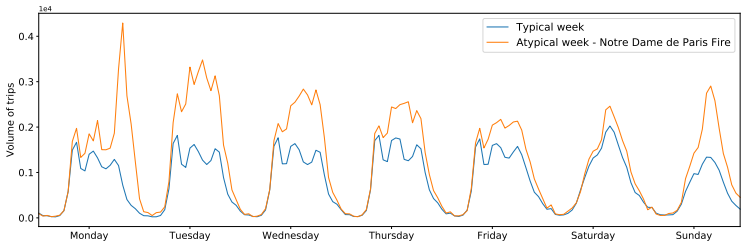
6.3.1 A1: Human Mobility Analysis during Abnormal Events

As an application, that our work makes possible is for instance detecting abnormal mobility situations that can occur in the city at fine spatiotemporal grained. For this,

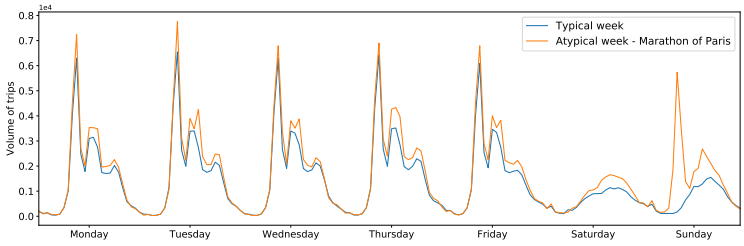
we segment the city of Paris into a set of squares of dimension 800m $\times$ 800m, with a temporal bin size of one hour. This spatio-temporal granularity makes it possible to analyze human mobility at a fine-grained scale. For each zone, we compute the *attraction demand profile*, which corresponds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



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



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



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

Figure 6.1: Typical/atypical weekly temporal demand profile during atypical events

First of all, on Wednesday, the 1st of May 2019, a bank holiday, the Labor Day march took place near Place d'Italie in Paris. Figure 6.1a shows in blue the typical attraction profile of this zone and in red the attraction profile of the week that includes the demonstration. Whereas, for all days, the attraction profile was similar to the typical profile, we can see that, on Labor Day, the attraction of the studied zone presents a high peak after midday.

As a second event, we studied the fire of the Notre Dame de Paris cathedral, on Monday the 15th of April 2019. Figure 6.1b shows in blue the typical attraction profile of this zone, and in red the attraction profile of the week that includes the abnormal event. We can see a high peak in the attraction profile right after the beginning of the fire on Monday 15th (around 6:30pm). Contrary to the previous example, this event also affected mobility the following days, when an attraction demand higher than

usual is observed in the corresponding area. This attraction demand progressively decreases after the event, but we notice an upsurge on Sunday, the Easter holiday, probably explainable by religious activities and nearby gatherings of tourists and worshipers visiting the area surrounding the cathedral after the fire on this special day.

Finally, we study another special event, the Marathon of Paris, on Sunday the 14^{th} of April 2019, with its start and end in the proximity of the Arc de Triomphe. A high peak on the attraction profile can be observed at the departure time of the marathon, at 9am, as shown in Figure 6.1c. A second peak, is observed few hours later, more spread over time and lower in magnitude compared to the first one, corresponding to the marathon arrival.

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

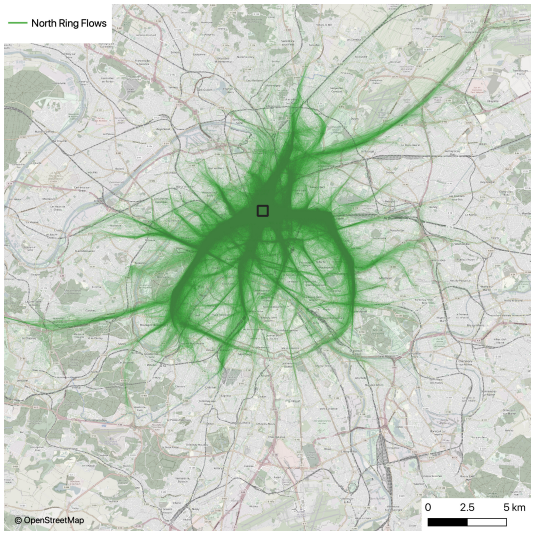
6.3.2 A2: Ring Road Trajectory Analysis

As a second case study, we leverage TRANSIT to perform a fine-grained trajectory analysis focused on the Paris *périphérique* (ring road). The mobility flow on this urban highway is usually very high, often leading to heavy congestion especially during peak hours. Transport authorities are traditionally very interested in the possibility of tracing and quantifying the flows of people moving along city major road axes. Such studies are necessary for urban planning purposes, infrastructure renewal and road maintenance, and can be extremely cost-demanding. Specifically, they are based on travel diaries or GPS trace collection, and generally end up capturing only a small sample of the flow actually traversing the major axis, with resulting limited accuracy. TRANSIT permits to leverage NSD to access a much larger and more representative sample of this specific population.

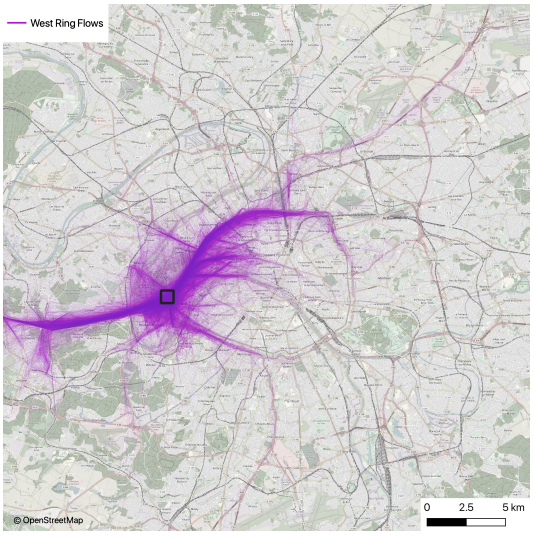
In our case study related to the Paris *périphérique*, we considered four different zones of interest: the east, west, north and south entries. The idea is to select a spatial zone and study all the trajectories passing by the respective zone. The enhanced trajectories $\widehat{\mathcal{M}}^i$ produced by TRANSIT on $\mathcal{D}_P$ allow us to capture at scale the origin, the destination, and the paths taken by the users passing by the studied zone, the kind of information usually expected in the aforementioned studies. The result for the four zones of the *périphérique* (east, west, north and south) are thus reported in Figure 6.2b, Figure 6.2c, Figure 6.2d and Figure 6.2a.

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

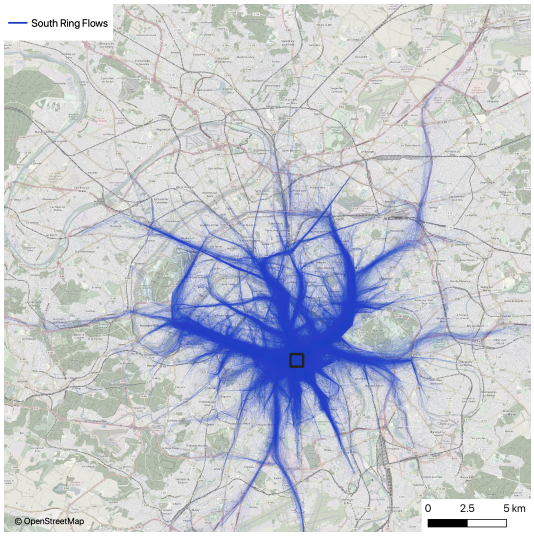
These results hint at the numerous perspectives brought by TRANSIT in the



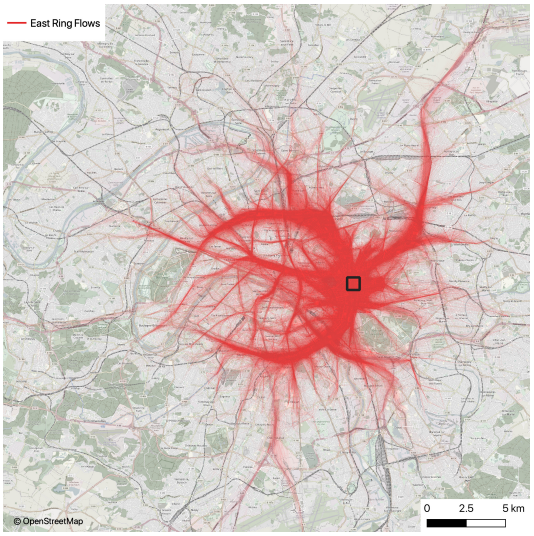
(a) North Ring Road



(b) West Ring Road



(c) South Ring Road



(d) East Ring Road

Figure 6.2: Heatmap of recurrent trips for the Paris ring-road (the black square shows the catchment area)

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

6.3.3 A3: Mobility based SIR for Studying COVID-19 Propagation

As a third application, we leverage NSD for providing mobility information and especially dynamic OD matrix that is integrated in a epidemiological model,*SIR*.

The model is based on the classical SIR model, developed in 1926 by Kermack and McKendrick [61] and which can be defined as follows:

$$\frac{ds(t)}{dt} = -\frac{\beta s(t) i(t)}{n(t)} \tag{6.1}$$

$$\frac{di(t)}{dt} = \frac{\beta s(t) i(t)}{n(t)} - \frac{\mu i(t)}{n(t)} \tag{6.2}$$

$$\frac{dr(t)}{dt} = \frac{\mu i(t)}{n(t)} \tag{6.3}$$

where, $s(t)$, $i(t)$, $r(t)$ are, respectively, the fraction of susceptible, infected and recovered population at time t. The equations 6.1 , 6.2 , 6.3 described the dynamics of the propagation in terms of the evolution in time of the number of people susceptible to be infected, infected and recovered. The rational behind this model is that the number of new people infected is proportional to the number of people susceptible to be infected $s(t)$ and infected $i(t)$. The proportional factor is β. This factor depends on the number of contacts between $s(t)$ and $i(t)$ and the probability that someone in $i(t)$ infects someone in $s(t)$. β factor is also called the incidence rate. In addition, the population recovered is proportional $r(t)$ to the population infected $i(t)$ with a proportional factor μ.

However, the classical SIR epidemic model proposed by Kermack and McKendrick [61] does not consider the flows that can occur in a region and which affect the epidemic dynamics. To overcome this limitation, we introduce the *mobility* and *social connectivity parameters* in our proposed model. Instead, considering the population as a whole, we will divide the studied regions into multiple locations and consider in the model the mobility between these locations. Beyond the epidemiological model, the innovative aspect of this work is to use the mobile phone data for capturing mobility dynamics of a region over time.

Let j ($j \subset l$) represent a set of locations (zones in a city for instance), which are connected to location i. Therefore, $\sum_j N_j$ is the maximum possible number of individuals connected to location i, from all the locations j. The parameter $c_{i,j}$ reflects the mobility of individuals from locations j to location i. This rate will be inferred using mobile phone data. Global transmission depends upon this mobility parameter of individuals from one location to another. Similar to local transmission, I_j is the number of individuals in the infected compartment in location j. Hence, total mobility of infected individuals from all the other connected locations to location i is $\sum_j c_{i,j} \frac{I_j}{N_j}$.

Considering the above description, the chances of transmission of infection from

all the connected locations to location i is $\sum_j c_{i,j} \frac{I_j}{N_j} \beta$. This transmission further depends upon the *social connectivity* (α) of all the individuals at location i. α captures the fact that the population infected within a region is proportional with the number of social interactions between the individuals of this region. Therefore, the proportion of healthy individuals at location i which can get infected from infected individuals from location j is $\frac{\alpha \sum_j c_{i,j} \frac{I_j}{N_j} \beta}{N_i + \sum_j c_{i,j}}$. Thus, the mean-field equations for the dynamics of the pandemic, based on the above discussed interactions are the following:

$$\frac{dS_i(t)}{dt} = -\frac{\beta S_i(t) I_i(t)}{N_i(t)} - \frac{\alpha S_i(t) \sum_j c_{i,j} \frac{I_j(t)}{N_j(t)} \beta}{N_i(t) + \sum_j c_{i,j}} \tag{6.4}$$

$$\frac{dI_i(t)}{dt} = \frac{\beta S_i(t) I_i(t)}{N_i(t)} + \frac{\alpha S_i(t) \sum_j c_{i,j} \frac{I_j(t)}{N_j(t)} \beta}{N_i(t) + \sum_j c_{i,j}} - \frac{\mu I_i(t)}{N_i(t)} \tag{6.5}$$

$$\frac{dR_i(t)}{dt} = \frac{\mu I_i(t)}{N_i(t)} \tag{6.6}$$

We perform various simulation experiments to explain the proposed model on *OD* matrix in a synthetic case and we then stress our model with a real world experiment. It is to be noted that, the model will behave as a standard SIR model in two cases, (i) if $\alpha = 0$, (ii) if the mobility is reduced to 100 percentile (that is no mobility allowed) from connected locations.

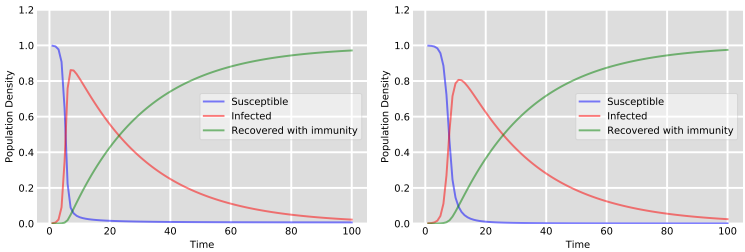
Pandemic Origins From Random Location

Fig. 6.3a to 6.3d display the influence of the *social connectivity parameter* 'α' while keeping the other parameters constant. It shows the pandemic dynamics with different values of α starting with $\alpha = 1$ to $\alpha = 0.1$. We observe that the peak of the infected compartment decreases significantly, as the α decreases, and it also takes longer to reach its peak. This indicates that there is a positive impact of lock-down in controlling a pandemic.

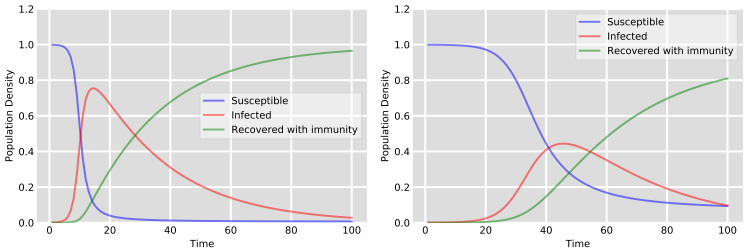
The effect of restricting the mobility from the top-X percentile of highly connected locations with other locations is shown in Fig. 6.3e to 6.3h. It displays the pandemic dynamics with different percentile of mobility restrictions of highly connected locations starting with 0% to 30% (keeping $\alpha = 0.5$). We observe that in the case of a pandemic, restricting the mobility from the top-10 percentile of highly connected locations can reduce the number of individuals who can get infected to 27%. Therefore, quarantine plays a vital role during pandemics.

Moreover, we can observe that the *social connectivity parameter* 'α' and *mobility* both play a fundamental role in determining the dynamics of the pandemics. Indeed, *Controlling mobility* reduces the fraction of infected individuals, and α delays the peak. Therefore, it is advisable to follow a dual strategy approach on the two above mentioned parameters during a pandemic outbreak.

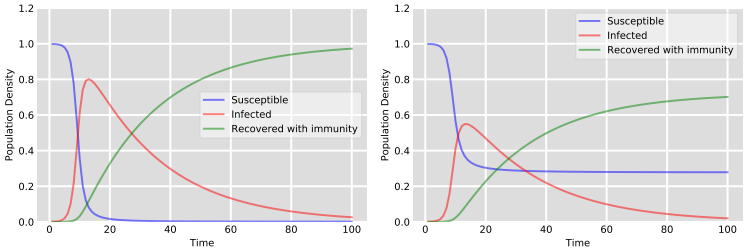
We also applied our model to real-time data of Rhône-Alpes region's COVID-19 cases. Figure 6.4 shows the actual number of cases and the cases forecast by the



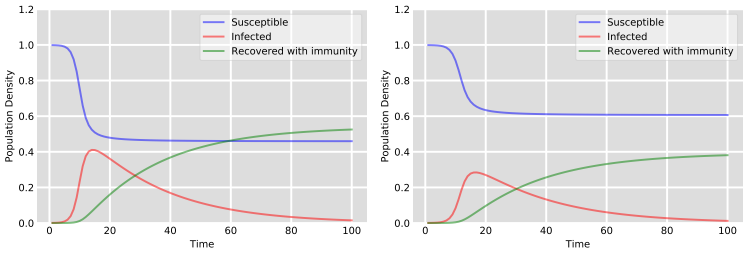
(a) $\alpha = 1$ (b) $\alpha = 0.6$



(c) $\alpha = 0.4$ (d) $\alpha = 0.1$



(e) 0% Locations Quarantine (f) 10% Locations Quarantine



(g) 20% Locations Quarantine (h) 30% Locations Quarantine

Figure 6.3: Pandemic Origin From Random Location: Effect of *Social Connectivity Parameter* 'α' (a), (b), (c), (d) and Quarantine Strongly Connected Locations (e), (f), (g), (h)

model using different values for α and mobility percentile. The region is divided into 14 sectors. For simulation purpose, we again considered the *OD* matrix between the sectors of the region obtained via network signaling data from Orange and census data[1]. This OD matrix has been built using the approach presented in Chapter 2.

For privacy matters, the number of COVID-19 cases is not reported in France at a fine spatio-temporal resolution in publicly available data[2]. Instead, the dataset

[1] https://www.insee.fr/fr/statistiques/4228434

[2] https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-infra-departementales-durant-lepidemie-covid-19/

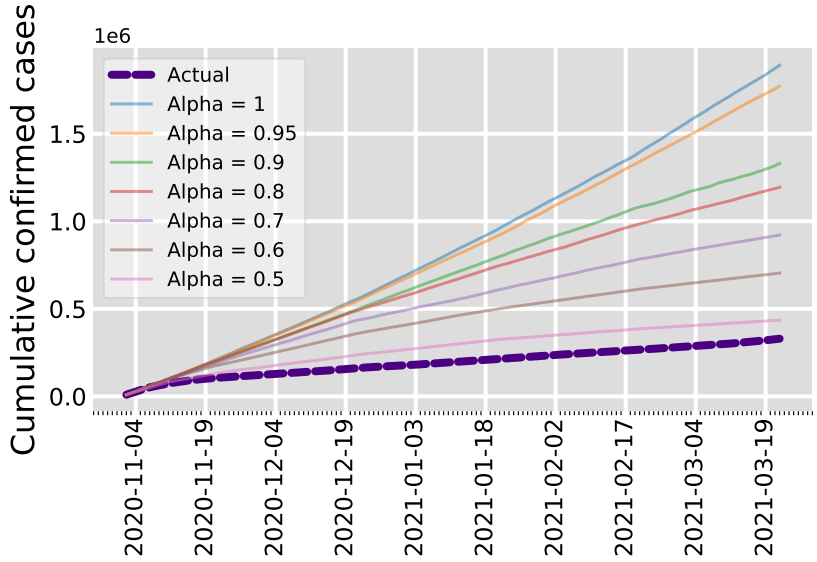


Figure 6.4: COVID-19 Cases In Rhône-Alpes Region In France.

only reports cumulative values of COVID-19 cases on a 7-days rolling window for each area of the administrative segmentation of the French territory. In addition, the number of cases is reported discretely, *i.e.*, as a range of values between a lower and upper bound containing the real value. Therefore, in order to obtain a daily estimation of the number of COVID-19 cases per each sector of the analyzed Rhone-Alpes region, two main assumptions have been made. On the one hand, we consider the number of cases as the mid value between the lower and upper bound of the reported range for the given area on a specific day. On the other hand, we replaced the 7 days rolling time window by the median day (*i.e.*, the 4^{th} day of the time window). As a result, to obtain the daily estimation of the number of cases, the cumulative reported estimation provided on a 7-days rolling time window is divided by 7. After summing this estimation for all the administrative areas belonging to a given sector of the OD matrix, we finally obtain an estimation of the number of COVID-19 cases per sector and per day. COVID-19 data cover the period from 2^{nd} November, 2020 to 19^{th} March, 2021.

For our simulation, the number of cases in all sectors is initialized as on 2^{nd} November, 2020. For the local transmission of the virus (within the sector), we consider the reproduction number $R_0 = 2.5$ [3]. The infection rate β and recovery rate μ are adjusted according to the value of R_0. The reported cases in Rhône-Alpes region in France as well as the forecast cases using the model, are shown in figure 6.4 until 22^{nd} March 2021. It can be noticed that the model predicted much higher cases of COVID-19 if no restrictions are introduced ($\alpha = 1$), while we can observe that, for $\alpha = 0.5$, the number of actual cases and forecast ones are quite close to each other. To explain this result, it is worth remembering that strong mobility restrictions were re-introduced in France by the end of October 2020[4], after the first lockdown ended during summer. The new restrictions contributed to keep low the number of COVID-19 infections (Actual). Moreover, it is reasonable to assume that mobility and social interactions were already significantly reduced at the beginning

[3] https://apps.who.int/iris/handle/10665/331443

[4] https://www.vie-publique.fr/en-bref/276947-covid-19-un-2e-confinement-national-compter-du-29-octobre-minuit

of this second lock-down, with respect to pre-pandemic behaviors, as a consequence of the first COVID-19 wave and previously imposed restrictive measures. In conclusion, this second case study confirms the applicability of the model to forecast a range of predicted number of cases. The latter can thus help the government and health agencies to understand the impact and introduce proportional interventions to restrict the expansion of the epidemic.

6.4 Challenges

Beyond the new opportunities offered by the mobile phone data, there are still several challenges that have to be tackled by the research community. Among these challenge one of the most important is to study the reliability of mobile phone data for human mobility study. Works from the literature [20, 50] and this thesis aim at characterizing the spatio-temporal bias of individual mobile phone trajectory by comparing these trajectories with GPS ground truth trajectories. The result show that the spatial bias can vary from 100 meters in urban area to several kilometers in rural areas. With TRANSIT, we can drastically reduce this spatial bias. At aggregated scale, the reliability of mobile phone data does not reach consensus in the research community. For population estimation, most of the works demonstrate the capability of mobile phone data for estimating accurately population distribution [31, 69]. However, the results are less convincing concerning travel demand estimation. On the one hand, Schneider *et al.* [98] show that the regular mobility patterns were correctly captured by the mobile network data compared to surveys. Several works [75, 19] also show OD matrix were equivalent to those obtained on surveys. On the other hand, Tizzoni *et al.* [114] observe statistical differences between commuting flows inferred from national census data and those observed from mobile phone data in three European countries, namely Portugal, Spain and France. In this thesis, this problem has been adressed, in Chapter 2 we demonstrate spatiotemporal bias when we used mobile phone data for estimating travel demand patterns. In Chapter 3, we improve, with TRANSIT, the travel demand inference and the approach showed good results when compared with surveys for inferring travel demand profiles. However, there is a need for finer validation before using TRANSIT in operational context.

Another challenge is related to preserving users' privacy in mobile phone datasets. Indeed, the mobile phone contains sensitive information related to the mobile phone user which raises privacy issue. This issue limits the accessibility and the diffusion of such a dataset. Thus, the research community investigates the problem of anonymization of mobile phone dataset. By analyzing fifteen months of human mobility data for one and a half million individuals, Montjoie *et al.* [79] find that human mobility traces are highly unique. Thus, the anonymization task is challenging. Gramaglia *et al.* develop GLOVES a trajectory anonymization approach which aims at hiding each trajectory with k similar trajectories so that the trajectory is not uniquely distinguishable from the k-others trajectories (k being a parameter of the approach). The approach relies on *generalization* and *suppression*. The *generalization* reduces data precision in space and time so that different mobile phone trajectories become identical. The *suppression* allows instead to remove some data, either individual samples or the whole trajectory if the anonymity criterion can not be fulfill.

The first challenge has already been discussed in Chapter 2 and Chapter 3. Related to the second challenge, we study an anonymization related problem. Instead of anonymizing trajectories as previously described, we study the problem of anonymizing an OD matrix inferred from NSD after applying TRANSIT and spatio-temporal aggregation. Indeed, if there are some OD-pairs that have a very low flow, there is a risk of re-identification of the user. This problem has not been tackled yet by the literature.

6.4.1 A6: Anonymization of Origin-Destination Matrix

Problem Formulation

We consider the problem of OD-matrix anonymization. Given an anonymity threshold k, all flows below k in the OD-matrix are suppressed. The goal is to find the best spatial grid for deriving the OD-matrix so that the number of flows lost is minimum and the spatial grid as fine as possible not to loose too much information on the OD-matrix. We consider a region partitioned in a uniform grid U containing $|U| = N \times N$ initial tiles. Let $8 \leq k \leq 16$ be our anonymity threshold. We perform spatial aggregation based on a quadtree, i.e., a tree-like data structure where each internal node n represents an area. There are two types of nodes: leaf nodes which are areas that can not be divided, these areas represent units of spatial information and non-leaf nodes which are areas that can be divided into four quadrants. $\mathcal{Q}$ is denoted as the set of quadtrees whose root represents the complete study area and whose vertices correspond to a non-empty set of tiles in U. Besides, for all $q \in \mathcal{Q}$ we note $\mathcal{L}(q)$ the set of the leaves of q and for all node $n \in q$ we note $|n|$ the number of tiles in U represented by n. For each $q \in \mathcal{Q}$, $\mathcal{L}(q)$ is a spatial partition of the study area. We apply spatial partition on origins and destinations separately. Origin aggregation aims at finding a spatial partition $\mathcal{L}(q^{ori}) = \{o_1, \ldots, o_i, \ldots\}$ for which the outgoing volumes v_o are the closest to a target volume v_{target}. Formally, q^{ori} represents the solution to the optimization problem, defined in Eq. 6.7:

$$q^{ori} = \arg\min_{q \in \mathcal{Q}} \sum_{o \in \mathcal{L}(q)} (v_o - v_{target})^2 \tag{6.7}$$

For each origin $o \in \mathcal{L}(q^{ori})$, destination aggregation aims at finding a spatial partition $\mathcal{L}(q_o^{dest}) = \{d_1, \ldots, d_i, \ldots\}$ that minimizes the generalization error defined in Eq. 6.8. This error, defined for a couple (o, d), corresponds to the *individual information loss* [72], which independently penalizes the generalization of each attribute. For OD trips, we only have two attributes namely their origin and destination and we may measure by $|o|$ and $|d|$ their spatial generalizations in number of tiles. This leads to the following loss function:

$$G(o,d) = \begin{cases} v_{od} \frac{|o|+|d|}{|U|} \text{ if } v_{od} \geq k \\ v_{od} \frac{|U|+|U|}{|U|} \text{ if } v_{od} < k. \end{cases} \tag{6.8}$$

Volumes $v_{od} < k$ must be suppressed, and are therefore counted as if they were aggregated at the highest level, leading to the maximal cost of 1 per attribute. The associated optimization problem is defined by:

$$q_o^{dest} = \arg\min_{q \in \mathcal{Q}} \sum_{d \in \mathcal{L}(q)} G(o,d). \tag{6.9}$$

Treating suppressed volumes as aggregated to the highest level leads to a disproportionate penalty and to solutions that have close to no suppression at all, resulting in coarse generalization. Trajectory data are known to consistently contain some hard-to-generalize outliers [39]. In order to allow the suppression of those outliers, we set a suppression threshold S interpreted as a maximal number of suppressed trips allowed, and we apply a coefficient $0 \leq \delta < 1$ to the cost of suppression. Generalization error for a trip becomes $G_\delta(o,d)$ defined as:

$$G_\delta(o,d) = \begin{cases} G(o,d) \text{ if } v_{od} \geq k \\ \delta G(o,d) \text{ if } v_{od} < k, \end{cases} \tag{6.10}$$

and the corresponding optimization problem becomes:

$$\begin{aligned} q_o^{dest} = \arg\min_{q \in \mathcal{Q}} & \sum_{d \in \mathcal{L}(q)} G_\delta(o,d), \\ s.t. & \sum_{\substack{d \in \mathcal{L}(q): \\ v_{od} \leq k}} v_{od} < S. \end{aligned} \tag{6.11}$$

Solution Proposed

All objective functions considered above are *modular*, meaning that for any partition $\mathcal{L}(q)$ of U, the objective function can be expressed as $\sum_{a \in \mathcal{L}(q)} g(a)$, where $g(a)_{a \in \mathcal{L}(q)}$ are *partial costs* that are independent of each other. In the absence of constraints as in Eq. 6.7 and Eq. 6.9, modularity makes it easy to recursively compute for any node $n \in q^U$ the smallest partial cost achievable $g^*(n)$: if n has no children, then $g^*(n) = g(n)$, else $g^*(n) = \min(g(n), \sum_{c \in children(n)} g(c))$. With this naive method, we solve Eq. 6.7 after visiting each node of q^U once, *i.e.*, in exactly $\frac{1}{3}(4|U| - 1)$ steps. Applied to the problem defined in Eq. 6.11 for each $o \in \mathcal{L}(q^{ori})$, this approach returns a solution $\tilde{q}_o^{dest}$ which is not guaranteed to respect the constraint. However, Eq. 6.11 can be recast as a type of knapsack problem. In this case, each node n of $\tilde{q}_o^{dest}$ is considered as an item with weight w being the volume that will get suppressed if n is split, and benefit b being the gain in generalization error induced by splitting n. The maximum capacity for the knapsack problem is then the suppression threshold S. Then, selecting which areas to split amounts to a knapsack problem with the following variants: *i)* the weight w may be zero if splitting n does not lead to suppressing additional volumes; *ii)* the benefit b may be negative if splitting causes too much suppression; *iii)* items follow a dependency tree: we can only split an area if its parent have been split. This problem has already been explored under the name of ordered knapsack problem in [57]. As our problem is small enough (an initial grid of $|U| = 128 \times 128$ gives $\frac{1}{3}(4|U| - 1) = 21\,845$ items for the knapsack problem), we can find the exact solution with a dynamic programming approach.

Case Study Lyon OD-matrix

For our case study, we work on the mobility tensor output by TRANSIT as described in Section 3.8. This mobility tensors can be seen as a set of OD-matrices, each OD-matrix being associated to one timeslot. The timeslot varies depending on the hour of the day so that OD matrices have comparable volume. Indeed, the night hours are merged in a single timeslot, during off peak each timeslot lasts two hours and in peak hour the timeslot lasts one hour. The mobility tensor covers one month data between 15^{th} March, 2019 and 16^{st} April, containing a total of 62 967 903 trips and is composed of 376 OD-matrices. The studied region can be modeled as a grid U was set with a mesh $m_U = 400\ m$ for a study area of $25600 \times 25600\ m^2$, which corresponds to $|U| = 64^2 = 4096$ tiles. We solved for $k \in \{8, 11, 16\}$ with $\delta = 0.01$ and $S = 10\%$ of total volume for each matrix. For each k, v_{target} is manually set based on performance obtained over one week of data. Due to execution time of the approach, we do not optimize the value v_{target}, instead we set v_{target} with reasonable value. We measure the precision of the aggregated data with the Mean Area of Origins (MAO) and Mean Area of Destinations (MAD) : $MAO = \frac{\sum_{od} v_{od} \times |o| m_U^2}{\sum_{od} v_{od}}$ and $MAD = \frac{\sum_{od} v_{od} \times |d| m_U^2}{\sum_{od} v_{od}}$. MAO and MAD represent the spatial precision of the aggregated data in m^2 and are tied to the total generalization error $G_{\delta,tot}$ with :

$$G_{\delta,tot} = \sum_{o,d} G_\delta(o,d) = (\frac{\sum_{od} v_{od}}{|U|}) * (\text{MAO} + \text{MAD}) + \delta * \sum_{\substack{o,d \\ v_{od} < k}} v_{od}. \tag{6.12}$$

k	approach	MAO (km^2)	MAD (km^2)	$G_{\delta,tot}$
8	our approach, $v_{target} = 500$	1.98	7.40	1.9e+05
8	naive 8*8 agg, reported	10.24	13.33	3.6e+05
8	naive 16*16 agg, reported	2.56	13.56.	2.6e+05
8	naive 32*32 agg, reported	0.64.	25.45.	4.1e+05
11	our approach, $v_{target} = 700$	2.78	9.15	2.2e+05
11	naive 8*8 agg, reported	10.24	14.57	3.8e+05
11	naive 16*16 agg, reported	2.56	17.59	3.1e+05
11	naive 32*32 agg, reported	0.64	33.80	5.4e+05
16	our approach, $v_{target} = 1000$	3.89	11.28	2.7e+05
16	naive 8*8 agg, reported	10.24	16.69	4.1e+05
16	naive 16*16 agg, reported	2.56	23.80	4.1e+05
16	naive 32*32 agg, reported	0.64	46.07	7.4e+05

Table 6.2: Performances of k-anonymization with our approach compared to naive tile aggregation, for various k

We compare our solution to a naive approach using a uniform spatial aggregation of 8×8, 16×16 and 32×32 initial tiles. Results are shown in Table 6.2. For all anonymity threshold considered $G_{\delta,tot}$ obtained with our approach is significantly better than the one obtained with competing approaches. We could further improve our approach by optimizing the parameter v_{target}.

6.5 Conclusion

In the literature, mobile phone data have been leveraged for a large bunch of applications such as elaborating the laws that govern human mobility at individual and aggregated scale, network-scale travel estimation or population density estimation. By overcoming the main limitations of the mobile phone trajectories *i.e.*, uncertainty in space, sparsity in time and oscillation effect, we unlock the potential of mobile phone data. To demonstrate this potential we develop, in this chapter several applications. The latter include fine grained human mobility analysis during abnormal events, ring road trajectories analysis and mobility based epidemiological model for modeling COVID-19 propagation. The approaches developed pave the way of new applications that make the analysis of human mobility at fine spatio-temporal granularity and in multimodal settings possible. These applications could be developed in future works and future directions of improvement of our own work have been discussed in the different chapter of this thesis. Besides a lot of opportunities, there are challenges that the mobile phone based approaches face. On the one hand, the research community and the work developed here demonstrate the importance of dealing with spatial and temporal bias on aggregated results inferred with mobile phone data. There is a need of strong validation of travel demand inferred with mobile phone data so that the latter can be considered as reliable. On the other hand, there is also a need for anonymization approach on mobile phone datasets. This would allow to protect the privacy of the user and facilitate the share within researcher and practitioner communities of mobile phone datasets.

Chapter 7

Conclusion

The research presented in this thesis has been centered around the study of mobile network data for estimating human mobility. These include methodologies able to overcome mobile phone limitations and produce enhanced (in space and time) mobile phone trajectories, map matching approaches able to map these trajectories to multimodal transportation network in urban environment. The latter allow new kind of application which are, for some of them, developed in this thesis as a proof of concept. Five specific research questions have been investigated in the previous chapters. This chapter aims at answering these questions.

In this final chapter, we draw conclusions in section 7.1 and limitations in section 7.2. We further discuss research directions that are worthwhile exploring in the future in section 7.3.

7.1 Answers to Research Questions

This thesis has been devoted to develop approaches for large scale urban mobility estimation. To conclude the thesis, we now provide answers to all the research questions raised in Chapter 1.

RQ1: To what extent are mobile phone data suitable for estimating human mobility? - Chapter 2

By applying Fekih *et al.* [33] framework on signalling data of 2 million mobile phone users, we show that NSD is feasible to robustly extract residents' trips and estimate the hourly trip distribution throughout the Lyon region, on the condition that spatio-temporal biases of cell phone signalling transactions are properly detected and removed. With our debiasing procedure, travel demand inferred by mobile phone data exhibits strong correlations with the one obtained with surveys. Moreover, by clustering the trip flows based on the temporal profile of the emitted demand of each zone and matching them with official land use data, we also unveil interesting and relevant heterogeneities in dynamic travel demand patterns related to trip production zones. All the above-mentioned results advocate for the suitability of mobile phone data for estimating human mobility.

RQ2: Can the repetitive nature of human mobility be used to improve in space and time human trajectories as observed through the bias of mobile phone data? - Chapter 3

The repetitive nature of human mobility has been largely demonstrated in the literature. Based on this property and by leveraging oscillation effect as triangulation, we design TRANSIT, a framework able to improve temporal granularity of mobile phone trajectories as well as reducing the spatial uncertainty. TRANSIT outperforms the other approaches of the literature on a ground truth dataset with both mobile phone and gps data for a set of volunteer in the Lyon metropolitan area. TRANSIT achieves an average spatial accuracy of 190m. Besides these satisfying results, TRANSIT is scalable, thus it can be used for large scale human mobility estimation.

RQ3: How can we estimate very fine mobility information *i.e.*, the path traveled on a multimodal transportation network, from mobile phone trajectories ? - Chapter 4

For answering this question, we studied, first, the challenging task of map-matching mobile phone trajectories to the transportation network. For mapping mobile phone trajectories to multimodal transportation network, we develop an HMM based map-matching which achieves low spatial accuracy. The spatial accuracy, the matching rate and the F1 score can be significantly improved by relying on TRANSIT before applying map-matching. However, we also showed that despite low spatial accuracy, the map-matching applied directly on multimodal-transportation network exhibit low matching rate (*i.e.*, the transportation mode is wrongly inferred by our approach). The results were much better when the map-matching was applied on each layer separately: road, public transport and train layers. We leverage the approach to infer popular paths per transportation mode in the city of Lyon.

RQ4: How can we derive aggregated mobility patterns along the main dimensions that characterize human mobility? - Chapter 5

For estimating aggregated mobility patterns, we consider analyzing the mobility

in three dimensions: origin, destination and time. We build a daily mobility tensor aiming at capturing the flows along these dimensions in the city of Lyon. The mobility tensor has been inferred by relying on the aggregation in space and time of the result of TRANSIT (Chapter 3). The low spatial error that exhibits TRANSIT allows us to use a fine spatial segmentation for the origin and destinations dimensions: the Lyon area has been divided in 625 areas, each area being a 800m$\times$800m square. Based on the approach from the literature, we apply non-negative sparse tucker decomposition to the daily mobility tensor. Then, we analyzed the resulting decomposition and showed that the approach were able to infer fine-grained temporal and spatial (origin and destination) patterns. The approach is also able to capture complex spatio-temporal dependencies. We also demonstrated that our approach is resistant to sampling effect. Indeed, even with 50% sampling ratio applied to the daily mobility tensor, most of spatio-temporal patterns were preserved.

RQ5: What kind of applications are made possible by our approach when applied on mobile phone data? - Chapter 6

The approaches developed in this thesis unlock the potential of mobile phone data. On the one hand, we provide TRANSIT able to perform trajectory segmentation as well as trajectory enhancement better than state of the art approaches. The latter allows to improve already existing studies/application related to large scale travel demand inference in urban environment with mobile phone data. Moreover, it allows new kinds of mobility studies including fine-grained trajectory based analysis which were previously reserved for GPS data. In addition, our multimodal map-matching approach paves the way for multimodality analysis at large scale and in urban environment using mobile phone data. Finally, we leverage the approaches developed in this thesis to feed an original case study which consists on an epidemiological model for studying COVID-19 propagation. The latter application demonstrates the potential of mobile phone data to be used for a large diversity of applications.

7.2 Limitations

The present work is constrained by several limitations. These limitations are related to the mobile network data. This work is based on the use of NSD that are not widely open. For privacy matters discussed above there is only few works on this kind of data and research on NSD is still at early stages. Moreover, it is very difficult to obtain large scale ground truth datasets. The drawbacks are twofold. On the one hand, larger dataset would add confidence on the obtained results and would also allow finer analysis for instance concerning the multimodal map-matching. On the other hand, the lack of large scale ground truth dataset prevent the use of supervised algorithms. Instead we have to rely on unsupervised approaches that exhibits in general lower performances. Still, despite the aforementioned limitations, the presented works bring several contributions and can be further developed for future work perspectives.

7.3 Future Directions

There are several improvements that can be done beyond the approaches developed in this thesis. These include the refinement of the parameters of our HMM based map-matching approach. Indeed, the transition matrix could be estimated using the real traffic conditions whereas the emission matrix could be estimated with antennas' coverage maps provided by Orange or by Bayesian inference using ground truth data. In addition, TRANSIT could be further improved taking into account the inter-individual regularity of human mobility *i.e.*, different users that take similar paths from one origin to one destination. This could be done by integrating to our framework additional clustering approach. On the tensor decomposition approach it could be interesting to investigate the decomposition obtained by adding the transportation mode dimension. This would allow to obtain mobility pattern per mode and these patterns could represent valuable knowledge for transportation planners. Finally, a lot of new applications could be developed on the basis of the work proposed in this thesis as discussed in Chapter 6.

Appendix A

Alternating proximal gradient approach

Let (g_1, g_2, g_3, g_4) denote $(\boldsymbol{\mathcal{C}}, \mathbf{O}, \mathbf{D}, \mathbf{T})$ for concision. Using a Proximal Gradient (PG) method, the algorithm updates the i-th variable of $\mathcal{G}$ in the s-th round as:

$$\begin{aligned}\mathbf{g}_i^{(s)} &= \underset{\mathbf{g}_i \geq 0}{\operatorname{argmin}} \left\langle \frac{\partial \mathcal{J}\left(\mathbf{g}_{<i}^{(s)}, \tilde{\mathbf{g}}_i^{(s)}, \mathbf{g}_{>i}^{(s-1)}\right)}{\partial \mathbf{g}_i}, \mathbf{g}_i - \tilde{\mathbf{g}}_i^{(s)} \right\rangle + \frac{\tau_i}{2} \left\| \mathbf{g}_i - \tilde{\mathbf{g}}_i^{(s)} \right\|_F^2 + \lambda_i \left\| \mathbf{g}_i \right\|_1 \\ &= \max \left\{ 0, \tilde{\mathbf{g}}_i^{(s)} - \frac{1}{\tau_i} \frac{\partial \mathcal{J}\left(\mathbf{g}_{<i}^{(s)}, \tilde{\mathbf{g}}_i^{(s)}, \mathbf{g}_{>i}^{(s-1)}\right)}{\partial \mathbf{g}_i}, \mathbf{g}_i - \tilde{\mathbf{g}}_i^{(s)} - \frac{\lambda_i}{\tau_i} \right\}\end{aligned} \tag{A.1}$$

where $\langle \cdot \rangle$ denotes the inner product, $\mathbf{g}_{<i}^{(s)}$ denotes $\{\mathbf{g}_1^{(s)}, \cdots, \mathbf{g}_{i-1}^{(s)}\}$ and $\mathbf{g}_{>i}^{(s)}$ denotes $\{\mathbf{g}_{i+1}^{(s)}, \cdots, \mathbf{g}_4^{(s)}\}$. The variable $\tilde{\mathbf{g}}_i^{(s)}$is a linear extrapolated point as follows:

$$\tilde{\mathbf{g}}_i^{(s)} = \mathbf{g}_i^{(s-1)} + w_i^{(s)} \left(\mathbf{g}_i^{(s-1)} - \mathbf{g}_i^{(s-2)} \right) \tag{A.2}$$

where $w_i^{(s)}$ is an extrapolation weight set according to [ref]. The parameter τ_i is a Lipschitz constant of $\frac{\partial \mathcal{J}(\mathbf{g}_i)}{\partial \mathbf{g}_i}$ with respect to $\mathbf{g}_i$, namely,

$$\left\| \frac{\partial \mathcal{J}(\mathbf{g}_{i_1})}{\partial \mathbf{g}_{i_1}} - \frac{\partial \mathcal{J}(\mathbf{g}_{i_2})}{\partial \mathbf{g}_{i_2}} \right\|_F \leq \tau_i \left\| \mathbf{g}_{i_1} - \mathbf{g}_{i_2} \right\|_F, \forall \mathbf{g}_{i_1}, \mathbf{g}_{i_2} \tag{A.3}$$

and λ_i is the regularization parameter of $\mathbf{g}_i$. Specifically, the gradients of $\mathcal{J}$ with respect to each component are calculated as:

$$
\begin{aligned}
\nabla_{\mathcal{C}}\mathcal{J} =& 2\mathcal{C} \times_o \left(\mathbf{O}^\top\mathbf{O}\right) \times_d \left(\mathbf{D}^\top\mathbf{D}\right) \times_t \left(\mathbf{T}^\top\mathbf{T}\right) \\
& - 2\mathcal{M} \times_o \mathbf{O}^\top \times_d \mathbf{D}^\top \times_t \mathbf{T}^\top \\
\nabla_{\mathbf{O}}\mathcal{J} =& 2\mathbf{O}\left(\mathcal{C} \times_d \left(\mathbf{D}^\top\mathbf{D}\right) \times_t \left(\mathbf{T}^\top\mathbf{T}\right)\right)_{(o)} \mathcal{C}_{(o)}^\top \\
& - 2\left(\mathcal{M} \times_d \mathbf{D}^\top \times_t \mathbf{T}^\top\right)_{(o)} \mathcal{C}_{(o)}^\top - \frac{1}{1\sigma_{WO}^2}\left(\mathbf{W} - \mathbf{O}\mathbf{O}^\top\right)\mathbf{O} \\
\nabla_{\mathbf{D}}\mathcal{J} =& 2\mathbf{D}\left(\mathcal{C} \times_o \left(\mathbf{O}^\top\mathbf{O}\right) \times_t \left(\mathbf{T}^\top\mathbf{T}\right)\right)_{(d)} \mathcal{C}_{(d)}^\top \\
& - 2\left(\mathcal{M} \times_o \mathbf{O}^\top \times_t \mathbf{T}^\top\right)_{(d)} \mathcal{C}_{(d)}^\top - \frac{1}{1\sigma_{WD}^2}\left(\mathbf{W} - \mathbf{D}\mathbf{D}^\top\right)\mathbf{D} \\
\nabla_{\mathbf{T}}\mathcal{J} =& 2\mathbf{T}\left(\mathcal{C} \times_o \left(\mathbf{O}^\top\mathbf{O}\right) \times_d \left(\mathbf{D}^\top\mathbf{D}\right)\right)_{(t)} \mathcal{C}_{(t)}^\top \\
& - 2\left(\mathcal{M} \times_o \mathbf{O}^\top \times_d \mathbf{D}^\top\right)_{(t)} \mathcal{C}_{(t)}^\top
\end{aligned}
\tag{A.4}
$$

The algorithm is given by the following pseudo-code:

Procedure 2 Alternating proximal gradient for sparse NTD

Data: daily mobility tensor $\mathcal{M}$
Initialization: $\mathbf{A} = (\mathbf{O}, \mathbf{D}, \mathbf{T})$, $\mathcal{C}^{-1} = \mathcal{C}$ and $\mathbf{A}^{-1} = \mathbf{A}$

for $k = 1, 2, \cdots$ **do**
  Set $\mathcal{C}^{k,-1} = \mathcal{C}^{k,0} = \mathcal{C}^0$ if $k = 1$, $\mathcal{C}^{k,-1} = \mathcal{C}^{k-1,N-1}$, $\mathcal{C}^{k,0} = \mathcal{C}^{k-1,N}$ otherwise.
  for $n = 1, \cdots, N$ **do**
    Update core tensor $\mathcal{C}$
    Choose $L_{\mathcal{C}}^{k,n}$ to be a Lipschitz constant of $\nabla_{\mathcal{C}} \mathcal{J}(\mathcal{C}, \mathbf{A}_{j<n}^{k}, \mathbf{A}_{j\geq n}^{k-1})$ about $\mathcal{C}$
    Choose $w_{\mathcal{C}}^{k,n} \geq 0$ and set $\tilde{\mathcal{C}}^{k,n} = \mathcal{C}^{k,n-1} + w_{\mathcal{C}}^{k,n}(\mathcal{C}^{k,n-1} - \mathcal{C}^{k,n-2})$

    Update $\mathcal{C}$ by $\mathcal{C} = \max\left\{0, \tilde{\mathcal{C}}^{k,n} - \frac{1}{L_{\mathcal{C}}^{k,n}} \nabla_{\mathcal{C}} \mathcal{J}(\tilde{\mathcal{C}}^{k,n}, \mathbf{A}_{j<n}^{k}, \mathbf{A}_{j\geq n}^{k-1}) - \frac{\lambda_{\mathcal{C}}}{L_{\mathcal{C}}^{k,n}}\right\}$

    Update factor matrice $\mathbf{A}_n$

    Choose L_n^k to be a Lipschitz constant of $\nabla_{\mathbf{A}_n} \mathcal{J}(\mathcal{C}^{k,n}, \mathbf{A}_{j<n}^{k}, \mathbf{A}_n, \mathbf{A}_{j>n}^{k-1})$ about $\mathbf{A}_n$

    Choose $w_n^k \geq 0$ and set $\tilde{\mathbf{A}}_n^k = \mathbf{A}_{n-1}^k + w_n^k(\mathbf{A}_{n-1}^k - \mathbf{A}_{n-2}^k)$

    Update $\mathbf{A}_n$ by $\mathbf{A}_n = \max\left\{0, \tilde{\mathbf{A}}_n^k - \frac{1}{L_n^k} \nabla_{\mathbf{A}_n} \mathcal{J}(\mathcal{C}^{k,n}, \mathbf{A}_{j<n}^{k}, \tilde{\mathbf{A}}_n^k, \mathbf{A}_{j>n}^{k-1}) - \frac{\lambda_n}{L_n^k}\right\}$

    Re-update if loss function has increased
$\mathcal{J}(\mathcal{C}^{k,n}, \mathbf{A}_{j\leq n}^{k}, \mathbf{A}_{j>n}^{k-1}) > \mathcal{J}(\mathcal{C}^{k,n-1}, \mathbf{A}_{j<n}^{k}, \mathbf{A}_{j\geq n}^{k-1})$

    Re-update $\mathcal{C}^{k,n}$ and $\mathbf{A}_n^k$ with $\tilde{\mathcal{C}}^{k,n} = \mathcal{C}^{k,n-1}$ and $\tilde{\mathbf{A}}_n^k = \mathbf{A}_n^{k-1}$
  end for
  Set $\mathcal{C}^k = \mathcal{C}^{k,N}$
stopping conditions
  Return($\mathcal{C}^k, \mathbf{A}_1^k, \cdots, \mathbf{A}_N^k$)
end for

Appendix B

Factor matrices obtained with R-NTF under 50% sampling ratio

B.1 Temporal Patterns

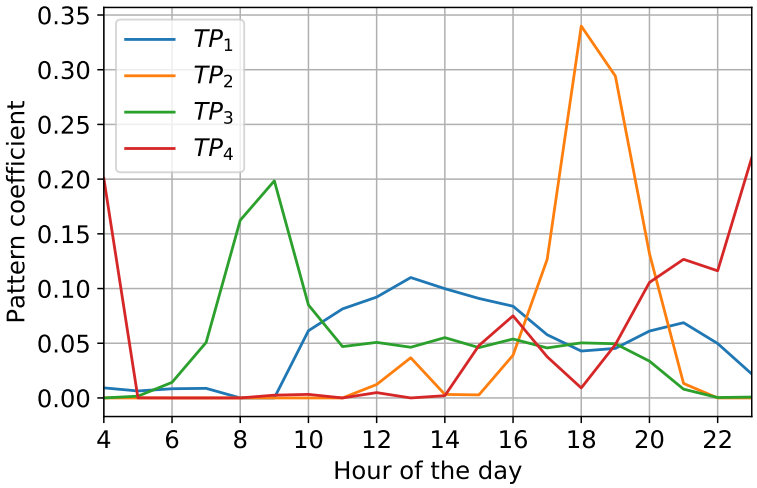
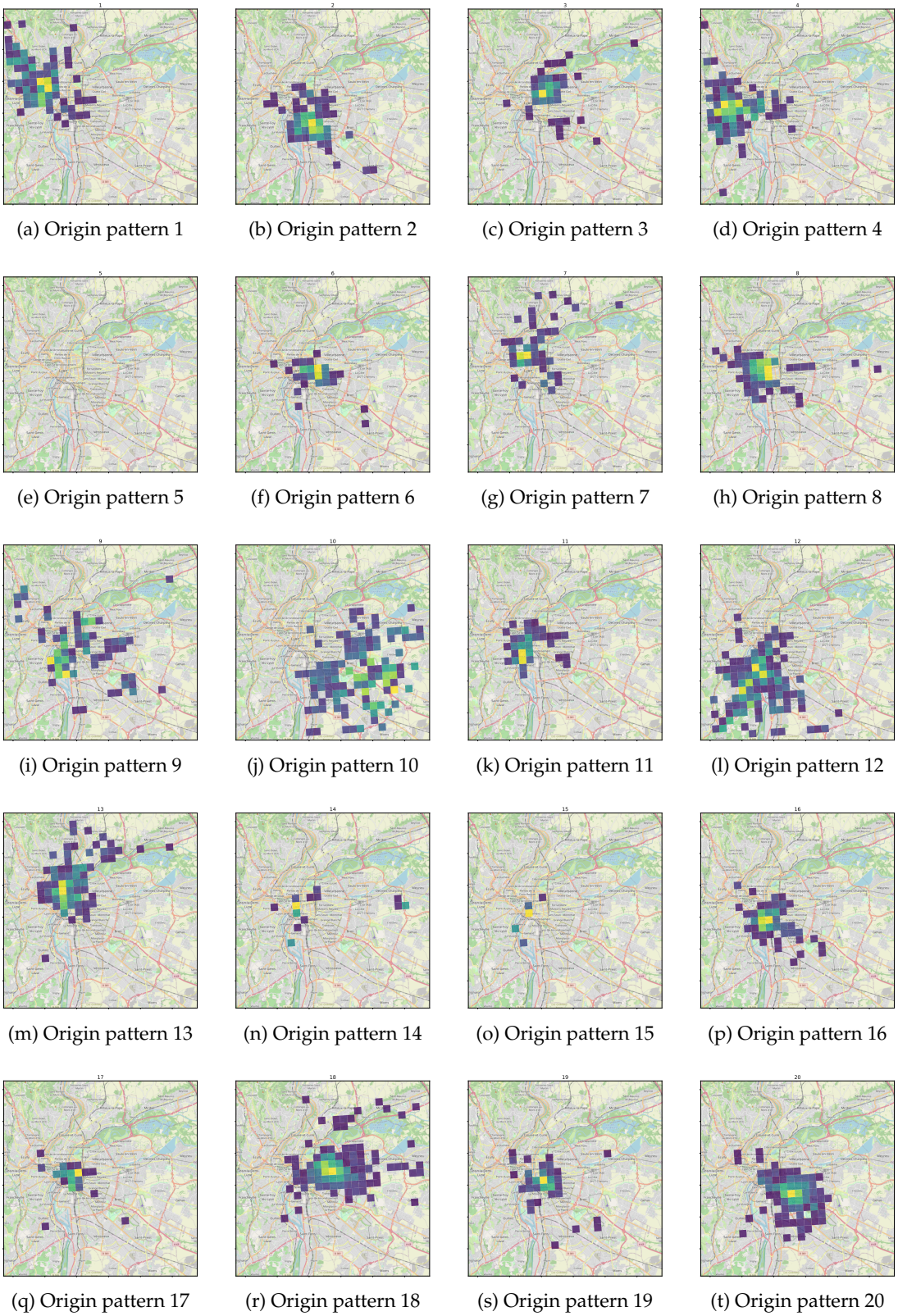


Figure B.1: Hidden temporal patterns

B.2 Origin Patterns



(a) Origin pattern 1 (b) Origin pattern 2 (c) Origin pattern 3 (d) Origin pattern 4

(e) Origin pattern 5 (f) Origin pattern 6 (g) Origin pattern 7 (h) Origin pattern 8

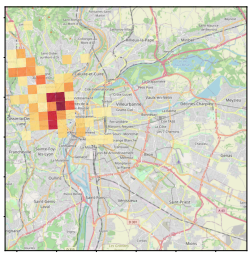
(i) Origin pattern 9 (j) Origin pattern 10 (k) Origin pattern 11 (l) Origin pattern 12

(m) Origin pattern 13 (n) Origin pattern 14 (o) Origin pattern 15 (p) Origin pattern 16

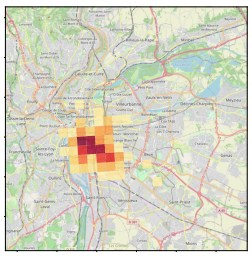
(q) Origin pattern 17 (r) Origin pattern 18 (s) Origin pattern 19 (t) Origin pattern 20

Figure B.2: Hidden origin patterns

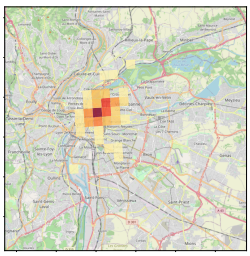
B.3 Destination Patterns



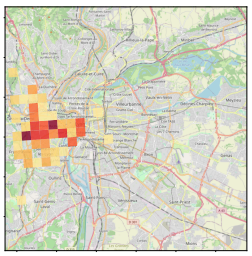
(a) Destination pattern 1



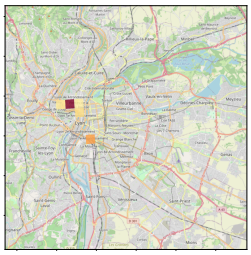
(b) Destination pattern 2



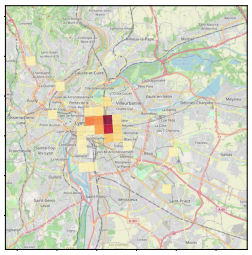
(c) Destination pattern 3



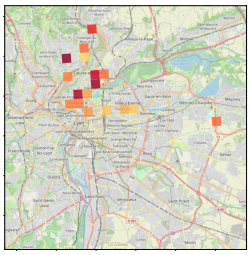
(d) Destination pattern 4



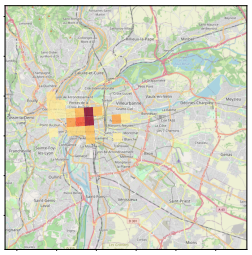
(e) Destination pattern 5



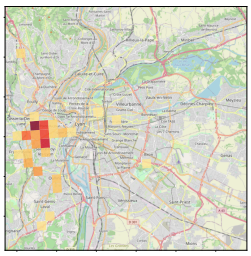
(f) Destination pattern 6



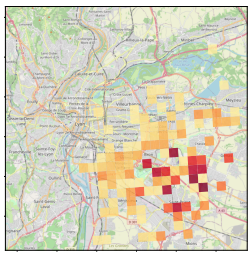
(g) Destination pattern 7



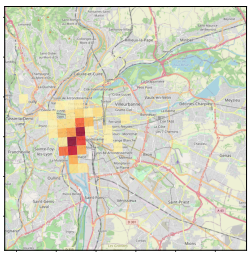
(h) Destination pattern 8



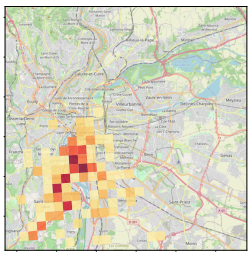
(i) Destination pattern 9



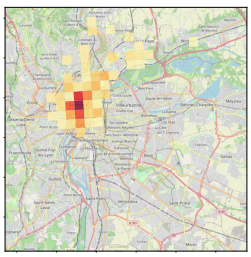
(j) Destination pattern 10



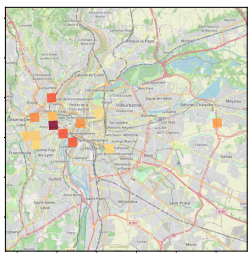
(k) Destination pattern 11



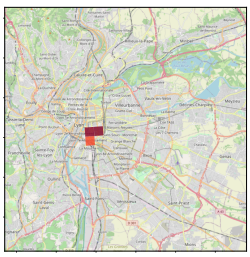
(l) Destination pattern 12



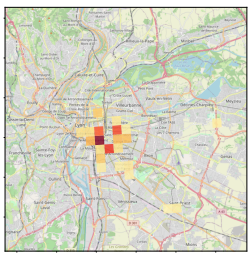
(m) Destination pattern 13



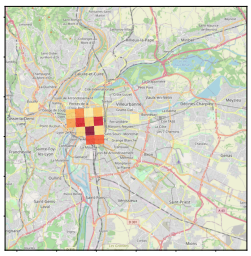
(n) Destination pattern 14



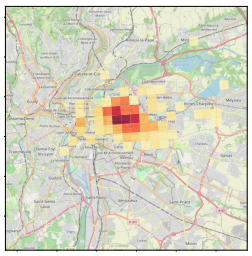
(o) Destination pattern 15



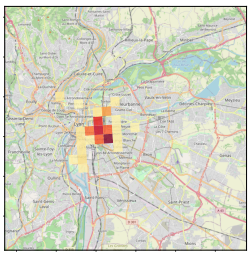
(p) Destination pattern 16



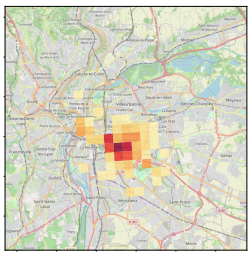
(q) Destination pattern 17



(r) Destination pattern 18



(s) Destination pattern 19



(t) Destination pattern 20

Figure B.3: Hidden destination patterns

Bibliography

[1] Rein Ahas et al. "Mobile positioning in space-time behaviour studies: Social positioning method experiments in Estonia". In: *Cartography and Geographic Information Science* 34.4 (Oct. 2007), pp. 259–273. ISSN: 15230406. DOI: 10.1559/152304007782382918. URL: https://www.tandfonline.com/doi/abs/10.1559/152304007782382918.

[2] Rein Ahas et al. "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones". In: *Journal of Urban Technology* 17.1 (Apr. 2010), pp. 3–27. ISSN: 1063-0732. DOI: 10.1080/10630731003597306. URL: http://www.tandfonline.com/doi/abs/10.1080/10630731003597306.

[3] Lauren Alexander et al. "Originâ"destination trips by purpose and time of day inferred from mobile phone data". In: *Transportation Research Part C* 58 (2015), pp. 240–250. DOI: 10.1016/j.trc.2015.02.018. URL: http://dx.doi.org/10.1016/j.trc.2015.02.018.

[4] Essam Algizawy, Tetsuji Ogawa, and Ahmed El-Mahdy. "Real-Time Large-Scale Map Matching Using Mobile Phone Data". In: *ACM Transactions on Knowledge Discovery from Data* 11.4 (July 2017), pp. 1–38. ISSN: 15564681. DOI: 10.1145/3046945. URL: http://dl.acm.org/citation.cfm?doid=3119906.3046945.

[5] Theo Arentze et al. "Data Needs, Data Collection, and Data Quality Requirements of Activity-Based Transport Demand Models". In: *Transportation Research Circular* E-C008 (Aug. 2000). ISSN: 0097-8515.

[6] Fereshteh Asgari et al. "CT-Mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network". In: *Computer Communications* (2016). ISSN: 01403664. DOI: 10.1016/j.comcom.2016.04.014.

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

[8] Danya Bachir et al. "Inferring dynamic origin-destination flows by transport mode using mobile phone data". In: *Transportation Research Part C: Emerging Technologies* 101 (Apr. 2019), pp. 254–275. ISSN: 0968-090X. DOI: 10.1016/J.TRC.2019.02.013. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18310519.

[9] Danya Bachir et al. "Using mobile phone data analysis for the estimation of daily urban dynamics". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2017, pp. 626–632. ISBN: 978-1-5386-1526-3. DOI: 10.1109/ITSC.2017.8317956. URL: http://ieeexplore.ieee.org/document/8317956/.

[10] Hillel Bar-Gera. "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel". In: (). DOI: 10.1016/j.trc.2007.06.003. URL: www.elsevier.com/locate/trc.

[11] Hugo Barbosa et al. "Human mobility: Models and applications". In: *Physics Reports* 734 (Mar. 2018), pp. 1–74. ISSN: 0370-1573. DOI: 10.1016/J.PHYSREP.2018.01.001. URL: https://www.sciencedirect.com/science/article/pii/S037015731830022X.

[12] Michele Berlingerio et al. "AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8190 LNAI.PART 3 (2013), pp. 663–666. DOI: 10.1007/978-3-642-40994-3{_}50. URL: https://link.springer.com/chapter/10.1007/978-3-642-40994-3_50.

[13] Geoff Boeing. "OSMNX: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks". In: *SSRN Electronic Journal* (May 2016). ISSN: 1556-5068. DOI: 10.2139/ssrn.2865501. URL: http://www.ssrn.com/abstract=2865501.

[14] Patrick Bonnel et al. "Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations". In: *Transportation Research Procedia*. Vol. 11. 2015. ISBN: 0000000000. DOI: 10.1016/j.trpro.2015.12.032.

[15] Loïc Bonnetain et al. "Can We Map-Match Individual Cellular Network Signaling Trajectories in Urban Environments? Data-Driven Study". In: *Transportation Research Record: Journal of the Transportation Research Board* 2673.7 (July 2019), pp. 74–88. ISSN: 0361-1981. DOI: 10.1177/0361198119847472. URL: http://journals.sagepub.com/doi/10.1177/0361198119847472.

[16] Loïc Bonnetain et al. "TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data". In: *Transportation Research Part C: Emerging Technologies* 130 (Sept. 2021), p. 103257. ISSN: 0968-090X. DOI: 10.1016/J.TRC.2021.103257. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X21002692.

[17] Noelia Caceres et al. "Traffic flow estimation models using cellular phone data". In: *IEEE Transactions on Intelligent Transportation Systems* 13.3 (Sept. 2012), pp. 1430–1441. DOI: 10.1109/TITS.2012.2189006.

[18] Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. "Urban Sensing Using Mobile Phone Network Data: A Survey of Research". In: *ACM Computing Surveys* 47.2 (Aug. 2014), pp. 1–20. ISSN: 15577341. DOI: 10.1145/2655691. URL: https://dl.acm.org/doi/10.1145/2655691.

[19] Francesco Calabrese et al. "Estimating origin-destination flows using mobile phone location data". In: *IEEE Pervasive Computing* 10.4 (Oct. 2011), pp. 36–44. DOI: 10.1109/MPRV.2011.41.

[20] Francesco Calabrese et al. "Real-time urban monitoring using cell phones: A case study in Rome". In: *IEEE Transactions on Intelligent Transportation Systems* 12.1 (Mar. 2011), pp. 141–151. DOI: 10.1109/TITS.2010.2074196.

[21] Francesco Calabrese et al. "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example". In: *Transportation Research Part C: Emerging Technologies* 26 (2013), pp. 301–313. ISSN: 0968090X. DOI: 10.1016/j.trc.2012.09.009.

[22] Cynthia Chen et al. *The promises of big data and small data for travel behavior (aka human mobility) analysis*. 2016. DOI: 10.1016/j.trc.2016.04.005.

[23] Guangshuo Chen et al. "Complete trajectory reconstruction from sparse mobile phone data". In: *EPJ Data Science* 8.1 (Dec. 2019), p. 30. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-019-0206-8. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-019-0206-8.

[24] Guangshuo Chen et al. "Enriching sparse mobility information in Call Detail Records". In: *Computer Communications* 122 (June 2018), pp. 44–58. ISSN: 0140-3664. DOI: 10.1016/J.COMCOM.2018.03.012. URL: https://www.sciencedirect.com/science/article/pii/S0140366417309234.

[25] Kimberley Chin et al. "Inferring fine-grained transport modes from mobile phone cellular signaling data". In: *Computers, Environment and Urban Systems* 77 (Sept. 2019), p. 101348. ISSN: 0198-9715. DOI: 10.1016/J.COMPENVURBSYS.2019.101348.

[26] Serdar Çolak et al. "Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities". In: *Transportation Research Record* 2526 (Apr. 2015), pp. 126–135. ISSN: 03611981. DOI: 10.3141/2526-14. URL: https://journals.sagepub.com/doi/10.3141/2526-14.

[27] Balázs Cs. Csáji et al. "Exploring the mobility of mobile phone users". In: *Physica A: Statistical Mechanics and its Applications* 392.6 (Mar. 2013), pp. 1459–1473. ISSN: 0378-4371. DOI: 10.1016/J.PHYSA.2012.11.040. URL: https://www.sciencedirect.com/science/article/pii/S0378437112010059.

[28] Somayeh Danafar, Michal Piorkowski, and Krzysztof Krysczcuk. "Bayesian framework for mobility pattern discovery using mobile network events". In: *25th European Signal Processing Conference, EUSIPCO 2017* 2017-January (Oct. 2017), pp. 1070–1074. DOI: 10.23919/EUSIPCO.2017.8081372.

[29] David L. Davies and Donald W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. ISSN: 01628828. DOI: 10.1109/TPAMI.1979.4766909.

[30] Pierre Deville et al. "Dynamic population mapping using mobile phone data". In: *Proceedings of the National Academy of Sciences* 111.45 (Nov. 2014), pp. 15888–15893. ISSN: 0027-8424. DOI: 10.1073/PNAS.1408439111. URL: https://www.pnas.org/content/111/45/15888https://www.pnas.org/content/111/45/15888.abstract.

[31] Rex W Douglass et al. "High resolution population estimates from telecommunications data". In: *EPJ Data Science* 4.1 (Dec. 2015), p. 4. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-015-0040-6. URL: http://www.epjdatascience.com/content/4/1/4.

[32] Oscar Egu and Patrick Bonnel. "Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon". In: *Travel Behaviour and Society* 19 (Apr. 2020), pp. 112–123. ISSN: 2214-367X. DOI: 10.1016/J.TBS.2019.12.003.

[33] Mariem Fekih et al. "A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France)". In: *Transportation* (Apr. 2020), pp. 1–32. ISSN: 15729435. DOI: 10.1007/s11116-020-10108-w. URL: https://doi.org/10.1007/s11116-020-10108-w.

[34] Mariem Fekih et al. "Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: a large-scale comparative study with travel survey and land use data". In: *https://doi.org/10.1080/19427867.2021.1945854* (June 2021), pp. 1–19. DOI: 10.1080/19427867.2021.1945854. URL: https://www.tandfonline.com/doi/abs/10.1080/19427867.2021.1945854.

[35] 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*. IEEE, May 2017, pp. 1–9. ISBN: 978-1-5090-5336-0. DOI: 10.1109/INFOCOM.2017.8057089. URL: http://ieeexplore.ieee.org/document/8057089/.

[36] Angelo Furno et al. "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas". In: *IEEE Transactions on Mobile Computing* 16.10 (Oct. 2017), pp. 2682–2696. ISSN: 1536-1233. DOI: 10.1109/TMC.2016.2637901. URL: http://ieeexplore.ieee.org/document/7779102/.

[37] Rahul Goel et al. "Mobility-based SIR model for complex networks: with case study Of COVID-19". In: *Social Network Analysis and Mining* 11.1 (Dec. 2021), pp. 1–18. ISSN: 1869-5450. DOI: 10.1007/S13278-021-00814-3/FIGURES/12. URL: https://link.springer.com/article/10.1007/s13278-021-00814-3.

[38] Marta C. González, César A. Hidalgo, and Albert-László Barabási. "Understanding individual human mobility patterns". In: *Nature* 453.7196 (June 2008), pp. 779–782. ISSN: 0028-0836. DOI: 10.1038/nature06958. URL: http://www.nature.com/doifinder/10.1038/nature06958.

[39] Marco Gramaglia and Marco Fiore. "Hiding mobile traffic fingerprints with GLOVE". In: *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT 2015* (Dec. 2015). DOI: 10.1145/2716281.2836111.

[40] David Gundlegård et al. "Travel demand estimation and network assignment based on cellular network data". In: *Computer Communications* 95 (Dec. 2016), pp. 29–42. ISSN: 0140-3664. DOI: 10.1016/J.COMCOM.2016.04.015.

[41] Emir Halepovic and Carey Williamson. "Characterizing and modeling user mobility in a cellular data network". In: *PE-WASUN'05 - Proceedings of the Second ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks* (2005), pp. 71–78. DOI: 10.1145/1089803.1089969.

[42] Hannah Ritchie and Max Roser. *Urbanization*. 2018. URL: https://ourworldindata.org/urbanization.

[43] Samiul Hasan, Xianyuan Zhan, and Satish V. Ukkusuri. "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, New York, USA: ACM Press, 2013, p. 1. ISBN: 9781450323314. DOI: 10.1145/2505821.2505823. URL: http://dl.acm.org/citation.cfm?doid=2505821.2505823.

[44] Andrea Hess, Ian Marsh, and Daniel Gillblad. "Exploring communication and mobility behavior of 3G network users and its temporal consistency". In: *IEEE International Conference on Communications* 2015-September (Sept. 2015), pp. 5916–5921. DOI: 10.1109/ICC.2015.7249265.

[45] Thomas Holleczek et al. "Traffic measurement and route recommendation system for mass rapid transit (MRT)". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2015-August (Aug. 2015), pp. 1859–1868. DOI: 10.1145/2783258.2788590.

[46] Haosheng Huang, Yi Cheng, and Robert Weibel. "Transport mode detection based on mobile phone network data: A systematic review". In: *Transportation Research Part C: Emerging Technologies* (Feb. 2019). ISSN: 0968-090X. DOI: 10.1016/J.TRC.2019.02.008. URL: https://www.sciencedirect.com/science/article/pii/S0968090X1831369X.

[47] Jin Huang and Ming Xiao. "State of the art on road traffic sensing and learning based on mobile user network log data". In: *Neurocomputing* 278 (Feb. 2018), pp. 110–118. ISSN: 18728286. DOI: 10.1016/J.NEUCOM.2017.03.096.

[48] Tin Ying Hui. "Investigating the Use of Anonymous Cellular Data for Intercity Travel Patterns in Alberta". In: (2017). DOI: 10.7939/R3QN5ZR65. URL: https://era.library.ualberta.ca/items/f13ee1b6-4bd7-4b57-801b-bd1025bf9945.

[49] Md. Shahadat Iqbal et al. "Development of origin–destination matrices using mobile phone call data". In: *Transportation Research Part C: Emerging Technologies* 40 (Mar. 2014), pp. 63–74. ISSN: 0968-090X. DOI: 10.1016/J.TRC.2014.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0968090X14000059/.

[50] Sibren Isaacman et al. "Ranges of human mobility in Los Angeles and New York". In: *2011 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2011* (2011), pp. 88–93. DOI: 10.1109/PERCOMW.2011.5766977.

[51] Masahiko Itoh et al. "Visual Exploration of Changes in Passenger Flows and Tweets on Mega-City Metro Network". In: *IEEE Transactions on Big Data* 2.1 (Apr. 2016), pp. 85–99. DOI: 10.1109/TBDATA.2016.2546301.

[52] George R. Jagadeesh and Thambipillai Srikanthan. "Online Map-Matching of Noisy and Sparse Location Data with Hidden Markov and Route Choice Models". In: *IEEE Transactions on Intelligent Transportation Systems* 18.9 (2017), pp. 2423–2434. ISSN: 15249050. DOI: 10.1109/TITS.2017.2647967.

[53] Andreas Janecek et al. "Cellular Data Meet Vehicular Traffic Theory: Location Area Updates and Cell Transitions for Travel Time Estimation". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 361–370. ISBN: 9781450312240. DOI: 10.1145/2370216.2370272. URL: https://doi.org/10.1145/2370216.2370272.

[54] Maxim Janzen et al. "Closer to the total? Long-distance travel of French mobile phone users". In: *Travel Behaviour and Society* 11 (Apr. 2018), pp. 31–42. ISSN: 2214367X. DOI: 10.1016/j.tbs.2017.12.001.

[55] Shan Jiang, Joseph Ferreira, and Marta C. Gonzalez. "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore". In: *IEEE Transactions on Big Data* 3.2 (June 2017), pp. 208–219. ISSN: 2332-7790. DOI: 10.1109/TBDATA.2016.2631141. URL: http://ieeexplore.ieee.org/document/7755745/.

[56] Meihan Jin et al. *A Conceptual and Semantic Modelling Approach for the Representation and Exploration of Human Trajectories Thèse soutenue le 18 septembre, 2017 devant le jury composé de*. Tech. rep. URL: https://hal.archives-ouvertes.fr/tel-01591177v1/document.

[57] D. S. Johnson and K. A. Niemi. "On Knapsacks, Partitions, and a new Dynamic Programming Technique For Trees." In: *Mathematics of Operations Research* 8.1 (1983), pp. 1–14. DOI: 10.1287/MOOR.8.1.1.

[58] 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)*. IEEE, 2001, pp. 23–27. ISBN: 0-7803-7194-1. DOI: 10.1109/ITSC.2001.948623. URL: http://ieeexplore.ieee.org/document/948623/.

[59] Chaogui Kang et al. "Intra-urban human mobility patterns: An urban morphology perspective". In: *Physica A: Statistical Mechanics and its Applications* 391.4 (Feb. 2012), pp. 1702–1717. ISSN: 03784371. DOI: 10.1016/J.PHYSA.2011.11.005. URL: https://ideas.repec.org/a/eee/phsmap/v391y2012i4p1702-1717.html.

[60] Panagiota Katsikouli et al. "Characterizing and Removing Oscillations in Mobile Phone Location Data". In: (2019), pp. 1–9. DOI: 10.1109/WoWMoM.2019.8793034{\"{i}}. URL: https://hal.inria.fr/hal-02110719.

[61] W. O. Kermack and A. G. McKendrick. "Contributions to the mathematical theory of epidemics 2. The problem of endemicity". In: *Bulletin of Mathematical Biology 1991 53:1* 53.1 (Mar. 1991), pp. 57–87. ISSN: 1522-9602. DOI: 10.1007/BF02464424. URL: https://link.springer.com/article/10.1007/BF02464424.

[62] Ghazaleh Khodabandelou et al. "Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata". In: *IEEE Transactions on Mobile Computing* (Sept. 2018). ISSN: 15580660. DOI: 10.1109/TMC.2018.2871156.

[63] Henk A.L. Kiers and Iven Van Mechelen. "Three-way component analysis: Principles and illustrative application". In: *Psychological Methods* 6.1 (2001), pp. 84–110. ISSN: 1082989X. DOI: 10.1037/1082-989X.6.1.84. URL: /doiLanding?doi=10.1037%2F1082-989X.6.1.84.

[64] Tamara G. Kolda and Brett W. Bader. *Tensor decompositions and applications*. Aug. 2009. DOI: 10.1137/07070111X.

[65] Xiangjie Kong et al. "LoTAD: long-term traffic anomaly detection based on crowdsourced bus trajectory data". In: *World Wide Web 2017 21:3* 21.3 (Aug. 2017), pp. 825–847. ISSN: 1573-1413. DOI: 10.1007/S11280-017-0487-4. URL: https://link.springer.com/article/10.1007/s11280-017-0487-4.

[66] Xiangjie Kong et al. "Urban traffic congestion estimation and prediction based on floating car trajectory data". In: *Future Generation Computer Systems* 61 (Aug. 2016), pp. 97–107. ISSN: 0167-739X. DOI: 10.1016/J.FUTURE.2015.11.013.

[67] A. Kuppam et al. "Special events travel surveys and model development". In: *Transportation Letters* 5.2 (2013), pp. 67–82. ISSN: 19427875. DOI: 10.1179/1942786713Z.0000000007. URL: https://www.tandfonline.com/doi/abs/10.1179/1942786713Z.0000000007.

[68] Richard J. Lee, Ipek N. Sener, and James A. Mullins. "An evaluation of emerging data collection technologies for travel demand modeling: from research to practice". In: *Transportation Letters* 8.4 (Jan. 2016), pp. 181–193. ISSN: 19427875. DOI: 10.1080/19427867.2015.1106787. URL: https://www.tandfonline.com/doi/abs/10.1080/19427867.2015.1106787.

[69] Maxime Lenormand et al. "Cross-Checking Different Sources of Mobility Information". In: *PLOS ONE* 9.8 (Aug. 2014), e105184. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0105184. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105184.

[70] Ilias Leontiadis et al. "From cells to streets: Estimating mobile paths with cellular-side data". In: *CoNEXT 2014 - Proceedings of the 2014 Conference on Emerging Networking Experiments and Technologies*. New York, NY, USA: Association for Computing Machinery, Inc, Dec. 2014, pp. 121–132. ISBN: 9781450332798. DOI: 10.1145/2674005.2674982. URL: https://dl.acm.org/doi/10.1145/2674005.2674982.

[71] Mingxiao Li et al. "Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data". In: *Computers, Environment and Urban Systems* 77 (Sept. 2019), p. 101346. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2019.101346.

[72] Yuting Liang and Reza Samavi. "Optimization-based k-anonymity algorithms". In: *Computers & Security* 93 (June 2020), p. 101753. ISSN: 0167-4048. DOI: 10.1016/J.COSE.2020.101753.

[73] Zhongqiu Liu and Chao Yang. "Exploring group-level human mobility from location-based social media check-in data". In: *Proceedings of 5th IEEE Conference on Ubiquitous Positioning, Indoor Navigation and Location-Based Services, UPINLBS 2018* (Dec. 2018). DOI: 10.1109/UPINLBS.2018.8559796.

[74] An Luo, Shenghua Chen, and Bin Xv. "Enhanced Map-Matching Algorithm with a Hidden Markov Model for Mobile Phone Positioning". In: *ISPRS International Journal of Geo-Information* 6.11 (2017), p. 327. ISSN: 2220-9964. DOI: 10.3390/ijgi6110327. URL: http://www.mdpi.com/2220-9964/6/11/327.

[75] Jingtao Ma et al. "Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data". In: *International Journal of Transportation Science and Technology* 2.3 (Sept. 2013), pp. 183–204. ISSN: 2046-0430. DOI: 10.1260/2046-0430.2.3.183. URL: https://www.sciencedirect.com/science/article/pii/S204604301630140X.

[76] Ed Manley and Adam Dennett. "New Forms of Data for Understanding Urban Activity in Developing Countries". In: *Applied Spatial Analysis and Policy* 12.1 (Mar. 2019), pp. 45–70. ISSN: 18744621. DOI: 10.1007/S12061-018-9264-8/FIGURES/9. URL: https://link.springer.com/article/10.1007/s12061-018-9264-8.

[77] Eduardo Martínez-Montes et al. "Concurrent EEG/fMRI analysis by multiway Partial Least Squares". In: *NeuroImage* 22.3 (July 2004), pp. 1023–1034. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2004.03.038.

[78] Benoît Matet et al. *A Lightweight Approach for Origin-Destination Matrix Anonymization*. 2021. ISBN: 9782875870827. URL: http://www.i6doc.com/en/..

[79] Yves-Alexandre de Montjoye et al. "Unique in the Crowd: The privacy bounds of human mobility". In: *Scientific Reports 2013 3:1* 3.1 (Mar. 2013), pp. 1–5. ISSN: 2045-2322. DOI: 10.1038/srep01376. URL: https://www.nature.com/articles/srep01376.

[80] Diala Naboulsi et al. "Large-Scale Mobile Traffic Analysis: A Survey". In: *IEEE Communications Surveys & Tutorials* 18.1 (2016), pp. 124–161. ISSN: 1553-877X. DOI: 10.1109/COMST.2015.2491361. URL: http://ieeexplore.ieee.org/document/7299258/.

[81] 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*. New York, New York, USA: ACM Press, 2009, p. 336. ISBN: 9781605586496. DOI: 10.1145/1653771.1653818. URL: http://portal.acm.org/citation.cfm?doid=1653771.1653818.

[82] Dragan Obradovic, Henning Lenz, and Markus Schupfner. "Fusion of Map and Sensor Data in a Modern Car Navigation System". In: *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* 45.1-2 (Nov. 2006), pp. 111–122. ISSN: 0922-5773. DOI: 10.1007/s11265-006-9775-4. URL: http://link.springer.com/10.1007/s11265-006-9775-4.

[83] Luis E. Olmos et al. "Macroscopic dynamics and the collapse of urban traffic". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.50 (Dec. 2018), pp. 12654–12661. ISSN: 10916490. DOI: 10.1073/pnas.1800474115. URL: www.pnas.org/cgi/doi/10.1073/pnas.1800474115.

[84] Vasyl Palchykov et al. "Inferring human mobility using communication patterns". In: *Scientific Reports 2014 4:1* 4.1 (Aug. 2014), pp. 1–6. ISSN: 2045-2322. DOI: 10.1038/srep06174. URL: https://www.nature.com/articles/srep06174.

[85] Luca Pappalardo et al. *An individual-level ground truth dataset for home location detection*. Tech. rep. 2020.

[86] Utpal Paul et al. "Understanding traffic dynamics in cellular data networks". In: *Proceedings - IEEE INFOCOM* (2011), pp. 882–890. DOI: 10.1109/INFCOM.2011.5935313.

[87] Chengbin Peng et al. "Collective human mobility pattern from taxi trips in urban area". In: *PLoS ONE* 7.4 (Apr. 2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0034487.

[88] Santi Phithakkitnukoon et al. "Inferring social influence in transport mode choice using mobile phone data". In: *EPJ Data Science 2017 6:1* 6.1 (June 2017), pp. 1–29. ISSN: 2193-1127. DOI: 10.1140/EPJDS/S13688-017-0108-6. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-017-0108-6.

[89] Raymond H. Putra et al. "Map matching with Hidden Markov Model on sampled road network". In: *undefined* (2012).

[90] Zhou Qin et al. "CellPred: A behavior-aware scheme for cellular data usage prediction". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1 (Mar. 2020), pp. 1–24. ISSN: 24749567. DOI: 10.1145/3380982. URL: https://dl.acm.org/doi/10.1145/3380982.

[91] Zhou Qin et al. "EXIMIUS: A measurement framework for explicit and implicit urban traffic sensing". In: *SenSys 2018 - Proceedings of the 16th Conference on Embedded Networked Sensor Systems*. New York, NY, USA: Association for Computing Machinery, Inc, Nov. 2018, pp. 1–14. ISBN: 9781450359528. DOI: 10.1145/3274783.3274850. URL: https://dl.acm.org/doi/10.1145/3274783.3274850.

[92] Yingchun Qu, Hang Gong, and Pu Wang. "Transportation Mode Split with Mobile Phone Data". In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* 2015-October (Oct. 2015), pp. 285–289. DOI: 10.1109/ITSC.2015.56.

[93] Mohammed A. Quddus, Robert B. Noland, and Washington Y. Ochieng. "A High Accuracy Fuzzy Logic Based Map Matching Algorithm for Road Transport". In: *Journal of Intelligent Transportation Systems* 10.3 (Sept. 2006), pp. 103–115. ISSN: 1547-2450. DOI: 10.1080/15472450600793560. URL: https://www.tandfonline.com/doi/full/10.1080/15472450600793560.

[94] 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". In: *Transportation Research Part C: Emerging Technologies* 15.5 (2007), pp. 312–328. ISSN: 0968090X. DOI: 10.1016/j.trc.2007.05.002.

[95] Injong Rhee et al. "On the Levy-Walk Nature of Human Mobility". In: *IEEE/ACM Transactions on Networking* 19.3 (June 2011), pp. 630–643. ISSN: 1063-6692. DOI: 10.1109/TNET.2011.2120618. URL: http://ieeexplore.ieee.org/document/5750071/.

[96] Nadine Rieser-Schussler. "Capitalising modern data sources for observing and modelling transport behaviour". In: *Transportation Letters* 4.2 (2013), pp. 115–128. ISSN: 19427875. DOI: 10.3328/TL.2012.04.02.115-128. URL: https://www.tandfonline.com/doi/abs/10.3328/TL.2012.04.02.115-128.

[97] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257?via%3Dihub.

[98] Christian M Schneider et al. "Unravelling daily human mobility motifs". In: *Journal of The Royal Society Interface* 10.84 (2013), p. 20130246. DOI: 10.1098/rsif.2013.0246. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0246.

[99] Gunnar Schulze, Christopher Horn, and Roman Kern. "Map-Matching Cell Phone Trajectories of Low Spatial and Temporal Accuracy". In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* 2015-Octob (2015), pp. 2707–2714. DOI: 10.1109/ITSC.2015.435.

[100] M. Shafiei, M. Nazemi, and S. Seyedabrishami. "Estimating time-dependent origin–destination demand from traffic counts: extended gradient method". In: *http://dx.doi.org/10.1179/1942787514Y.0000000048* 7.4 (Sept. 2014), pp. 210–218. ISSN: 19427875. DOI: 10.1179/1942787514Y.0000000048. URL: https://www.tandfonline.com/doi/abs/10.1179/1942787514Y.0000000048.

[101] Zbigniew Smoreda, Ana-Maria Olteanu-Raimond, and Thomas Couronné. "Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment". In: *Transport Survey Methods: Best Practice for Decision Making* (Jan. 2013), pp. 745–768. DOI: 10.1108/9781781902882-041.

[102] Chaoming Song et al. "Limits of Predictability in Human Mobility". In: *Science* 327.5968 (2010), pp. 1018–1021. ISSN: 0036-8075. DOI: 10.1126/science.1177170. URL: https://science.sciencemag.org/content/327/5968/1018.

[103] Chaoming Song et al. "Modelling the scaling properties of human mobility". In: *Nature Physics 2010 6:10* 6.10 (Sept. 2010), pp. 818–823. ISSN: 1745-2481. DOI: 10.1038/nphys1760. URL: https://www.nature.com/articles/nphys1760.

[104] Yiwei Song et al. "Miff Human mobility extractions with cellular signaling data under spatio-Temporal uncertainty". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.4 (Dec. 2020), pp. 1–19. ISSN: 24749567. DOI: 10.1145/3432238. URL: https://dl.acm.org/doi/10.1145/3432238.

[105] Peter R. Stopher and Stephen P. Greaves. "Household travel surveys: Where are we going?" In: *Transportation Research Part A: Policy and Practice* 41.5 (June 2007), pp. 367–381. ISSN: 0965-8564. DOI: 10.1016/J.TRA.2006.09.005.

[106] Lijun Sun and Kay W. Axhausen. "Understanding urban mobility patterns with a probabilistic tensor factorization framework". In: *Transportation Research Part B: Methodological* 91 (Sept. 2016), pp. 511–524. ISSN: 0191-2615. DOI: 10.1016/J.TRB.2016.06.011. URL: https://www.sciencedirect.com/science/article/pii/S0191261516300261.

[107] Lijun Sun and Kay W. Axhausen. "Understanding urban mobility patterns with a probabilistic tensor factorization framework". In: *Transportation Research Part B: Methodological* 91 (Sept. 2016), pp. 511–524. ISSN: 01912615. DOI: 10.1016/j.trb.2016.06.011.

[108] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. "A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis". In: *IEEE Transactions on Knowledge and Data Engineering* 22.2 (Feb. 2010), pp. 179–192. DOI: 10.1109/TKDE.2009.85.

[109] Abdel Aziz Taha and Allan Hanbury. "An Efficient Algorithm for Calculating the Exact Hausdorff Distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.11 (Nov. 2015), pp. 2153–2163. ISSN: 01628828. DOI: 10.1109/TPAMI.2015.2408351.

[110] Huachun Tan et al. "Short-Term Traffic Prediction Based on Dynamic Tensor Completion". In: *IEEE Transactions on Intelligent Transportation Systems* 17.8 (Aug. 2016), pp. 2123–2133. ISSN: 15249050. DOI: 10.1109/TITS.2015.2513411.

[111] Jinhui Tang et al. "Cross-space affinity learning with its application to movie recommendation". In: *IEEE Transactions on Knowledge and Data Engineering* 25.7 (2013), pp. 1510–1519. DOI: 10.1109/TKDE.2012.87.

[112] Haiyan Tao et al. "Re-examining urban region and inferring regional function based on spatial–temporal interaction". In: *International Journal of Digital Earth* 12.3 (Mar. 2019), pp. 293–310. ISSN: 17538955. DOI: 10.1080/17538947.2018.1425490.

[113] Arvind Thiagarajan et al. *Accurate, low-energy trajectory mapping for mobile devices*. 2011. URL: https://dl.acm.org/citation.cfm?id=1972485.

[114] Michele Tizzoni et al. "On the Use of Human Mobility Proxies for Modeling Epidemics". In: *PLOS Computational Biology* 10.7 (2014), e1003716. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1003716. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003716.

[115] Jameson L. Toole et al. "Inferring land use from mobile phone activity". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, New York, USA: ACM Press, 2012, pp. 1–8. ISBN: 9781450315425. DOI: 10.1145/2346496.2346498. URL: http://dl.acm.org/citation.cfm?doid=2346496.2346498.

[116] Jameson L. Toole et al. "The path most traveled: Travel demand estimation using big data resources". In: *Transportation Research Part C: Emerging Technologies* 58 (Sept. 2015), pp. 162–177. ISSN: 0968-090X. DOI: 10.1016/J.TRC.2015.04.022. URL: https://www.sciencedirect.com/science/article/pii/S0968090X15001631.

[117] Roberto Trasarti et al. "Discovering urban and country dynamics from mobile phone data with spatial correlation patterns". In: *Telecommunications Policy* 39.3-4 (2015), pp. 347–362. ISSN: 03085961. DOI: 10.1016/J.TELPOL.2013.12.002. URL: https://hal.archives-ouvertes.fr/hal-03346123.

[118] M. Alex O. Vasilescu and Demetri Terzopoulos. "Multilinear Analysis of Image Ensembles: TensorFaces". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2350 (2002), pp. 447–460. ISSN: 16113349. DOI: 10.1007/3-540-47969-4{_}30. URL: https://link.springer.com/chapter/10.1007/3-540-47969-4_30.

[119] A. Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13.2 (Apr. 1967), pp. 260–269. ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1054010. URL: http://ieeexplore.ieee.org/document/1054010/.

[120] Dianhai Wang et al. "Nonnegative tensor decomposition for urban mobility analysis and applications with mobile phone data". In: *Transportmetrica A: Transport Science* (2019). ISSN: 23249943. DOI: 10.1080/23249935.2019.1692961.

[121] Jingyuan Wang et al. "Discovering urban spatio-temporal structure from time-evolving traffic networks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8709 LNCS. Springer Verlag, 2014, pp. 93–104. ISBN: 9783319111155. DOI: 10.1007/978-3-319-11116-2{_}9.

[122] Jingyuan Wang et al. "Understanding Urban Dynamics via Context-aware Tensor Factorization with Neighboring Regularization". In: (Apr. 2019). DOI: 10.1109/TKDE.2019.2915231. URL: http://arxiv.org/abs/1905.00702http://dx.doi.org/10.1109/TKDE.2019.2915231.

[123] Ming-Heng Wang et al. "Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data". In: *International Journal of Intelligent Transportation Systems Research* 11.2 (May 2013), pp. 76–86. ISSN: 1348-8503. DOI: 10.1007/s13177-013-0058-8. URL: http://link.springer.com/10.1007/s13177-013-0058-8.

[124] Ming Heng Wang et al. "Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data". In: *International Journal of Intelligent Transportation Systems Research 2013 11:2* 11.2 (Apr. 2013), pp. 76–86. ISSN: 1868-8659. DOI: 10.1007/S13177-013-0058-8. URL: https://link.springer.com/article/10.1007/s13177-013-0058-8.

[125] Yilun Wang, Yu Zheng, and Yexiang Xue. "Travel time estimation of a path using sparse trajectories". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2014), pp. 25–34. DOI: 10.1145/2623330.2623656.

[126] Christopher E White, David Bernstein, and Alain L Kornhauser. "Some map matching algorithms for personal navigation assistants". In: *Transportation Research Part C: Emerging Technologies* 8.1-6 (Feb. 2000), pp. 91–108. ISSN: 0968-090X. DOI: 10.1016/S0968-090X(00)00026-7. URL: https://www.sciencedirect.com/science/article/pii/S0968090X00000267.

[127] Peter Widhalm et al. "Discovering urban activity patterns in cell phone data". In: *Transportation* 42.4 (2015). ISSN: 15729435. DOI: 10.1007/s11116-015-9598-x.

[128] Jean Wolf, Marcelo Oliveira, and Miriam Thompson. "Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey:" in: *Transportation Research Record: Journal of the Transportation Research Board* 1854 (Jan. 2003), pp. 189–198. ISSN: 03611981. DOI: 10.3141/1854-21. URL: https://journals.sagepub.com/doi/10.3141/1854-21.

[129] Wei Wu et al. "Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling". In: *2014 IEEE 15th International Conference on Mobile Data Management*. IEEE, July 2014, pp. 321–328. ISBN: 978-1-4799-5705-7. DOI: 10.1109/MDM.2014.46. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6916937.

[130] Wei Wu et al. "Studying Intercity Travels and Traffic Using Cellular Network Data". In: (2013).

[131] Yangyang Xu. "Alternating proximal gradient method for sparse nonnegative Tucker decomposition". In: *Mathematical Programming Computation 2014 7:1* 7.1 (May 2014), pp. 39–70. ISSN: 1867-2957. DOI: 10.1007/S12532-014-0074-Y. URL: https://link.springer.com/article/10.1007/s12532-014-0074-y.

[132] Yanyan Xu, Riccardo Di Clemente, and Marta C. González. "Understanding vehicular routing behavior with location-based service data". In: *EPJ Data Science* 10.1 (Dec. 2021), pp. 1–17. ISSN: 21931127. DOI: 10.1140/epjds/s13688-021-00267-w. URL: https://doi.org/10.1140/epjds/s13688-021-00267-w.

[133] Xiao Yong Yan et al. "Diversity of individual mobility patterns and emergence of aggregated scaling laws". In: *Scientific Reports* 3 (Sept. 2013). ISSN: 20452322. DOI: 10.1038/srep02678.

[134] Hao Yang and Hesham Rakha. "A novel approach for estimation of dynamic from static origin–destination matrices". In: *Transportation letters* 11.4 (July 2017), pp. 219–228. ISSN: 19427875. DOI: 10.1080/19427867.2017.1336353. URL: https://www.tandfonline.com/doi/abs/10.1080/19427867.2017.1336353.

[135] Meng Yu. "Improved positioning of land vehicle in its using digital map and other accessory information". In: *The Hong Kong Polytechnic University* (2006). URL: http://ira.lib.polyu.edu.hk/handle/10397/2858.

[136] Chao Zhang, Quan Yuan, and Jiawei Han. *Bringing Semantics to Spatiotemporal Data Mining*. Tech. rep. URL: http://chaozhang.org/slides/slides-icde17.pdf.

[137] Chen Zhao, An Zeng, and Chi Ho Yeung. "Characteristics of human mobility patterns revealed by high-frequency cell-phone position data". In: *EPJ Data Science* 10.1 (Dec. 2021), p. 5. ISSN: 21931127. DOI: 10.1140/epjds/s13688-021-00261-2. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00261-2.

[138] Yi Zhao et al. "CellTrans". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.3 (Sept. 2019), pp. 1–26. ISSN: 2474-9567. DOI: 10.1145/3351283. URL: https://dl.acm.org/doi/10.1145/3351283.

[139] Yi Zhao et al. "Urban scale trade area characterization for commercial districts with cellular footprints". In: *ACM Transactions on Sensor Networks* 16.4 (Oct. 2020), pp. 1–20. ISSN: 15504867. DOI: 10.1145/3412372. URL: https://dl.acm.org/doi/10.1145/3412372.

[140] Ziliang Zhao et al. "Understanding the bias of call detail records in human mobility research". In: *International Journal of Geographical Information Science* (2016). ISSN: 13623087. DOI: 10.1080/13658816.2015.1137298.

[141] Yu Zheng. "Trajectory Data Mining: An Overview". In: *ACM Trans. On Intelligent Systems and Technology* 6.3 (2015). DOI: 10.1145/2743025. URL: http://dx.doi.org/10.1145/2743025.

[142] Yu Zheng et al. "Mining interesting locations and travel sequences from GPS trajectories". In: *WWW'09 - Proceedings of the 18th International World Wide Web Conference* (2009), pp. 791–800. DOI: 10.1145/1526709.1526816.

[143] Yu Zheng et al. "Urban computing: Concepts, methodologies, and applications". In: *ACM Trans. Intell. Syst. Technol* 5.38 (2014). DOI: 10.1145/2629592. URL: http://dx.doi.org/10.1145/2629592.

[144] George Kingsley Zipf. "The P 1 P 2 D Hypothesis: On the Intercity Movement of Persons". In: *American Sociological Review* 11.6 (Dec. 1946), p. 677. DOI: 10.2307/2087063.