École des Ponts
ParisTech

THÈSE DE DOCTORAT
de l'École des Ponts ParisTech

FLOATING CAR DATA MINING TO FEATURE OUT MOBILITY PATTERNS: INDIVIDUAL-CENTERED AND PLACE-BASED ANALYSES

École Doctorale Ville, Transports et Territoires

Spécialité : Transport

Thèse préparée au Laboratoire Ville Mobilité Transport
UMR T9403, École des Ponts, Univ Gustave Eiffel

Thèse soutenue le 27 Janvier 2022, par

Danyang SUN

Dirigée par Fabien LEURENT et co-encadrée par Xiaoyan XIE

Composition du jury :

Cristina, PRONELLO Professor, Politecnico di Torino	*Rapporteur*
Rosaldo J.F., ROSSETTI Associate Professor, Universidade do Porto	*Rapporteur*
Éric, CORNELIS Senior Research Associate, Université de Namur	*Examinateur* *et Président du jury*
Daniel J., GRAHAM Professor, Imperial College London	*Examinateur*
Jimmy, ARMOOGUM Senior Researcher, Université Gustave Eiffel	*Examinateur*
Fabien, LEURENT Professor, École des Ponts ParisTech	*Directeur de thèse*
Xiaoyan, XIE Research Associate, CY Cergy Paris Université	*Co-encadrant de thèse*

ABSTRACT

The presence and movement of human beings in space and time constitute their mobility: it is a physical phenomenon and also a socioeconomic phenomenon, since people choose their locations and their trips between them to satisfy their needs and desires. The scientific knowledge of mobility as a physical phenomenon involves causalities and patterns. At the individual level, occupation of places and making trips along the day, with activity purposes and transport means, give rise to statistical regularities from day to day – most notably the home to work commuting. Then, in a statistical population of individuals, there are statistical regularities at the interindividual level – in other words, clusters of individuals on the basis of their mobility profiles. Furthermore, the spatial occupation and movements of individuals aggregate in local volumes at the zone level and origin-destination flows between zones: such aggregate patterns interplay with local land-use and the related territorial configuration, contributing to the land-use and transport interaction.

Owing to the recent surge in sensing technologies, geolocation has become a ubiquitous service and trajectory data are now massively available, thereby empowering the observation of human mobility at the individual level, with large spatial and temporal coverage and possibly significant penetration rates. Among various information sources, Floating Car Data (FCD) pertain to vehicle-based mobility and yield discretized trajectories composed of digital traces of positions in space and time for the "vehicle" entity, accurate geolocations and timestamps, and also speeds and driving directions. Most of the academic literature has concentrated on methods for processing raw trajectory data in order to characterize traffic conditions or to model traffic-relevant parameters from an engineering aspect. Only a few studies have shifted the focus of FCD data mining towards semantic-oriented excavation by exploring behavioral representations and interpreting activity contexts, possibly in relation to land-use descriptors.

This thesis aims to explore and analyze mobility patterns (in the sense of motifs and forms) by leveraging FCD to contribute a better understanding of vehicle-based movements. More specifically, mobility patterns are studied at two levels: the individual level of human behaviors of the "authors" of the trajectories, and the more global level of spatial relations and structure on the basis of aggregated mobility features in space. Furthermore, another fold of the thesis objective is to build up methodological approaches for trajectory mining, contributing to

broadening the way of using trajectory data in mobility analytics. This concerns the algorithms and methods to be developed for processing the trajectory, reconstructing the information, and developing analytical models, aiming to translate the data into understandable mobility knowledge.

The first part of the thesis pertains to individual mobility. Three research questions are addressed, each in a specific chapter: first, daily patterns of trip-making at the vehicle level, second the multiday regularity of individual place frequentation, third the estimation of travel times from individual trajectories. Chapter 2 discovers vehicle usage patterns of individuals based on their digital footprints. It aims to expand the semantic exploration of mobility patterns by capturing how vehicles are used in people's daily mobility. A topic-modeling approach is developed for the discovery by regarding the mobility profiles of individuals as documents, trips as words, and the vehicle usage types as latent topics to be determined, thereby constituting a vehicle usage typology. Then, Chapter 3 investigates significant places to mobility makers, by identifying the "anchoring" geolocations and further extrapolating their meaningful representations such as homes, workplaces, and other secondary places. It takes the advantage of multiple-day tracking of FCD to investigate individual mobility regularities and habits by using unsupervised learning. Next, Chapter 4 proposes a novel approach for travel time estimation by building a stochastic model to exploit FCD materials. It aims to allow for simple but robust estimation of the key factors in traffic conditions, along with the goal to fill the research gap in measuring the reliability. This issue is recalled particularly as it plays a vital role in individual decision making and also the massive FCD traces serve like sensors over the network, making it especially suitable and powerful in addressing such a problem.

The second part of the thesis pertains to places and spatial relations. Three research questions are also addressed, again each in a specific chapter: first the functions of space by mobility activities, second the territory structure by core-periphery patterns, third the estimation of Origin-Destination flows by leveraging digital trajectory data. Chapter 5 reveals the functional occupation of urban areas by looking at related vehicle movements. A multi-view cooperative clustering is designed to identify the space typology by integrating different facets of characteristics in terms of the composition of activity visits, temporal flows of trip generation and attraction, and spatial connections in trip distance distribution. Then, Chapter 6 studies the

spatial organization of a territory, with a particular emphasis on the fundamental jobs-housing spatial relations. It establishes a data-driven method to recognize employment core areas by spatial density distribution and identify corresponding residential catchment areas by core-periphery patterns. Bonded spatial communities are identified by applying graph-partition algorithms to find sub-regions with denser inner exchanges in the home-work network. Third, Chapter 7 investigates the spatial interaction between places. It deals with the estimation of the Origin-Destination matrix flows based on two kinds of data: vehicle trajectory data and local traffic counts, along with a developed Bayesian assignment framework to account for the heterogeneous sampling rate issues of such data. This study aims to leverage the modern data to quantify the spatial interactions in traffic flows, by avoiding the conventional sophisticated process of traffic assignment modeling.

In a concluding chapter, the research outreaches are summarized and the limitations are discussed, along with future perspectives. Overall, this thesis expands the “mobility analytics” especially on “pattern recognition” from a data mining standing point. It contributes to overcoming traditional limitations on extensive mobility analysis in terms of inter-day variations and large-scale observations by employing massive digital trajectories and artificial intelligence. Through various applications, this thesis shows the feasibility of mining semantic context behind individual mobility at a micro-level and the possibility of capturing grouped phenomena reflected in geographical spaces at a macro-level. Empirical findings were obtained in terms of vehicle usage ways, mobility regularities, spatial functions, and place relations via case studies using real-world data. Fundamental metrics in traffic conditions: travel times and travel demand flows were also proved obtainable from those trajectories with good applicability and effectiveness. However, this thesis pays particular attention to vehicle-related mobility based on FCD. Future work can bring with other modes of transportation to have a more complete investigation of the mobility system. The discovered patterns and trends may also be further investigated with examination on the influencing factors for exploring the explanations.

Keywords: Mobility Pattern; Trajectory Data Mining; Floating Car Data; Machine Learning; Statistical Regularities of Individual Mobility Behavior; Mobility-Based Spatial Structure

Acknowledgment

Writing this thesis has not been easy but a truly memorable accomplishment of my life, especially being through the journey with most of the time under the global strike of the Covid pandemic, which has left a unique mark on the experience, a stamp of this special epoch. Fortunately, the journey has not been walked through alone in my own efforts. With the tremendous support and love from my surroundings, I have been accompanied both practically and emotionally to accomplish the steps I've made until today.

My first and deepest gratitude goes to my supervisor, Prof. Fabien Leurent, who has inspired me by his devotion and dedication to research. His insights, guidance, patience, and support, greatly assisted me in completing this thesis. I am, therefore, fortunate to have him as my supervisor. I would also like to thank Dr. Xiaoyan Xie, as my co-supervisor, for her continued guidance, suggestions and support of my research. Her insights helped me overcome the research challenges on the way.

I would like to acknowledge my jury members, Prof. Pronello, Prof. Rossetti, Prof. Armoogum, Prof. Cornelis and Prof. Graham, for their commitment to read my work and participate in my final defense committee. I am sincerely grateful for all their commitments that contribute to the improvement of this thesis, which further aid the completion of the Ph.D. study and inspire me in the future directions.

I would like to thank my hosting lab, Laboratoire Ville Mobilité Transport, and all my colleagues (especially, Sophie C., Virginie D., Sandrine V., Pierre Z., Biao Y. Sheng L. and Wei K.) for their kind support, help, and fruitful exchange during my entire study. The journey to a foreign country was challenging at the beginning, but with them, after 3 years, I am happy to say that I haven't felt alone. Paris has become an unforgettable place that stays with me in my life, with the connections to them.

I would also like to give my thanks to the research chair – Lab recherche environnement VINCI & ENPC, for supporting and funding the research, with particular gratitude to Dr. Nicolas Coulombel for hosing the chair and assisting me with the research activities. I would also like to thank Aleia and Coyote, for the data and assistance, and Chair ENPC-IdFM in facilitating the research.

Finally, this is truly the opportunity for me to write down the heartfelt gratitude that has always been kept inside me to my dear family, who has been with me for all the journeys since I was born, my girlfriend, who has been the source of my strength and accompanied me during the last 10 years, and a long list of my beloved friends. The thesis wouldn't be accomplished, or not even initiated, without the encouragement, support, and love from them. The thesis is for them all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1. Context and Motivation: Interests in Human Mobility

The term mobility commonly refers to the entity's presence in space along time and the movement from one space-time point to another. The needs and desires of performing various social activities motivate human beings to choose their locations and travel between places. These movements, therefore, reflect not only physical phenomena but also social phenomena. While mobility phenomena are not formed randomly. Causalities are constituted over time with various behaviors, generating the mobility patterns which represent rules, trends, or relationships that are recognizable or predictable. Until now, observing the human mobility patterns has been one of the main contributors in characterizing modern society and shows much usefulness in developing knowledge on understanding the human behaviors and the interplay with the environment. At the individual level, people occupy places for societal meanings and make trips for activity purposes under different transport means, the repetition of which gives rise to statistical regularities from day to day. Much evidence has also proved that human mobility is highly predictable (Song et al., 2010) and centered around a few anchoring places (Andrade et al., 2019; Herder & Siehndel, 2012). A most notable example would be the home-to-work commuting pattern, which shapes the fundamental daily mobility routine of a majority of the active population with employment. Then, in a statistical population of individuals, there are statistical similarities at the inter-individual level, which yields for group-based regularities - in other words, clusters of individuals with similar ways of trip-making on the basis of their mobility profiles. Furthermore, the impact of mobility needs on space also plays an important role in the territory evolutions (X. Liu et al., 2015; Wegener & Fürst, 2004). The spatial occupation and movements of individuals aggregate in local volumes at the zone level and origin-destination flows between zones: such aggregate patterns interplay with the local built environment and the related territorial configuration, contributing to the land-use and transport interaction. Urban areas with different functions could be reflected from relevant mobility characteristics and over time the mobility situations would progressively orient the development of new facilities and make impacts on the territorial configurations. For those reasons, crucial research interests arise with respect to studying human mobility patterns, with a good capture of up-to-date situations, problems and trends.

The comprehension of mobility phenomena has long been a key issue in scientific and social investigations. Early attention was paid to the physics of motions in the space and the dynamics over time by summarizing the characteristics (Giannotti & Pedreschi, 2008). Lately, mobility-related research has been further approached extensively in many subdivided domains with different emphasized perspectives, including urban geography, transport engineering, behavioral sociology, and spatiotemporal economics, etc. Among them, one important primary step is to capture the current situations of how people travel in the territory, from which the extensive secondary studies can be stemmed. Barbosa et al. (2018) made a review of the history and models developed for mobility understanding. Geographers were argued to be the pioneers in modeling "traveling patterns" from the 1950's, with quantitative and theoretical studies ranging from daily trips for performing social activities to inter-regional migration in long-term evolutions. While taking an overview of all domains of contributions lately, we can qualitatively assign the developed knowledge into physical models and behavioral models. As for the physical ones, a common aim of them was to reproduce the mobility patterns or population flows, including individual models such as Random Walks, Brownian Motion (Einstein, 1956), Lévy Flight (Brockmann et al., 2006), etc. and aggregated models such as Gravity Model (Zipf, 1946), Intervening Opportunities Model (Stouffer, 1940), and The Radiation Model (Simini et al., 2012). Physical distance has been regarded as a crucial factor in such theories in modeling mobility generations. While for the behavioral models, the economists have dug into the analysis about the individual decisions and preferences with respect to time and cost rather than plain physical distance. Based on that, random utility maximization theories have been developed to explain choice making in mobility and facilitating the forecasting of transportation demand (Ben-Akiva et al., 1985). As for the spatial level, the formulation of transportation with land use arrangement has been investigated through various urban economical models, such as the fundamental monocentric model (Boarnet, 1994) which presupposed the existence of a geographic space with decreasing multi-functionality as its radius from center increases. Overall, these efforts strived to provide explanations to relevant questions related to mobility phenomena, such as what initiates a trip, what determines the choice of transport and destination, and how we can predict human movements.

In the past, one of the main obstacles that hindered the mobility-related research was the availability of data. The majority of traditional mobility studies relied on travel surveys, such as census, household travel surveys, and so on. Although these surveys hold the advantage in collecting social-demographic attributes to the travelers, they have the main restriction in collecting information spanning more than a few days, due to the complex and high cost of survey conducting. This highly limits the investigation on behavioral regularities over time, while, in fact, evidence has been given on the predictability of human mobility (Song et al., 2010). The spatial coverage is also limited, due to the burden of sampling at a national wide, causing problems for understanding traveling patterns in some areas. Moreover, the low update frequency of those data is another major barrier for modern mobility studies, due to the increasingly rapid life evolutions, thanks to technology advancements. The typical 5-10 years updating interval of the travel surveys can no more respond well to the current mobility situations and provide meaningful insights for future planning. Besides, other traditional sources of data, such as loop detectors, vehicle log books, and video cameras are subject to similar problems as concerns both spatial coverage and temporal resolution. More recently, a significant increase has been witnessed in the amount of mobility-related publications, with an emerging hit of the keyword: mobility pattern (Barbosa et al., 2018; A. Wang et al., 2021). One major cause is the widespread adoption of location-based technologies, which significantly facilitate the collection of mobility information. Such a new wave expands the focus of mobility studies from a geographical standing point to an informatic exploration perspective. Therefore, new forms of mobility data are required to overcome the traditional limitations and push forward the study of mobility patterns.

1.2. New Era of Mobility Analytics

1.2.1. Modern Trajectory Data

Since the end of the 20th century, the rapid advancement of information technologies has made profound changes in peoples' lives. Sensing technologies and powerful computing devices have enabled us much new potential in observing the city and its dynamics (Aguilera & Boutueil, 2018; Peuportier et al., 2019). A new concept of urban computing has been brought to our horizon since the last decade, which aims to integrate computer science to tackle convention urban challenges in transportation, environment, civil engineering, economic, and other related fields based on a data-centric framework. Zheng et al. (2014) made an overview of the current advancements and future perspectives related to the concept of urban computing, in which he pointed out human mobility data is one of the most fundamental sources of data to study those issues. In this thesis, we make a specific emphasis on the domain of mobility. Owing to the growing diffusion of location-based technologies (GPS, GSM, Wi-Fi, etc.), digital traces are collected from more and more mobility entities including people, cars, and animals and at a large spatial-temporal scale, with in-depth trace information at high interception frequency and possibly significant sampling rates. Besides, the new era of information also brings much more data into digitalization, such as digital cartography layers, Open-Street-Map, and digitalized information of many other objects that can be fused to facilitate understanding the mobility "semantics". In all above, trajectory data are the most essential source for tracking and studying mobility movements. Parallel with the notion from traditional household travel surveys, the trajectory data collect the information of human movement traces, but empowered by digital geo-locations, the modern form now enables the possibility to track the complete paths, instead of simple origins and destinations. Extra advantages include being cost-effective, being easy-updatable, enabling multi-day monitoring, and long-term extensive analysis for mobility habits and regularities (Calabrese et al., 2013).

The modern trajectory data of human mobility exist in many forms, which can be generally categorized into two groups: active recording and passive recording. The examples of active recording include geotagged tweets, travel log applications, check-in data, etc., while the passive recording, being the most prevailing form for the analysis, includes the traces stemming from various receivers, such as cell phone (mobile phone data), vehicle on-board devices (Floating Car Data, Automatic Vehicle Location Data), transit-ticket validation machines (Smart Card Data), etc. A common definition of the data form can be provided as "a chronological series of location points at discrete time intervals associated with diverse context-varied attributes". More specifically, the data structure can be formulised as

$$\textbf{Trajectory}: \{(P_{T_1}, T_1, A_{T_1})), (P_{T_2}, T_2, A_{T_2}), \cdots, (P_{T_n}, T_n, A_{T_n})\},$$

where P_{T_n} denotes the spatial position at the time T_n , and A_{T_n} denotes the associated attributes, such as travel speed, heading, or status.

This thesis takes advantage of the availability of massive Floating Car Data (FCD) to study mobility patterns with respect to roadway vehicles: the associated traffic phenomena and the territorial aspects of such mobilities. These FCD are digital traces of positions in space and time for the "vehicle" entity, which produces trajectories in short intervals. More specifically, each trace point includes the attributes of vehicle identifiers (anonymized), geo-coordinates, moving speeds, driving headings, and timestamps. The recording frequency is around 30s to 60s, subject to the model configuration of different GPS receivers. The raw dataset is typically organized in a sequence of logs, each of which corresponds to a trace of a specific vehicle. The dataset is sourced from Coyote (https://www.moncoyote.com), which is a major roadway service information provider in France. The marketing share of the users is up to 5% of the local vehicle population, which indicates a considerably large sampling rate in terms of real-world data. Knowing such a large number of trajectories, it becomes possible to deduce sufficiently robust statistical indicators for different themes of mobility analysis. More so, as FCD are normally dedicated to collecting vehicle trace data, most of which including our case has the advantage for collecting high-quality speed status and detailed route information, making it more superior in analyzing vehicle mobility than conventional GPS records. Thus, applying the FCD could bring new valuable information to the research of vehicle mobility by its larger scale, higher accuracy, and denser information. One should also note that FCD can be regarded as a typical example of the modern trajectory data to showcase the potential and applicability in studying mobility patterns.

1.2.2. Processing and Treatment Issues

Thanks to the “ubiquity” of modern sensing technologies, digital trajectory data provide us unprecedented information to understand human mobility. While, such rich but massive data, in turn, calls for systematic research on methods for basic processing and mining tasks to obtain understandable knowledge. A huge volume of publications has been made in the last two decades for exploring the use of trajectory data, however, as a new rising thing, innovations and contributions from these studies are various but fragmentary. In 2015, a detailed survey of trajectory mining techniques was provided by Zheng (2015), which synthesized the trajectory data exploration into the following 3 common steps:

1) Data processing: This step consists of a few fundamental processing tasks to serve for the following high-level studies and applications. Typical examples of these tasks involve noise filtering for eliminating unacceptable errors, map-matching for recovering the geographical context, and trajectory segmentation for detecting meaningful trips and stay points.
2) Data management: The dataset has to be managed properly to account for the large corpus issue of massive trajectory data, which will further serve the need for efficient data retrieval in many industrial applications. It mainly concerns indexing building and data compression techniques.
3) Data mining: The mining tasks are various, serving different needs of knowledge extraction. Generally, these tasks deal with the issue of trajectory uncertainty for inferring the missing information like paths, pattern exploration for detecting correlations or trends, and category classification for identifying transport modes or activity types.

More recently, to enhance the use of trajectory data, a new concept of approaches has attracted more attention, which shifts the interest from raw trajectory processing to semantic-based trajectory processing. Parent et al. (2013) made a survey of the new ideas and relevant techniques. Overall, the new trend of trajectory processing aims to combine the raw movement tracks with more relevant contextual data to enrich the information. It particularly focuses on the interpretation of behavioral representation when conducting the mining of the trajectory data, to better serve the developments of many use-case-oriented applications.

Besides the advantages, the powerfulness of the modern data also draws the concern of many other problems, spanning from information missing or errors to data analytical difficulties (Calabrese et al., 2013; J. Wang et al., 2019). These are the challenges to be further dealt with, but the exploration of using modern trajectory data has to be further pushed ahead. Overall, we can summarize the challenges into the following aspects.

1) Erroneous and missing information: Compared to traditional travel surveys, a significant drawback of the modern trajectory data is missing socio-demographic attributes of the respondents due to privacy concerns. Such information is important in modeling underlying behavior mechanisms. To account for this, it is required to

develop semantic-oriented mining methods to investigate the historical records and make rational interpretations to enrich the trajectory representation. Besides, the data also suffer the errors caused by signal transmission, environment turbulence, etc. All these noises need to be carefully addressed before making analytics.

2) Data representativeness issue: At the current stage, the trajectory data are collected separately and independently by different information services providers or associations, e.g. the FCD used in our thesis is provided by Coyote, which is a roadway information services provider in France. This may lead to a quite low sampling rate for each source of data. The collected samples may be biased to a certain group of people like the mainstream users of the service and not able to represent the situations for the full population of all socio-demographic backgrounds. Besides, data may also be heterogeneously sampled over space causing a skewed distribution at some places and a sparse sampling in other areas. Therefore, a careful adjustment by either sampling expansion or calibration by other complementary information is necessary, especially when dealing with demand issues.
3) Computation complexity: the huge volume of trajectory data brings rich information but meanwhile it also increases the complexity and computing load for the analytics. Careful indexing has to be made for data storage and retrieval. The forms of data are also not always easy to be processed. The transition between different formats such as vectors, matrices, and graphs should be made to be suitable for different study purposes.
4) Privacy concerns: Although the data are mostly anonymized under the current trend of regulations, hackers may still be able to retrieve sensitive information hidden in the trajectory of a user. To this end, a series of technologies have been developed to protect data security for both the scenario of real-time cases and the scenario of historical records. This would rather concern the attention from the domain of information security. A more detailed review can be found in Zheng (2015).

1.2.3. Related Studies and Research Gaps

Apart from the technical processing side, it is also important to take a look at the knowledge side for the specific problems tackled using trajectory data. A wide range of studies can be found on this basis, concerning various domains, e.g., traffic control and management, city planning, land-use evaluation, commercial modeling, emergency management, epidemic modeling, social phenomena understanding, etc. (Barbosa et al., 2018; J. Wang et al., 2019). To make a qualitative synthesis, A. Wang et al. (2021) categorized these contributions into a hierarchy consisting of 3 levels which are 1) discovering the phenomenon 2) identifying and explaining the difference, and 3) prediction and implication. It should be noted that although the data structure is in common to some extent among different forms of trajectory data, the applicational orientation and use cases might bear differences subject to the tracked entities. To this end, as well as the emphasis of the thesis is to explore the utilization of Floating Car Data, we take a particular review of the specific related work based on Floating Car Data as follows.

Floating car data, emerging as an easily deployed alternative for collecting vehicle traces, has attracted much attention from researchers all around the world. We can generally summarize the related work into 4 main ranges of issues: traffic state issues, demand issues, geography issues, and behavioral issues.

First, in the literature, FCD are found mostly used to determine the traffic state. This is due to the fact that FCD originally provides good indicators of the traffic, e.g., speeds and time intervals between two places. By categories, this part can be further subdivided into three major aspects: 1) estimation of traffic speed (Fabritiis et al., 2008; Fusco et al., 2016; Rempe et al., 2017); 2) estimation of travel times (Hunter et al., 2009; Jenelius & Koutsopoulos, 2013; Rahmani et al., 2015, 2017) and 3) determination of traffic conditions, such as congestion (Brockfeld et al., 2007; X. Liu & Ban, 2013), bottleneck (Altintasi et al., 2017) and incident (Houbraken et al., 2017). Although FCD do not provide direct information on density for flow estimation, some studies have also been conducted to model the traffic flow by either deriving a fundamental diagram between speed, density, and flow (Sunderrajan et al., 2016) or building functions based on the conservation law (Seo & Kusakabe, 2015). Further than those above, by knowing those traffic state parameters, many other studies tried to manifest the added value of FCD into extended traffic analysis, involving modeling of traffic emission (Russo et al., 2021), estimation of parking searching time (Mannini et al., 2017), regulation of traffic lights (Astarita et al., 2017), and evaluation of safety effects (Kieć et al., 2018).

Second, many studies use FCD to conduct demand estimation or forecasting, benefitting the need for transportation planning. One essential metric for the demand analysis in transportation to is derive the Origin-Destination (OD) matrices (Dewulf et al., 2015; Nigro et al., 2018). As FCD store the path information, it is not complicated to reconstruct the OD matrices between spatial zones. The OD matrices can also be reconstructed by different time periods so as to model the demand situations between peak and off-peak times. However, an important problem that comes with this issue is the heterogeneity of sampling rates among different OD pairs, namely sampling biases problems or unrepresentativeness issues, which were overlooked in many above studies. Yang et al. (2017) and Gómez et al. (2015) have made efforts on such an issue by deriving a prior-matrix from FCD and using supplementary information such as local link flows as constraints to correct the biases. However, these two studies were all based on simulated data, the applicability on real-word data is still pending to be examined.

Third, FCD is also found used in geography studies. One common use case is to detect hotspots based on trace clustering, such as detecting popular areas of pick-ups and drop-offs (Jahnke et al., 2017; X. Liu et al., 2015; Tang et al., 2015) and discovering points of interest (Angkhawey & Muangsin, 2018; Y. Liu et al., 2021; Qi et al., 2011). In fact, the mobility situations are largely correlated with the environment, and analyzing human mobility activities could help understand many geographical issues. Based on that, a few researchers further explored spatial structure behind the observed mobility phenomena, including exploring the land uses (Y. Liu et al., 2012; N. J. Yuan et al., 2014), analyzing accessibility issues(Q. Li et al., 2011), and delineate the borders of activity basins (X. Liu et al., 2015; Rinzivillo et al., 2012).

Lastly, we also found a branch of research using FCD to model behaviors, but largely related to travel behaviors. Route choice modeling is the most prevailing subcase in this direction. The path observed in FCD could provide a good reference for analyzing the various choice sets (Bekhor et al., 2006; Ciscal-Terry et al., 2016) and modeling the rationale behind the decision making (Dabbas et al., 2021; D. Li et al., 2016). The large number of observations also facilitates the modeling potentials in improving the traditional utility theories by including more individual characteristics or also enabling training machine learning methods in predicting the choices (R. Yao & Bekhor, 2020). Besides, some researchers also investigated the destination choice problems by looking at the historical records (Xue et al., 2013) and making recommendations for the fastest routes (J. Yuan et al., 2010).

To summarize, extensive methods have been developed in processing trajectory data, but previous mining methods mostly concentrate on using direct information collected in the data. Further work should be expanded towards a semantic-oriented mining by deepening the exploration of behavioral representation and interpreting activity context to enrich the knowledge. As for the applications, unlike human-based trajectory data, FCD are mostly used for determining traffic status or modeling traffic relevant parameters from an engineering standing point. Although some studies have shed light on revealing behavioral patterns of route choices and geographical patterns of areas of interest based on FCD, the efforts are still not sufficient and subject to further exploration to have a more complete understanding of vehicle-based human mobility. Relevant issues may concern the vehicle usage in individual mobility and spatial structure in shaping collective mobility activities or vice-versa. There are also many specific technical gaps according to each particular research problem to be addressed, we will take a closer look of those in the following chapters when addressing the detailed questions. In sum, the great potential of FCD in studying mobility patterns has not been fully excavated.

1.3. Research Issues and Objectives

Acknowledging the context and research gaps, this thesis aims to explore mobility patterns (in the sense of motifs, forms) by leveraging the modern trajectory data (FCD) to contribute a better understanding of vehicle-based human movements in the frame of territory. Since the spatial configuration and human movement behaviors are the two key elements in sequencing mobility phenomena, which are reciprocally interacted, mobility patterns are studied at two levels in this thesis: the elementary level of understanding behaviors of the “author” of the trajectories, and the more global level of capturing the geographical aggregated mobility phenomena.

More specifically, at the individual level, the objective is to mine out patterns of individual movements to characterize statistical regularities in mobility behaviors. As the behavioral characteristics could be studied from plenty of aspects, we delineate the study scope with a concentration on a few fundamental aspects. Specific research questions are addressed with respect to vehicle usage types pertaining to daily trip-making, traveling regularities discovered along a longitudinal time period, and mobility “anchoring” places where the mobility activities are centered around. We particularly pay attention to mine the behavioral representations conveyed in the Floating Car Data and take the advantage of multiple-day tracking to investigate the mobility dynamics over time. Travel time is also modeled in this thesis with a novel approach as it plays as one of the most important factors in individual decision-making. Besides, FCD are especially suitable for addressing such a problem, however, the applicability of using real-world data remains to be further investigated as well as with a few analytical gaps to be filled.

Regarding the spatial level, the goal is to characterize the structure of the territory according to the vehicle mobility that takes place there (more precisely, which is sampled there). Specific research questions include revealing the functional occupation of a space from human activities, identifying spatial relations between places to find out core areas and bonded communities, and quantifying the spatial interaction intensity by estimating the traffic flow between places. These questions aim to explore different facets of the territorial mobility morphology, joining together to provide a systematic understanding of the issue.

Besides obtaining the knowledge of mobility patterns at the two levels, another fold of the thesis objective is to build up methodological approaches for trajectory mining, contributing to broadening and deepening the way of using trajectory data in mobility analytics. This concerns the algorithms and methods to be developed for processing the raw data, reconstructing the information, and developing analytical models, aiming to translate the data into understandable mobility knowledge.

Overall, the outreach of the thesis is to expand the mobility analytics of patterns from a data observing standing point, by employing new forms of trajectory data and the state of arts of artificial intelligence to overcome traditional limitations and explore new potentials. As the specific focus is on mining Floating Car Data, due to data availability at the stage of doing this thesis, the research problems investigated are oriented to roadway traffic-based mobility patterns. This situates the empirical finding of this thesis in a different position from person-based mobility study, which is another mainstream of mining Mobile phone GPS traces. However, it should be noted that the proposed ways of pattern constructing, the developed methods and algorithms, and the applications frameworks are transferrable on studying the trajectories from other sources of mobility entities in a much referential way.

1.4. Research Approaches

The analytical framework is founded based on how to use trajectory data to study mobility patterns, an overall illustration of which is displayed in Figure 1.1. Consistent with the two folds of the objectives, research approaches are developed for both the methodological knowledge in modeling mobility patterns and the instrumental approaches in mining trajectory data. More specifically, 4 key methodologies considerations are described as follows.

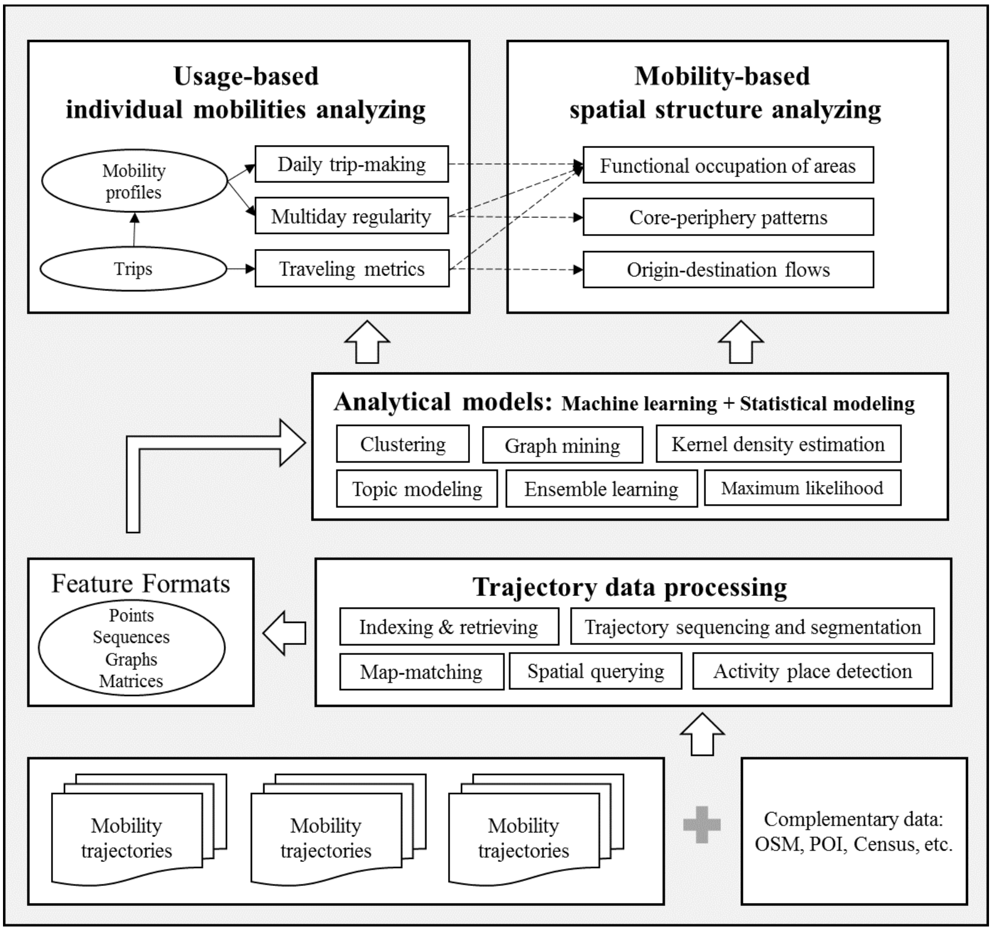


Figure 1.1 Illustration of the analytical framework

Trajectory data processing: An analytical platform is set up to process the vast amount of FCD within the territory of the Great Paris Region spanning multiple days. The platform is built from two aspects: technical architecture and functional architecture. Technically, the processing framework is mainly developed based on Python scripting with aid of GIS tools for spatial coding and visualization. A data-lake composed of trajectory data and other relevant contextual data is built up in the backend for data storage of and rapid information retrievals, supporting the dataflow connection among various processing operations. The data-lake is managed by MongoDB, a document-based database program, which supports the storage of data in different types and nested structures, e.g., geo-coordinates vs timestamps. Functionally, algorithms are developed to enable a series of mining tasks, including trajectory sequencing and segmentation, map-matching, activity place identification, individual usage profile constructing, and spatial attribute assembling. Besides those well-established basic algorithms such as spatial-joining (*Spatial Join - GIS Wiki*, n.d.), map-matching (Newson & Krumm, 2009), etc., the core processing algorithms and pipelines are either originally designed or adjusted with customization for the fitness in processing the FCD of our case. More details about the principles of specific algorithms are explained in the following chapters of studies, according to the needs of addressing corresponding research problems. It should also be mentioned that the large corpus of the massive data would hinder the analytical explorations due to the computational complexity with the high time cost of each trial. For example, the FCD dataset of a whole day in our case for the Paris Region would yield a mass of 1.4 million records for around 62,000 vehicles and costs a few hours in major steps of processing such as trajectory processing on a single PC with a conventional configuration. Bearing the computation efficiency in mind, a pre-sampling in downsizing the dataset is often used for experimenting with the developed methodologies and making relevant tuning of the algorithms. According to the objectives of different sub-issues, different data samples are used in the case studies. Detailed settings are provided in the corresponding chapters.

Applicability of machine learning techniques: Thanks to the ease of information acquisition, modern trajectory data like FCD can provide us a large amount of observations to capture mobility situations and their evolution, based on which artificial intelligence can be leveraged to facilitate mobility analytical problems. The applicability of machine algorithms is exploited in this thesis, to investigate the feasibility of overcoming traditional limitations of mobility analytics and also explore the potentials brought by automated learning. The input requirements, the selection of algorithms, and the parameter tuning are discovered and tailored to tackle the proposed mobility issues. In particular, unsupervised learning techniques are mainly explored and applied due to the fact that common forms of trajectory are anonymized by eliminating most of the labels that can reflect the direct semantic meanings of personal mobility. For privacy concerns, any information relevant to the socio-demographic background and the personal attributes of the individual is not contained. To this end, the analytical process should not be tackled against a specific individual, however, the results can be focused on grouped phenomena. Besides the original format, trajectory data can be further transformed into other formats, such as points, profiles, graphs, matrices, and so on, according to which, specific machine learning algorithms, including clustering[1], kernel density modeling[2], topic modeling[3], graph mining[4], and ensemble learning[5] are applied in this thesis.

Usage-based understanding of individual mobilities: Methodologies are investigated on how to model individual mobility patterns, in particular, related to behaviors, preferences, and decision makings. Instead of simply using the mined-out features and feeding them into machine learning algorithms and statistical models, we make an original contribution in building a hierarchy to model the individual vehicle usage for solving the research questions progressively. In fact, mobility phenomena can be deconstructed by an organization of units at different levels. At a lower level, mobility patterns are presented as movements in a mix of spatial and temporal features, where we can introduce the trips as units. The trips here are defined as traveling movements for achieving certain activity purposes. To this end, a set of trip characterizing features are further modeled, which includes during-trip features (such as times windows, spatial distances, traveling speed, and path categories), before-trip features (such as trip purposes and traffic conditions for decision makings,) post-trip features (such as the consequences of the trips, staying places, and activity interpretations) and other possible spatial-temporal related parameters. On top of that, we model the vehicle usage of an individual according to the aggregation of all its relevant trips, which can be regarded as a usage profile that tracks its signature ways of trip-making. These profiles can be assembled at different scopes to understand the individual mobility pattern from different perspectives, such as the activity routine of a weekday/weekend or the visiting frequentation to a certain place. Lastly,

[1]·Clustering: a task of grouping a set of observations into groups according to similar characteristics or patterns

[2]·Kernel density modeling: a non-parametric way of estimating the probability density distribution of a random variable

[3]·Topic modeling: a probabilistic model for discovering recurring patterns as the abstract “topics”, commonly used in documentation processing.

[4]·Graph mining: techniques for exploiting graph data for properties, relationships, and structures of the elements (nodes and links)

[5]·Ensembling learning: techniques that aggregates multiple learning algorithms or the results by different data sources to obtain better learning performance.

machine learning and statistic modeling are conducted based on these profiles to explore vehicle usage ways, regularities, and preferences.

Mobility-based understanding of spatial structure: Methodologies are as well investigated on how to study mobility patterns in a spatial territory. We know that the configuration of human activities in space generates mobility, in connection with the intensity of land use by the functions of housing and production i.e. jobs. Many works have proved that investigating people's mobility activities can help to understand urban or territory morphology (Y. Liu et al., 2012; Rinzivillo et al., 2012; N. J. Yuan et al., 2014), thus inspiring us for further exploration along this direction. From a trajectory basis, we approach the issue of understanding spatial structures from 3 aspects, combined together towards a holistic view of the mobility phenomena of a territory. The first is to interpret the functional occupations of a place or an area by looking at the human activities carried out there. Places holding different functions, such as commercial zones and residential areas, tend to show up different patterns, in terms of the time of flows, spatial accessibility, and the staying characteristics of activities by duration and frequency of occurrence. By mining and combing those different facets of mobility characteristics using trajectory data, a typology of different geographical spaces can be characterized to differentiate their functions. Secondly, we characterize the structure of the territory by seeking spatial relations and identifying a hierarchy of places according to the movement currents. A common concept of such an issue is to explore the core-periphery patterns. To this end, the spatial relation is approached by firstly identifying core activities areas and then determining the catchment areas to each core from the mobility exchange between spaces. Lastly, to quantify the spatial interactions, we utilize the trajectory data to estimate the travel demand in flow between different origins and destinations. More than that, we pay particular attention to account for a representativeness bias issue in travel demand estimation when using FCD, which is commonly existed in the sampling process in most trajectory data. As it is an inherent problem that originated from the data collection, we seek other sources of information such as link flow counts at a few spots by roadside cameras for the calibration. Overall, the key approach for this research problem is to take advantage of observed path use from trajectory data to obtain trip assignment among places and estimate the origin-destination flow matrix directly and efficiently. This actually contributes to capturing the dynamic evolution of travel demand and its distribution over a territory by using the easy-updatable numerical data.

1.5. Thesis Outline

The overall research issue has been addressed in 6 articles prepared for referred scientific journals or conferences, which consists of 3 sub-research issues for individual-centered analyses and 3 sub-research issues for place-based analyses. This thesis is organized in the format of an article-based manuscript, with each chapter corresponding to an article in addressing a specific research problem. Below, a short overview of the main issue addressed in each chapter is provided.

Chapter 2 presents a study on discovering vehicle usage patterns of individuals based on their digital FCD footprints. It expands the semantic exploration of individual mobility patterns by capturing how vehicles are used in individual daily mobility and making a characterization of the usage ways. To achieve that, a trajectory sequencing model is firstly developed to segment the trajectory into meaningful legs, namely, trips. Trip physical types are then identified by conducting a clustering analysis based on departure time, trip distance and driving speed. Upon that, a mobility profile can be built for each vehicle by aggregating pertaining trips into a vector of counts per type in terms of the identified trip physical types and the geographical locations of destinations. Lastly, a topic-modeling approach is developed based on Latent Dirichlet Allocation to discover vehicle usage patterns by regarding profiles as documents, trips as words, and usage types as latent topics to be determined, thereby constituting a vehicle usage typology. An application was conducted for the Paris Region and identified five major vehicle usage types, among which three types were associated with local usage within specific areas and the other two had hybrid patterns between different areas. The prevailing pattern of vehicle usage was found on short-medium trips around peri-center and near suburban areas. Overall, this study showcases a data-driven framework to help understand vehicle daily usage patterns and their differentiation over a territory. In addition, the proposed topic-modeling-based method deals efficiently with the data sparsity and high dimension problems arising in the exploration of mobility usage patterns, along with the consideration of both physical characteristics and geographical context and the potential scalability to include further information.

Chapter 3 aims to investigate significant places to mobility makers, by identifying the "anchoring" geolocations and further extrapolating their functional types. In this study, we take the advantage of multiple-day tracking of FCD to explore the mobility dynamics over time. It demonstrates a way to investigate the individual mobility regularities and habits by historical trajectories. In particular, the geolocations of significant places are identified by detecting the stay points from trajectories of an individual and making a density-based clustering to cluster them into represented places. The functional types of the locations are discovered based on a two-level hierarchy method, which is to primarily identify the activity types of each visit according to stay characteristics, and secondarily discover the place types by assessing their profiles of activity composition and frequentation by investigating the occurrence likelihoods of each kind of activities. An applicational study was conducted in the Paris Region. Consequently, seven types of significant places were derived, the prominent characteristic of which can be further categorized into home place, work place, and other types of secondary places. Most of the vehicles analyzed were detected with home places while around half of them were found with work places based on frequent vehicle usage. The results of the proposed method were also compared with those from the commonly used rule-based extraction method and showed a highly consistent matching. Comparing to previous efforts, the proposed methodology shows good applicability in identifying significant places without prior knowledge in terms of both pre-labeled data for referential training and the expertise on setting specific rules for place recognition. Moreover, it also enables the detection for more diverse situations, as well as identifying the places for not only the primary ones (home and work places) but also the other secondary ones.

Chapter 4 proposes a novel approach for travel time estimation by building a stochastic model to exploit FCD materials. It aims to allow for simple but robust estimation of key factors in traffic conditions, along with the goal to fill the research gap in measuring the reliability along with the travel time. This issue is studied particularly as it plays a vital role in individual decision making and also the massive FCD traces serve like sensors over the network, making it especially suitable and powerful in addressing such travel time estimation. In this study, probabilistic specifications of Gaussian random variables are postulated to model the link travel time. A Maximum Likelihood Estimation method is devised to estimate the stochastic parameters based on massive FCD observations which contain the travel time information between spatial intervals. A numerical experiment was conducted to demonstrate the method's applicability under both urban settings and highway settings. The stochastic method was also compared with a conventional method, which straightforwardly took the average of all observed point-wise speeds from FCD. Results indicated that the stochastic model was able to deliver a more reliable estimation and required fewer observations to reach higher precision. Moreover, the pointwise average estimation was found tending to provide a higher speed than the stochastic model with less certainty. This may lead to an underestimation of the travel time, implicating the limitation of the current applications adopting such a straightforward estimation. The proposed stochastic model achieved a good computation efficiency, showing the applicability of being a modular work serving for estimation on large-scale network or adding up to obtain the path travel time.

From **Chapter 5**, this thesis starts to explore the aggregated mobility patterns in geographical space. This chapter presents a study to investigate spatial functional occupations by looking at related vehicle movements. A mobility-related typology of territorial zones is built to categorize and differentiate the space roles and uses by human activities. To be specific, the spatial functions of a territory are discovered by zonal divisions. For each zone, mobility-related attributes are mined out from 3 different views to describe the pertaining vehicle usage in terms of the composition of activity stays by frequentation, temporal flows of trip generation and attraction, and spatial complementarities in trip distance distribution. Then, a multi-view cooperative clustering analysis is conducted to identify the zone typology by integrating the explorations from the above different views of attributes. Such a method is particularly employed due to the fact of multiple facets of characteristics that a place could show up while interacting with mobility activities. More hidden patterns are also expected to be spotted, which are, otherwise, likely to be diluted if all features are concatenated into a single view. An applicational study was dealt with in the Paris Region using census-based zonal divisions. As a result, five mobility types of zones were characterized, including residential-oriented areas, business-oriented areas, and commercial-amenity mixed areas, etc., with each holding a different orientation of mobility usage. An evaluation was done by comparing the discovered areas with the common recognition of their social functions, which showed a consistent matching. Overall, this study provides a big-data instance to study and territorial functional divisions from mobility trajectories. The result could help benefit future planning and many other place-based applications.

Chapter 6 further extends the spatial exploration to study the relational patterns between spaces. It presents a study of mining trajectory data to recognize core activity areas of a region and identify their catchment areas. It aims to deconstruct the mobility structure of a territory by analyzing the core-periphery patterns and revealing agglomerate spatial communities. More specifically, this study pays particular attention to jobs-housing spatial relations considering the fundamental role of home-work commuting in one’s mobility life. The home and work places are first identified using the method developed in Chapter 2. A spatial density distribution analysis is conducted to identify the employment cores and sub-cores area. A jobs-housing graph network is then built based on the home and work flow counts to investigate the connection between core areas and other spatial zones. The catchment areas are identified from the jobs-housing network by applying the graph-partition algorithms to find communities (sub-networks) with a denser exchange internally and a sparser communication externally. The proposed methodology was applied in the Paris Region to evaluate its applicability. 10 employment core zones were identified according to the spatial density, which consisted of those commonly acknowledged business centers such as La-Defense, Boulogne-Billancourt, Versailles, Roissy, Rungis-Orly, etc. The community results were obtained under 2 different settings: the original jobs-housing network and the calibrated one, which is adjusted by making a zone-specific sample expansion according to local automobilist population from Census data to compare for potential sampling biases. The 2 results were consistent on the overall spatial distribution of the detected communities, although the specific constitution of some areas was different. The differences before and after the calibration can reflect that certain commuter groups were underrepresented in the sample data, which implicates further consideration in future data collection. However, both of the two results showed a good spatial adjacency of zones when forming communities as well as with employment cores embedded inside, although this is not a necessary property of the graph partition algorithms. Such a finding can be proof of the effectiveness of the proposed method for its good capability in discovering densely connected sub-regions while maintaining spatial cohesion. Overall, this study offers a pipeline for studying geographical relations directly from the traces without relying on external information or prior expertise. The analysis can be easily replicated with updated data inputs, showing a practical potential of capturing up-to-date territory evolutions, benefiting quick responses to the mobility changes.

Chapter 7 presents a study to explore the spatial interaction of mobility flow between places. It aims to leverage the modern data to quantify the spatial interactions in traffic flows, while accounting for the heterogeneous sampling rate issues of such data, which is recognized as a common problem remaining to be further investigated in the literature. This study deals with the estimation of the Origin-Destination matrix flows based on two kinds of data: vehicle trajectory data and local traffic counts. A step-to-step Bayesian formulation is derived for demonstrating the relationship between the link probe sampling rates and the fractional contributions from the sampling rates on different OD pairs. The unknown OD matrix is estimated by applying cross-entropy minimization using a prior matrix from the probe trajectories, along with the Bayesian assignment rules on link sample rates as the constraints. The methodology was applied using Floating Car Data and camera link flow counts for a numerical experiment. The results show that the method can achieve a robust estimation of OD matrices, even using different prior matrices. The issue of the heterogeneous sampling rates can be well addressed with link count constraints, effectively correcting the unknown bias in the probe sampling. It was also found that using an informative prior matrix using link counts to calculate OD pair specific sample rates would contribute to a more reliable estimation. The case study using real data also proves the feasibility of mining observed trajectory data to obtain the assignment fractions and estimate the OD matrix inversely, avoiding the conventional sophisticated process of traffic assignment modeling.

Chapter 8 provides an overview of the research presented herein, summarizes the major conclusions, contributions, and discusses the limitations along with the recommendations for future work.

CHAPTER 2. DISCOVERING VEHICLE USAGE TYPOLOGY BASED ON DAILY FOOTPRINTS

Adaptation of the paper published in:

Sun, D., Leurent, F., & Xie, X. (2021). Discovering vehicle usage patterns on the basis of daily mobility profiles derived from floating car data. *Transportation Letters*, 13(3), 163-171.

The candidate contributed to designing the method, performing the study and writing the manuscript.

Abstract

Digital traces of mobility entities such as vehicles and pedestrians are increasingly available nowadays, bringing great potential for mobility analysis. This paper presents a novel approach for establishing vehicle usage patterns by Floating Car Data based on their daily mobility making. Firstly, mobility representative features were recovered from trajectories via trip segmentation and characterization in terms of time window, travel distance, and average speed, as well as the geographical sector of the trip destination. Trips pertaining to each vehicle were then aggregated as a vector of counts per type in order to obtain the mobility profile of the vehicle. Based on these profiles, a topic modelling approach using Latent Dirichlet Allocation was developed to discover the patterns of vehicle daily usage, thereby constituting a typology. An application was conducted for the Paris Region and identified five major vehicle usage types, among which three types were associated with local usage within specific areas and the other two had hybrid patterns between different areas. The prevailing pattern of vehicle usage was found on short-medium trips around pericentre and near suburban areas. Overall, this study offered a data-driven framework to help understand vehicle daily usage patterns and their differentiation over a territory.

Keywords: Mobility pattern; Vehicle daily usage; Trajectory segmentation; Floating Car Data; Machine learning

2.1. Introduction

Over the years, more and more territories have been facing mobility challenges. One of the key causes is the inconsistency between mobility behaviours of people and the spatial structure of the region (X. Liu et al., 2015), which, therefore, prompts a need to understand people's travelling characteristics to reveal the underlying issues. A type of such travelling characteristics is referred to as a mobility pattern, which can be used to either characterize an individual phenomenon of movements or depict a spatial function interacted with human movements.

In the past studies, the availability of data has been a major concern for mobility analysis. Traditional data collection methods, such as household surveys, loop detectors, vehicle diaries, and video cameras are inherently limited as concerns both spatial coverage and temporal resolution (D. Sun et al., 2014). With the growing diffusion of GPS devices, Floating Car Data (FCD) have emerged as a new data source to address mobility analysis in a systematic and cost-effective way. Since no special equipment is required to set up, this technology can be widely deployed and has the potential in providing data for large scale network and up-to-date mobility demand.

FCD-related studies in existing literature can be generally grouped into two categories of research subjects: “physical issues” for recovering the state of road network performance and “behavioural issues” for revealing the microeconomic mechanism behind trips. More specifically, the first part focusing on traffic state analysis can be subdivided into three major issues: 1) estimation of traffic speed (Fabritiis et al., 2008; Fusco et al., 2016; Rempe et al., 2017); 2) estimation of travel times (Hunter et al., 2009; Jenelius & Koutsopoulos, 2013; Rahmani et al., 2017); 3) determination of traffic conditions (Altintasi et al., 2017; Brockfeld et al., 2007). As for behavioural analysis, several studies have been targeted to route choices either to model it or to perform statistical analysis (Ciscal-Terry et al., 2016; D. Sun et al., 2014; Zhu & Levinson, 2015). These studies addressed many traffic issues, however, so far, limited work has made use of FCD to investigate mobility issues in terms of vehicle usage patterns. According to the “mobility pattern” introduced beforehand, vehicle usage patterns can hereby be defined as types or repeated ways in which the vehicles are used to conduct mobility activities among the population. In fact, FCD records enable for spatial-temporal trajectories, which can be analysed to search for typical vehicle usages. So far, the issue of vehicle usage has been commonly investigated by using conventional vehicle diaries, especially for freight vehicles, results of which could further contribute to multiple sector analyses such as business monitoring, activity configuration and many other economic analyses (*Survey of the Use of Road Freight Vehicles (TRM)*, 2018). However, such analysis was limited due to the time-intensive nature and in-complete path capturing of conventional vehicle diaries. Therefore, new kinds of digital data such as FCD are expected to eliminate the restrictions and improve such studies (Nguyen et al., 2017). Overall, in the literature, there was a significant knowledge gap in studying vehicle usage types from FCD. Only limited reference was found to classify vehicles into types using GPS records collected in a more specialized way than anonymized FCD (Simoncini et al., 2016, 2018; Z. Sun & Ban, 2013).

Sun and Ban (2013) developed a method using Support Vector Machine (SVM) to distinguish delivery trucks from passenger cars, using field collected GPS data from traffic mobile sensors with a frequency of every 1s and 3s for passenger cars and trucks respectively. The result showed that acceleration- and deceleration-based features were more salient than speed-based features to perform the classification. However, this study only classified the vehicles into two classes and only limited data on arterial streets were involved. Besides, the data used in this study were recorded with a rather high frequency, while a more common practice is to transfer less data with less frequency, e.g. every minute or even longer. In Simoncini et al. (2016), the authors claimed that they were the first effort in tackling the problem of vehicle classification using low-frequency GPS data, collected from the installed devices in commercial fleets. A binary SVM classifier was developed to distinguish light-duty vehicles from larger ones (vehicle-mass-wise), based on a combination of features that were identified to be most predictive using a recursive feature elimination procedure. A more recent study by Simoncini et al. (2018) was conducted to classify vehicles from a lower frequency GPS data collected from connected vehicles by every minute. The authors employed recurrent neural networks to categorize the vehicles into the types of small-duty, medium-duty, and heavy-duty vehicles. An approach based on Long Short-Term Memory (LSTM) recurrent neural networks was proposed to learn effective hierarchical and stateful representations for temporal sequences, which outperforms the existing state-of-the-art. However, the employed data mining approaches, SVM and LSTM, were all based on supervised learning, which requires the vehicle type labels. While, due to the increasing need of privacy and regulation of data protection, the trend of the data available to wide research use would be fully anonymous, which prompts the need of unsupervised data-driven techniques in exploring such kind of data. Another issue among these studies is that the focal point was on vehicle model type rather than the usage type of mobility making, which is a more direct reflection of mobility activities.

Therefore, the objective of this study was to explore vehicle usage patterns based on FCD, from an activity-oriented perspective. The time unit was set as one day, considering the natural time frame and cyclic pattern of people activities. More specifically, the research aim was twofold: first, to reveal the usage patterns of vehicle on one day, second, to build a data-driven analysis framework employing unsupervised learning. An application was conducted over the Paris Region. This paper is a further development of our previous study (D. Sun et al., 2020), which focused on trip-making patterns only and did not address geographical patterns. In the present paper, we address both trip-making and geographical features, and we propose topic modelling to search for usage patterns: such an approach is more powerful than conventional clustering algorithms.

The structure of the paper is as follows. The data structure is presented in section 2. Section 3 describes the methodology of trajectory processing and vehicle usage patterns discovering. The application setting and corresponding results are presented in section 4, followed by a concluding discussion in Section 5.

2.2. Methodology

2.2.1. Trajectory Sequencing into Trips

2.3.1.1. Problem Definition

To recover the mobility information from FCD, it is important to sequence the trajectories for each vehicle and segment them into meaningful trips. This is because the prevailing FCD for the general public users does not have the dedicated indicator in the data to indicate whether a trip terminates or not. A trip is pre-defined as a set of consecutive traces by which the vehicle makes for a certain purpose of movement. A temporal stop is made between two trips which indicates the time duration for the activity. Many efforts in the literature have been made on trajectory segmentation into trips based on GPS traces, however, most of the methods and thresholds were investigated for the personal trips based on mobile phone GPS (Gong et al., 2014, 2015). Only limited work was found on the trajectory segmentation on the vehicle basis (H. Chen et al., 2017; Lin et al., 2016; Sarti et al., 2017). Among those works, the time interval between two succeeding points was most widely adopted as the decisive feature for identifying trip ends. Other criteria such as the distance and the interval average speed were also included in some studies. However, thresholds of those criteria varied from one study to another subjecting to the data source and the local context, with a time interval ranging from 90s to 600s and a distance from 50m to 150m.

2.3.1.2. Trip Detection Method

It can be generally assumed that the time interval between two trips is longer than the that of device recording. Many confounding scenarios may still exist, such as signal loss and instable signal transmission etc. Although there is no perfect way to detect the trips explicitly, a sole arbitrary interval threshold would result in much error in the trajectory segmentation. Thus, a statistical trip detection method with multiple checking conditions was developed in this study to interpret trip ends, the process of which is depicted in Figure 2.2. The major criteria are explained as follows. The time interval t between two consecutive records was firstly compared with a time threshold to distinguish normal recording pauses from overlong intervals. The time threshold was drawn at the value where there was a significant distinction by inspecting the statistical distribution of all intervals. As the cases with overlong intervals may contain some signal loss scenarios which should not represent the termination of trips, the underground road locations were employed from the OpenStreetMap to exclude those falling-in stopping points from the trip detection. The last criterion was used to account for the cases by unstable signal transmission and other asynchronous triggers such as engine-off events, which are still quite many considering trips are relatively rare to the recording magnitude. A gap distance ratio i was proposed for such a determination as the formula given below:

$$i = \frac{observed\ distance}{expected\ distance} = \frac{distance\ (P_a\ to\ P_b)}{\frac{1}{2}(V_a + V_b) * t} \quad (2\text{-}1)$$

where $distance\ (P_a\ to\ P_b)$ indicates the planar distance between point a and point b of an interval, V_a and V_b indicate the recorded speed of the two points respectively, and t represents the time duration of the interval. The overall assumption for this criterion is that for each interval, the vehicle with a trip end should have a much less observed distance than expected compared to those without making a trip end. It is assumed that *i* should be close to 1 for on-journey scenarios and, otherwise, 0 for those most likely having a major trip end for an activity. However, confounding cases may still exist with a ratio value in between caused by harsh driving events (Sarti et al., 2017). A Kernel Density Estimator (KDE) was therefore used to model the density distribution and infer the ratio ranges for different scenarios. It should be noted that, although not all trips may be detected, this model aims to capture those ones with major trip end activities without involving too many "fake" ones.

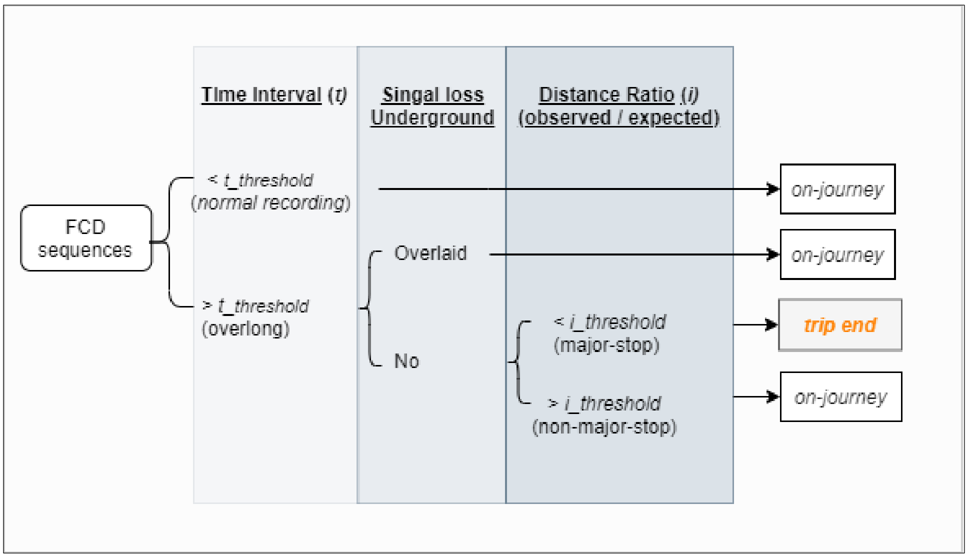

Time Interval (t)
Singal loss Underground
Distance Ratio (i) (observed / expected)
FCD sequences
< t_threshold (normal recording)
> t_threshold (overlong)
Overlaid
No
< i_threshold (major-stop)
> i_threshold (non-major-stop)
on-journey
on-journey
trip end
on-journey

Figure 2.1 Process of the trip detection method

2.2.2. Discovering Vehicle Usage Patterns

2.3.2.1. Problem Definition

Considering trips are normally used as the units for analysing travel behaviours, a trip profile can be built for each vehicle to represent its mobility pattern and analyse the usage type. Respective definitions are given hereafter:

- **Trip pattern:** A trip pattern (*w*) is a certain type of trips with certain characteristics.

- **Vehicle trip / mobility profile:** A vehicle trip profile (*d*) is a set of counts per trip pattern (*w*), depicting the trips made by the vehicle on a given day. ($d = \{n_{w_1}, \dots n_{w_N}\}$). The vehicle trip profile complemented by the counts of the trip destination points according to the geographical sectors is called the "vehicle mobility profile".

- **Vehicle usage typology:** The Vehicle usage typology (*T*) is the set of types or patterns ($t \in T$) of daily mobility, each of which exhibits recurrence among the population of vehicles. Each type (*t*) corresponds to a group of similar usage of vehicle in mobility making.

2.3.2.2. Method: A Two-Step Identification

A two-step identification method was proposed for discovering the vehicle usage typology: *Step 1* was to identify the types of elementary trips; and *Step 2* was to identify vehicle usage types on the basis of vehicles' mobility profiles

Step 1: trip type identification by K-means clustering. The K-means clustering algorithm was employed to partition trips into k different homogeneous groups, depending on their trip features during the travelling. To begin with, feature selection and standardization are required. To describe trips, we considered the following independent features in the recovered trip dataset: the departure time, trip distance, and average driving speed. Of course, many other attributes of trips may also be considered if related information can be fused from other data sources. Since the range of values of the above features may vary widely due to their respective scales, all features were standardized to weigh each dimension equally.

Another key issue of clustering analysis is to determine the number of clusters. It is commonly acknowledged that there is no unambiguous answer to this question. The optimal number k of clusters is relatively subjective and depends on the methods and the data used for partitioning. In this study, two widely used methods, the elbow method (Thorndike, 1953) and average silhouette method (Rousseeuw, 1987) were employed to determine the optimum cluster number for the k-means clustering. By combining the results of good candidates from the two analyses, a final judgement could be made to determine the optimal number of k.

To further understand the characteristics of the clusters and check whether the partitioning result was logical, a further characterization analysis was conducted to explore significant variations between clusters. By finding out prominent characteristic differences among each cluster, such a process could help define the types (Ren et al., 2018). Different approaches can be employed to make the comparison between clusters. Considering the feature dimensionality, partitioned trip clusters were characterized by a 3D scatter comparison.

Step 2: vehicle usage type identification by Latent Dirichlet Allocation. The above-defined vehicle trip profiles were built by counting the trip frequency that falls in each trip type. Besides the trip-making profiles according to the k clusters, the trip destination sector i.e. a geographical feature, can also be incorporated and identified by overlaying with the geographical morphology divisions of a territory (an example of the Paris Region can be found in *Section 2.3.1*). As a consequence, a cross-tabulation was assembled as the mobility profile for each vehicle, as shown in Table 2.1, with rows i and columns j as the geo-sectors and trip travelling patterns respectively. Each cell in the table represents the trip frequency of a certain trip pattern *wij*.

Table 2.1 Mobility profile built for each vehicle

Geo-Sectors (w_i)	**Trip-making profile by traveling patterns (w_j)**			
	Trip type 1	Trip type 2	…	Trip type *j*
Sector 1	n_{w11}	n_{w12}	…	n_{w1j}
Sector 2	n_{w21}	n_{w22}	…	n_{w2j}
…	…	…	…	…
Sector *i*	n_{wi1}	n_{wi2}	…	n_{wij}

We may regard the vehicle trip profile as a travel document consisting of various trip patterns as key words. Then, the topic modelling method can be used to explore the patterns of vehicle trip profiles just as the way of finding document topics based on text words. In this study, the widely used Latent Dirichlet Allocation (LDA) algorithm was employed to determine the vehicle usage typology. LDA is a generative probabilistic model (Blei et al., 2003), whose goal is to find a set of recurring patterns as the hidden topics for the document collection. It assumes that the words of each document could arise from those hidden topics, while each topic presents a set of keywords with corresponding importance (by distribution). Besides, it also holds the advantage in dealing with the sparse matrix, which fits the processing need with geographical features involved (N. J. Yuan et al., 2014; Zhao et al., 2019).

The document in LDA is treated as a bag of words, which can be managed as a set of trip patterns $\{w_{ij1}, \dots w_{ijN}\}$ in our case. The Bayesian model can then be described as Equation 2, with vehicle mobility profiles treated as documents d, trip pattern treated as different words w, and vehicle usage types treated as the topics t to be detected.

$$p(w \mid d) = \sum_{t \in T} p(w|t) * p(t|d) \quad \text{(2-2)}$$

where, $p(t|d)$ represents the topic distribution in document d, denoted as θ, which follows a Dirichlet distribution with a prior parameter α; $p(w|t)$ represents the word distribution in topics, denoted as φ, which is also Dirichlet distributed with another prior parameter β. T is the total number of topics. The goal of the modelling is to estimate the posterior distribution $p(\theta, \varphi \mid W, \alpha, \beta)$, which can be accomplished by likelihood maximization through different approaches such as Gibbs sampling and variational Bayes inference (Blei et al., 2003).

By feeding the vehicle profile collection as the input, the LDA outputs a set of vehicle types, with each representing a pattern of vehicle daily usage, described by a distribution over different trips. At the same time, for each vehicle profile, a distribution over each usage pattern contributing to the profile will also be computed. The pattern with the highest weight could be assigned as the deterministic usage type for the corresponding profile.

Lastly, to build an LDA model, the optimal number of topics has to be pre-assigned to the algorithm as an input parameter. Similar to K-means, there is no explicit answer to such a question. The commonly used practice is to try out with different candidate numbers and compare with the model performance of probability of likelihood. The log-likelihood and perplexity scores were used for the performance evaluation, with the model implemented using the Scikit-learn environment as in (Hoffman et al., 2010).

2.3. Application and Results

2.3.1. Settings

The dataset of FCD on February 7th, 2019 (Thursday) over the Paris Region was adopted for carrying the case study. As the goal is to investigate regional vehicle motility patterns, only vehicles residing in the territory should be selected, considering that the principle of traditional surveys is based on households. The residential vehicles of this region were selected when the first trip origin was inner the region on February 7th, and on February 8th and 9th as well. Another data cleaning was done by removing the data stemming from the problematic devices with an uncommon recording frequency such as those with outdated configuration (5 minutes recording frequency) and inflated with erroneous recordings. Besides, due to the inefficiency to do the data exploration on the whole large dataset, which is up to 62,681 vehicles with records over 1.4 million, random sampling was done based on the unique vehicle IDs to downsize the dataset. As a result, 5000 random vehicles were selected in this study, with 4113 detected as residential vehicles and 3749 with valid recordings further.

Besides the FCD, the data of the urbanization morphological division over the Paris Region were obtained from IAU 2017 (*Découpage Morphologique d'Île-de-France*, 2017) to help distinguish the geographical pattern of vehicle mobility. Upon that, the Paris Region was divided into 4 morphological sectors for simplification according the rate of urbanized spaces, work and population density, which could reflect an urbanization level of the land occupation (Proulhac, 2019). The geographical areas of the sectors and corresponding descriptions are illustrated in **Figure 2.1**.

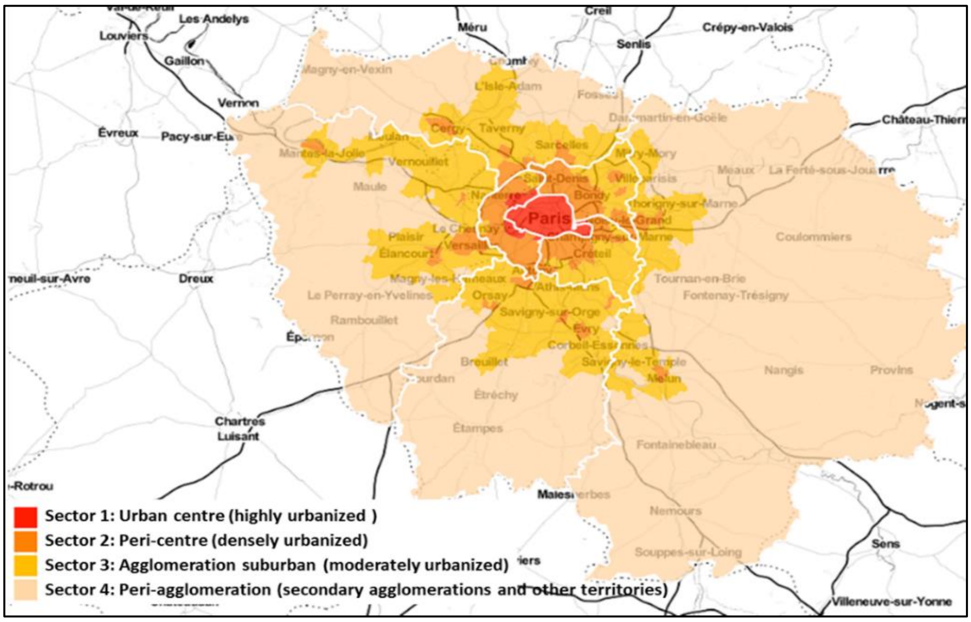


Figure 2.2 Morphological division of the Paris Region

2.3.2. Trajectory Segmentation Results

The trip detection method was implemented for this case. Thresholds were analysed and obtained based on the full dataset to ensure the general applicability. By inspecting the distribution of time intervals, a threshold of 90s was determined, which is consistent with recording frequency considering some fluctuations. As for the gap distance ratio i, the assumption was confirmed by checking empirical distributions of selected intervals that can represent the certain scenarios. Figure 2.3(a) and 2.3(b) shows the ratio distribution of very short intervals (<=90s) and very long intervals (>2,400s) respectively, which confirms that the i is close to 1 for on-journey scenarios and 0 for those making trip ends for long activities.

Confounding cases with mixed scenarios can be seen from the empirical distribution of the interval range (90-2,400s), shown in Figure 2.3(c). A Gaussian-distributed kernel density estimator was run to model the distribution mixture of mixed scenarios. The result is shown in Figure 2.3(d), which shows a mixture of 3 major distributions. As we were interested in detecting major trip ends, only the distribution (centred 0) was considered. The first local minimum at the value of 0.2 was adopted as the upper bound threshold for the detection.

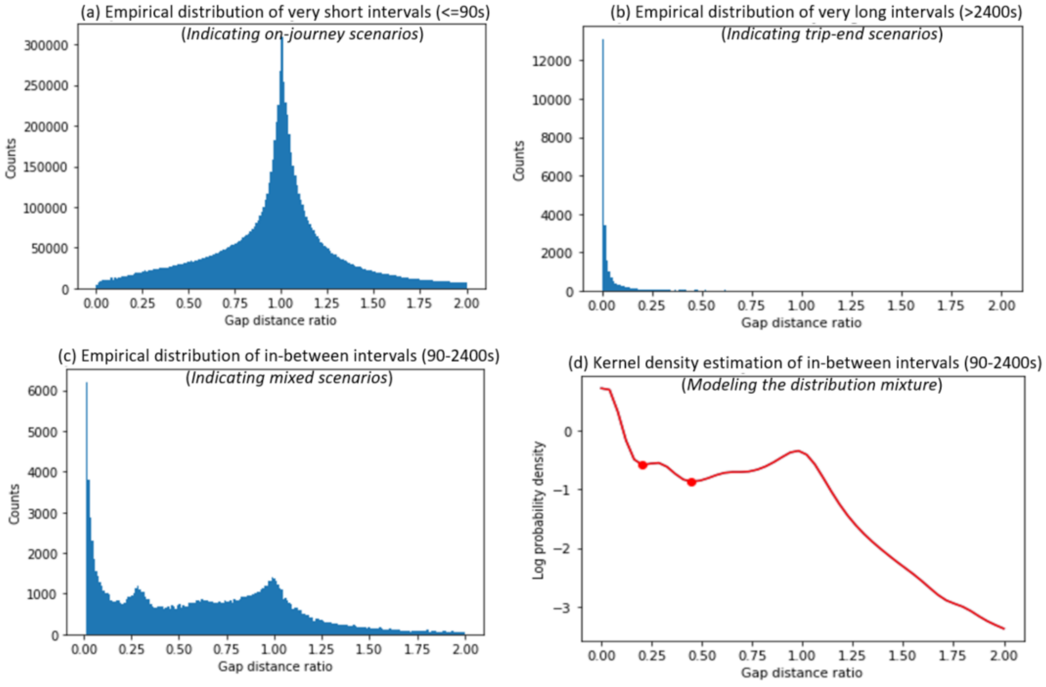
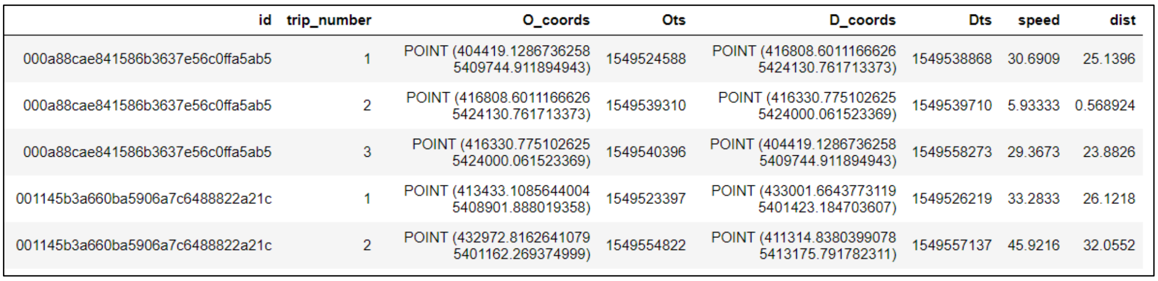


Figure 2.3 Gap distance ratio *i* : (a) *i* <= 90s, (b) *i* > 2400s, (c) *i* ∈ (90, 2400]s, and (d) kernel density estimation of *i*

Consequentially, a new dataset of segmented trips was generated along with a series of describing features computed from raw FCD. The layout of the data constituted by single trips is illustrated in Figure 2.4, which includes trip frequency (*trip_number*: the sequence order of the trip), origin-destination locations (*O_coords and D_coords*: geo-coordinates in WGS 84), time window (*Ots* and *Dts*: timestamps in Unix time), trip distance (*dist* in km) and pointwise average driving speed (*speed* in km/h). As a consequence, 12,928 trips were identified for the 3,749 vehicles with an average trip frequency of 3.44, which is consistent with the EGT household travel survey conducted in 2010 with an average car trip number extracted as 3.56 (B. Yin, 2019). The overall statistics of trips from FCD is summarized in Table 2.2.



id	trip_number	O_coords	Ots	D_coords	Dts	speed	dist
000a88cae841586b3637e56c0ffa5ab5	1	POINT (404419.1286736258 5409744.911894943)	1549524588	POINT (416808.6011166626 5424130.761713373)	1549538868	30.6909	25.1396
000a88cae841586b3637e56c0ffa5ab5	2	POINT (416808.6011166626 5424130.761713373)	1549539310	POINT (416330.775102625 5424000.061523369)	1549539710	5.93333	0.568924
000a88cae841586b3637e56c0ffa5ab5	3	POINT (416330.775102625 5424000.061523369)	1549540396	POINT (404419.1286736258 5409744.911894943)	1549558273	29.3673	23.8826
001145b3a660ba5906a7c6488822a21c	1	POINT (413433.1085644004 5408901.888019358)	1549523397	POINT (433001.6643773119 5401423.184703607)	1549526219	33.2833	26.1218
001145b3a660ba5906a7c6488822a21c	2	POINT (432972.8162641079 5401162.269374999)	1549554822	POINT (411314.8380399078 5413175.791782311)	1549557137	45.9216	32.0552

Figure 2.4 Sample of reconstituted single trips

Table 2.2 Data description of recovered trips

	Average	Median	1st Quartile	3rd Quartile
# Trips per vehicle	3.44	3.00	2.00	5.00
Speed (km/h)	29.55	25.91	14.34	40.95
Distance (km)	22.61	13.58	4.22	31.58
Trip Duration (h)	1.06	0.54	0.25	1.11

2.3.3. Vehicle Usage Patterns

2.3.3.1. Trip Types

The recovered 12,928 trips were clustered into 5 clusters, and visualized in Figure 2.5 from 2 different viewing angles. The feature statistics of each type is summarized in Table 2.3. By comparing with the features, the clusters can be characterized as 5 trip types as follows.

- Trip type 1 (T1): Morning relative short distance trips with low speed (M-short trips)

- Trip type 2 (T2): Evening relative short distance trips with low speed (E-short trips)

- Trip type 3 (T3): Morning relative medium distance trips with medium to high speed (M-medium trips)

- Trip type 4 (T4): Evening relative medium distance trips with medium to high speed (E-medium trips)

- Trip type 5 (T5): Long distance trips, with departure time most around morning (Long trips)

Table 2.3 Identified trip types with features

Trip type	Counts	Departure time (h)	Speed (km/h)	Distance (km)
T1	3706	8.5	14.9	7.7
T2	3911	15.8	18.6	9.7
T3	2496	6.9	41.4	32.0
T4	1935	15.4	55.0	32.9
T5	880	9.6	50.3	93.5

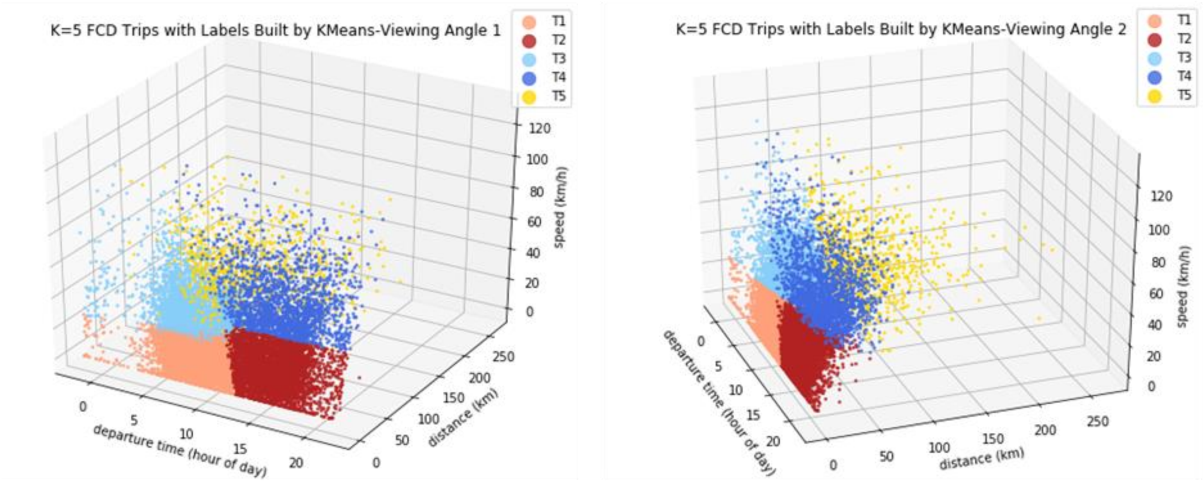


Figure 2.5 Trip clusters resulting from k-means with 2 view angles

2.3.3.2. Vehicle Usage Types

The result obtained by the LDA method was compared with that by running K-means on the vehicle profiles, which was proposed in our previous study (D. Sun et al., 2020) and used as the baseline method for the comparison.

The K-means method resulted in an optimal partition of 4 clusters *(VC1 to VC4*, also indicating vehicle usage types by K-means) with the statistics over different trip types summarized using boxplots in Figure 2.6. However, only the trip travelling patterns along the j_{th} dimension (trip counts of n_{wj}) were used to form the input vectors for the vehicle profile, due to the incapability in processing the high-dimensional sparse matrix if involving the geographical features. While for the LDA solution, 5 usage types were detected. Patterns of mobility-making for each type are visualized by heatmaps in Figure 2.7, where the rows represent the trip types by travelling patterns and columns represent the destination geo-sectors. The colour gradient in the heatmap indicates the probability of travel, with the darker the higher probability for that trip pattern to be made. The results by the LDA method exhibit an overall consistency with the K-means clusters obtained on the sole basis of the trip profiles (i.e. the *j–th* dimensions of travelling patterns). Except for topic 5, the other four topics correspond to a counterpart among the clusters from K-means, with similar pattern along the x-axis. However, the LDA method can further discriminate the vehicle mobility patterns in terms of geographical dimensions (y-axis in Figure 2.7), which outperforms K-means. Besides the superiority in processing sparse high-dimension data, the LDA method also describes vehicle usage in a probabilistic way over the different trip patterns, which is more practical than using mean values for the representation.

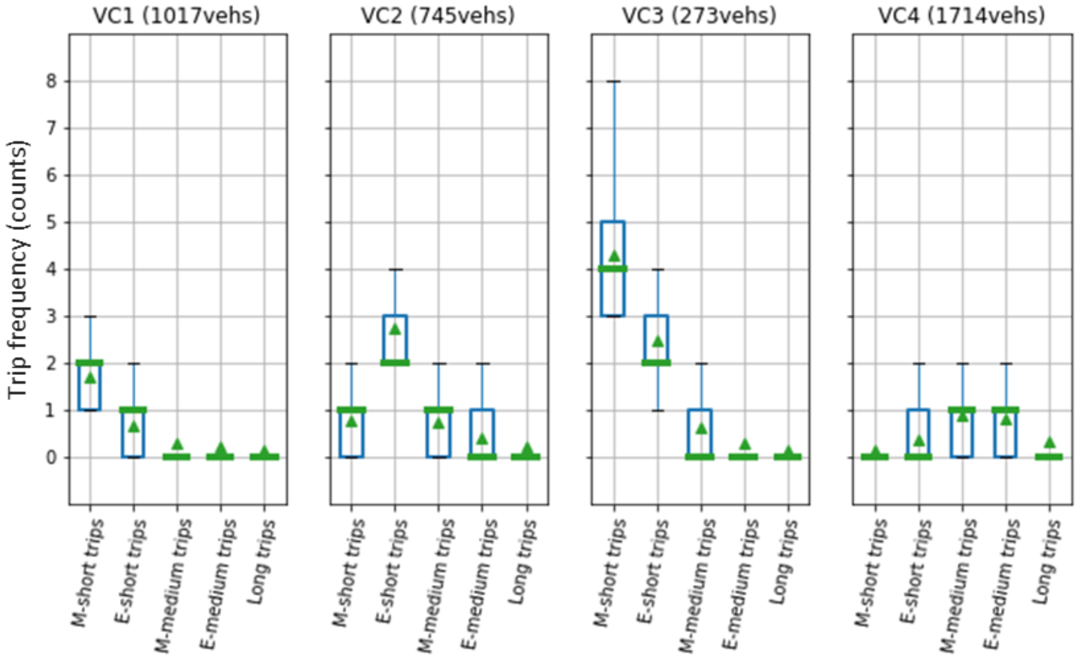
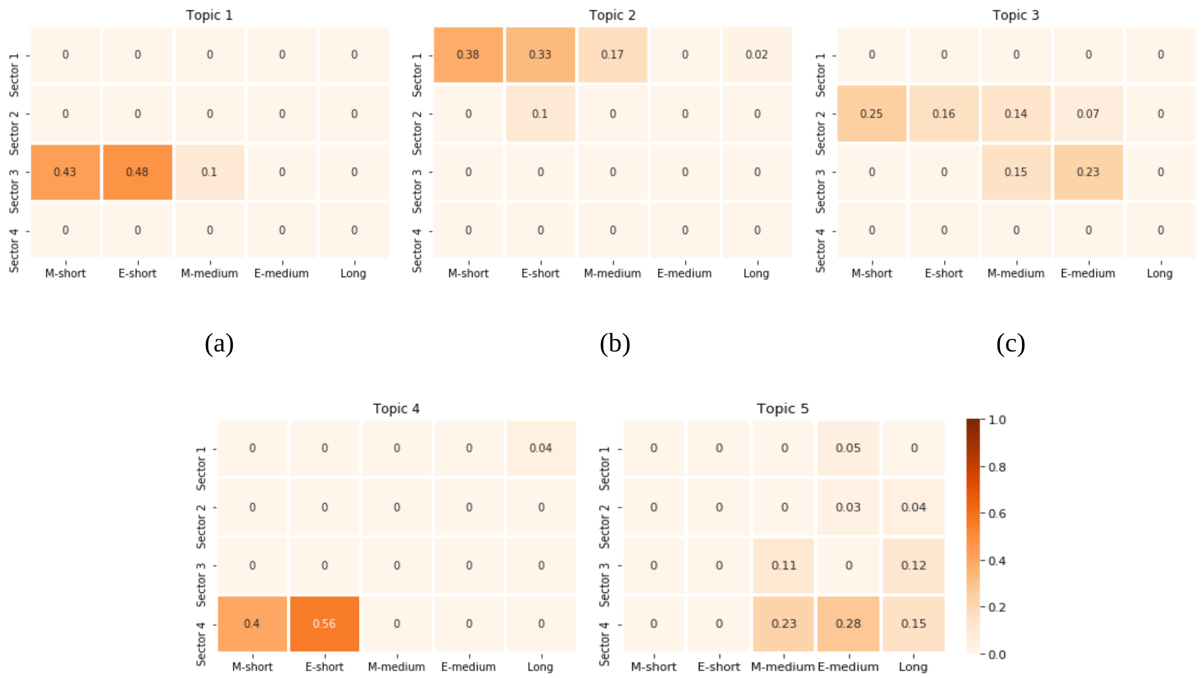


Figure 2.6 Vehicle trip-making types built by K-means



(a) (b) (c)

(d) (e)

Figure 2.7 Trip patterns and geographical sectors of 5 vehicle usage types (5 topics)

2.3.3.3. Vehicle Usage Typology Annotation

Based on the underlying profiles, the five LDA-discovered topics can be interpreted as the following vehicle usage types:

- Topic 1 concerns suburban local oriented usage (Figure 2.7(a)). This topic is mostly constituted of the vehicle usage travelling to sector 3 in short trips with a balanced morning and evening temporal distribution.

- Topic 2 concerns urban centre connected usage (Figure 2.7(b)). This topic mainly has the trips to sector 1, with a minor portion towards sector 2. The overall pattern is short-medium travelling with a bit more activity tendency in the morning. Some evening trips to sector 2 may represent the back to home trips, revealing a commuting linkage between sector 2 and sector 1.

- Topic 3 can be summarized as pericentre-suburban circulating usage (Figure 2.7(c)). This topic mainly consists of the trips to sector 2 and sector 3, which indicates a relative higher circulating behaviour between these two areas. The usage pattern shows a major tendency of short trips when travelling to sector 2 and a mixed but higher probability of longer trips when travelling to sector 3.

- Topic 4 pertains to peri-agglomeration local oriented usage (Figure 2.7(d)). Similar to topic 1, this topic shows a significant localized usage within sector 4 with short trips and balanced temporal distribution.

- Topic 5 can be inferred as long-distance oriented usage (Figure 2.7(e)). This topic has a prominent distribution over long distance travelling trip patterns. Although the trip destination varies among the 4 sectors, it overall represents the intra-sub-region travelling behaviours. A higher distribution shows up in sector 4, which is consistent with the higher vehicle dependency for long distance activity in urban outskirts.

Table 2.4 shows the population of each topic and sub-population in each frequency range. Each vehicle was assigned to its dominant topic, which had the highest probability among all the topics. Overall, topic 1 and topic 3 contribute the most to the collection of profiles, which indicates a comparatively leading vehicle usage population around pericentre and near suburban areas based on the studied data. Since the topic matrix of each usage type (as shown in Figure 2.7) could only show the trip probability distribution for making trips, the pattern of total travelling frequency was further investigated. The sub-types in this cross-tabulation (Table 2.4) could reflect the frequency pattern indicating higher activity-used vehicles or commuting oriented ones. A configural frequency analysis was then employed to examine the association between the types and the trip frequency intensity. More details of the method can be referred to from Sun et al., (2020) and Von Eye, (2003). As a result, a statistically significant higher trip frequency was found among suburban and peri-agglomeration oriented vehicle usage (Topics 1 and 4). Lower trip frequency was found significantly over-represented for pericentre-suburban circulating vehicle usage and long-distance travelling vehicle usage (Topics 3 and 5), while the prior one may indicate a high commuting pattern for that type of usage.

Table 2.4 Vehicle trip frequency counts over different detected usage types

Vehicle typology (topics)	**Trip frequency level**			
	Low (<= 2 trips)	Medium (3-5 trips)	High (>=6 trips)	Total vehicles
Topic 1	29%	29%	42%*	1074
Topic 2	40%	31%	29%	564
Topic 3	45%*	30%	25%	1066
Topic 4	30%	24%	46%*	344
Topic 5	48%*	31%	21%	701

* configuration “significant more”, tested at a confidence level of 99% by a Z score test.

2.4. Conclusion and Discussion

This research developed an explorative approach on FCD to discover daily usage patterns of vehicles. The objective was to identify vehicle usage types so as to get insights in major vehicle usage groups and their behavioural tendency over a large-scale regional territory. A statistical trip detection model was firstly proposed to recover mobility features at the trip level from vehicle trajectories. Trip types were then identified through a clustering analysis based on departure time, trip distance and driving speed. Upon that, a mobility profile was built for each vehicle in terms of identified trip types and the geographical sectors of trip destinations. Lastly, a topic-modelling approach was developed based on LDA to discover vehicle usage patterns by regarding profiles as documents, trips as words and usage types as latent topics to be determined.

The proposed method has been applied over the Paris Region for a case study. As a result, five major vehicle usage types were identified based on the generated vehicle profile collections: three types were associated with short trip usage within local areas, whereas the other two showed hybrid travelling patterns between different areas. A comparative leading usage was found on short-medium trips around pericentre and near suburban areas, where urban sprawl grew with an increasing habitation but a relative lack of mass public transit coverage. Although sampling bias is a common problem for the currently available FCD datasets, such method can be expected to provide more representative results with the progressive generalization of connectivity among future vehicles.

The practical contribution of this study was twofold. One was showing a big data driven instance in obtaining mobility behaviour knowledge based on modern mobility footprint data. The other was that the proposed approach provided a way to analyse vehicle mobility via usage segmentation and territorial differentiation. By linking the vehicle typology to other specific phenomena of interest such as emission, accidents, roadway usage etc., the relation outcomes could further aid future roadway planning, policy making and many other user-based mobility applications. Methodologically, the proposed topic modelling dealt efficiently with the data sparsity and high dimension problems arising in the analysis of mobility profiles. The analytic framework is transferrable to other territories and can be easily scaled for more scenarios in terms of larger regions and day-to-day analysis.

As the data available for this study were restricted to vehicle usage, future work can be done to extend the analysis for other trip modes so as to get a bigger picture the mobility system. By integrating other data sources, more features can be used to describe mobility and discover its patterns. Results could also confront with that of the field surveys or other sources of information with prior knowledge. Lastly, other machine learning methods may also be explored and compared to improve the identification of vehicle usage types.

CHAPTER 3. DISCOVERING INDIVIDUAL SIGNIFICANT PLACES BASED ON MOBILITY REGULARITIES

Adaptation of the paper published in:

Sun, D., Leurent, F., & Xie, X. (2021). Mining Vehicle Trajectories to Discover Individual Significant Places: Case Study using Floating Car Data in the Paris Region. *Transportation Research Record*, 2675(8), 1–9.

The candidate contributed to designing the method, performing the study and writing the manuscript.

Abstract

In this study we discovered significant places in individual mobility by exploring vehicle trajectories from Floating Car Data. The objective was to detect the geo-locations of significant places and further identify their functional types. Vehicle trajectories were firstly segmented into meaningful trips to recover corresponding stay points. A customized density-based clustering approach was implemented to cluster stay points into places and determine the significant ones for each individual vehicle. Next, a two-level hierarchy method was developed to identify the place types, which firstly identified the activity types by mixture model clustering on stay characteristics, and secondly discovered the place types by assessing their profiles of activity composition and frequentation. An applicational case study was conducted in the Paris region. As a result, 5 types of significant places were identified, including home place, work place, and 3 other types of secondary places. The results of the proposed method were compared with those from a commonly used rule-based identification, and showed a highly consistent matching on place recognition for the same vehicles. Overall, this study provides a large-scale instance of the study of human mobility anchors by mining passive trajectory data without prior knowledge. Such mined information can further help to understand human mobility regularities and facilitate city planning.

Keywords: Trajectory mining; Significant place identification; Mobility pattern; Floating Car Data

3.1. Introduction

Mobility has been growing rapidly in recent decades. Understanding human mobility patterns has been widely acknowledged as a critical task in studying urban dynamics and facilitating sustainable development (Zheng et al., 2014). With the rising implementation of location acquisition technologies (GPS, GSM, Wifi, etc.), massive data of location traces are becoming available and are providing fundamental knowledge with which to explore insights into people's movements. However, such raw data merely trace geo-locations with timestamps, and therefore need to be mined to recover meaningful information from their content. One key challenge is then to recognize meaningful locations that are important to mobility makers (Suzuki et al., 2019). Such locations can be termed as "significant places" including home place, work place and other regular visit places which shape the mobility pattern of an individual (Ahas et al., 2010). Evidence from many studies also indicates that human mobility is predictable and centered around a few base places (Andrade et al., 2019; Herder & Siehndel, 2012), while home and work places are commonly considered as the two critical base places in activity chain modeling (Valiquette & Morency, 2010) and transport planning (Chakirov & Erath, 2012). Therefore, identifying the significant places of individuals is a fundamental issue in mining mobility traces and can help us to better understand human mobility regularities.

In the previous literature, significant places have been described by different authors using different terminologies, including anchor points, core stops, meaningful locations, etc. (Ahas et al., 2010). However, a common core definition refers to the locations where people regularly stay and carry out activities. These relevant works can generally be summarized into 3 kinds. The first kind of work is on activity location classification, which mainly aims to tag trajectory segments and locations with semantic meanings that are learned or trained from GPS data with labeled indicators. Corresponding methods include supervised classifiers, such as random forest models (Witayangkurn et al., 2015), and sequence inference models such as Conditional Random Field (Liao et al., 2007), Dynamic Bayesian Network (J. Yin et al., 2004) and Integer Linear Programming (Suzuki et al., 2019). A common issue of these studies is the requirement to have prior knowledge of the trajectory meanings. Ground truth or labels are commonly required in the classification and parameter estimation, which is hard to obtain, especially in large scale applications. The second group is about place inference via geographical databases.

Points of interests have been widely incorporated in order to infer activity place types (Furletti et al., 2013; Huang et al., 2010). Other data sources containing similar information may also be explored, such as the Yellow Pages (X. Cao et al., 2010). However, this kind of place inference has rarely involved home place detection due to the lack of residence information in these datasets. The last kind involves mining place representations from human mobility data, in line with the interest of this study and further to the current trend of privacy protection which limits the information collected beyond anonymized traces. Ashbrook et al. (Ashbrook & Starner, 2003) inferred significant places by clustering GPS terminated points into location clusters based on a variant of the K-means algorithm. A further study by Zhou et al. (C. Zhou et al., 2007) proposed a density- and join- based clustering approach called DJ-cluster to detect significant places from GPS traces, which showed improved precision compared to the K-means. However, these two studies mainly focused on the identification of the geo-location of significant places, while the issue of place type inference has not been well explored. Besides these studies, mobile phone calls have been analyzed by Ahas et al. (Ahas et al., 2010) and Vanhoof et al (Vanhoof et al., 2018) to detect home, work and secondary places. Rule-based approaches were proposed for such an extraction based on the call-related local features of time of day and repetition over days. Chakirov and Erath (Chakirov & Erath, 2012) developed a similar approach based on decision rules to identify activities and their primary location based on smart card payment data from the public transportation system. One common issue of these rule-based methods is the difficulty of determining the cut-off thresholds, which may vary from one application to another. Besides, most of the above studies modeled on a relatively small sample size with only a few selected users. The need for a method suited to large scale data would therefore present itself when city-wide applications are considered.

The objective of this study was to identify significant places on the basis of unsupervised mining of mobility data. Vehicle trajectories traced by Floating Car Data (FCD) were used for the exploration in the Paris region and aimed at showcasing the method applicability on GPS based trajectory data in reality. More specifically, the geo-locations of significant places of each vehicle were firstly recovered from raw trajectories. After that, a two-level hierarchy approach was developed to identify the types of significant places by analyzing the activity type based on the characteristics of each visit, and then to infer the place type of the individual upon its activity composition and frequentation. Comparing to previous efforts, the contribution of this work was to identify significant places on a large scale without prior knowledge in terms of both of the pre-labeled data for referential training and the expertise on setting specific rules for place recognition. More so, we aimed to identify the geo-locations and functional types of these significant places for not only the primary ones (home and work places) but also the other secondary ones.

3.2. Significant-Place Geo-Location Detection

Trajectories collected by FCD are represented as sequences of timestamped traces with no direct semantic meanings. This section describes the mining task conducted to process the trajectory and recover the geo-locations of significant places. Hereby, some preliminary notions are provided: 1) A trip indicates a segment of a trajectory which a vehicle makes for a certain purpose of movement; 2) The temporal interval between two trips indicates the time duration of an activity; 3) A stay point indicates the geo-position at the end of a trip; 4) A place indicates a space where people carry out activities, the geo-location of which can be represented by a cluster of sufficiently adjacent stay points.

3.2.1. Dataset

FCD in the Paris region were analyzed in this study with a time span of 14 days in February 2019. Trajectories of 168,308 unique vehicles with a total of over 15 million logs were obtained within the Paris region during the time period. However, with computation efficiency in mind, a pre-selection was performed to downsize the dataset. Consequently, 10,000 vehicles with a frequent usage, that is, defined as having been used on at least 7 distinct days, were finally adopted in the following analysis. It should be noted that this does not affect the methodological findings in a meaningful way as the significant places were identified on the level of each individual vehicle.

3.2.2. Stay Point Detection

The first step of mining was to sequence the trajectories of each vehicle into trips and extract the stay points where the trips end. Generally, the time interval between two succeeding points is most widely adopted as the decisive feature for identifying trip ends, with other criteria incorporated by some studies including distance and average interval speed (Gong et al., 2014, 2015). However, thresholds of those criteria varied from one to another subject to the data source and local context. In this study, we further developed the trajectory sequencing method that was proposed in our previous paper (D. Sun et al., 2021a), described in *Chapter 2*, to segment the trips and make an additional extraction on the stay points. To summarize, the method adopts two main criteria: the time interval t and the interval distance ratio i, which leverages both distance and speed information recorded in the data with the formula given as below (**Equation 3-1**).

$$i = \frac{observed\ distance}{expected\ distance} = \frac{distance\ (P_a\ to\ P_b)}{0.5*(V_a+V_b)*t} \qquad (3\text{-}1)$$

where $distance\ (P_a\ to\ P_b)$ indicates the planar distance between point a and point b of an interval, and V_a and V_b indicate the recorded instantaneous speed of the two points respectively. The process of the algorithm is illustrated as shown in **Figure 3.1**, where the principles of the thresholds can be referred to from *Section 2.2.1*. For qualified intervals, the geo-position of P_a was used to approximate the geo-location of the true stay point as it was the last recorded point of the trip. It should be noted that although not all trips could be detected, this algorithm aimed to capture those with major trip end activities and without involving too many “false positives”. As a result, a new dataset of detected stay points was generated along with corresponding descriptive features including geo-locations, arrival times, and activity durations. A total of 307,623 trips were identified for the 10,000 vehicles.

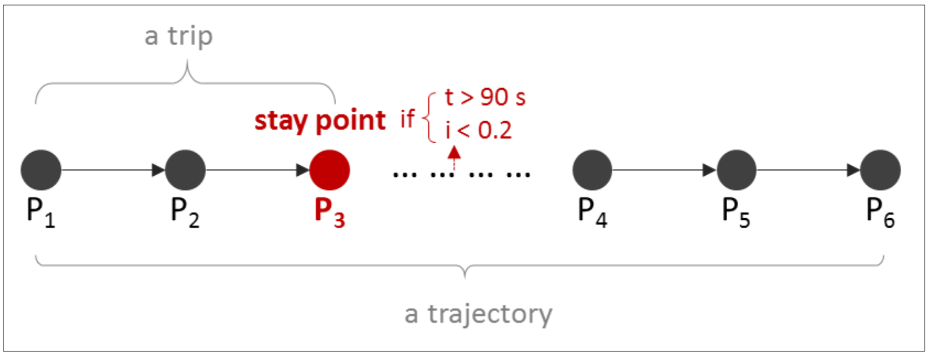


Figure 3.1 Trip and stay point detection

3.2.3. Stay Point Clustering to Extract Significant Places

Stay points at the end of trips can indicate the place where activities are carried out. However, due to the complicated nature of geo-locations, stay points with slightly different geo-positions may correspond to the same place. The second step of mining was therefore conducted to cluster those adjacent points into spatial clusters, using their centroids to represent the geo-locations of meaningful places (Ye et al., 2009). The process is illustrated in **Figure 3.2**. In the meantime, significant places can also be extracted by selecting spatial clusters with multiple visits by the same vehicle. To achieve the above goals, density-based clustering methods can be employed to detect spatial clusters based on geo-coordinates (Vanhoof et al., 2018; Ye et al., 2009). The widely used DBSCAN clustering method was adopted in this study. Such a method requires two input parameters, namely distance threshold (*eps*) and minimum number of points (*minPts*) to form a cluster. Considering the semantic interpretations, the *eps* was set as 150 m and the *minPts* was set as 2. One common issue of this algorithm is that some large clusters may be formed by straight cluster chains if many places are densely connected in urban settings. As well as adjusting the thresholds, we customized the clustering process by running it for individual vehicles separately to avoid such chaining cases, the results of which were visually inspected on their geographical distributions and fitted well in our application. As a result, the geo-locations of 44,882 significant places were identified with a vehicle average of 4.48.

It should also be noted that places were treated separately with different vehicles by bundling place identifiers with vehicle identifiers. This is because one same place can have different meanings for different visitors, e.g. a shopping mall may represent the work place for vehicle user 1, while just playing the role of an entertainment stop for others.

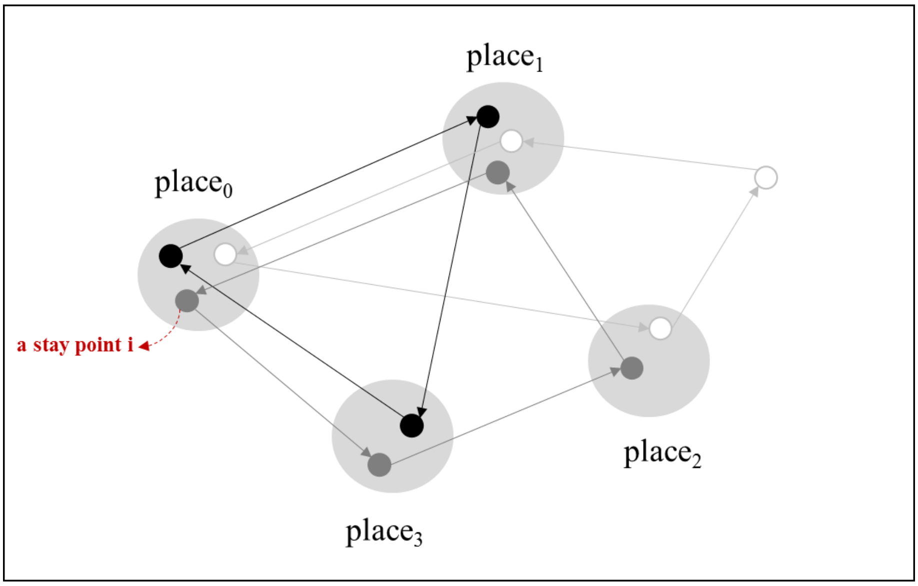


Figure 3.2 Stay point clustering into places

3.3. Significant-Place Type Identification

Places may play different roles in people's life, with some places visited more frequently and others less. Investigating visit characteristics at places can help to differentiate their types so as to understand individual mobility regularities. In this section, the significant places extracted earlier were examined for each vehicle with the aim of identifying their functional types such as home place, work place and other types of frequent visits.

3.3.1. Method: A Two-level Identification

In previous works, features related to significant place recognition can generally be summarized into two groups: 1) temporal circumstances: including time of the day, day of the week (Tongsinoot & Muangsin, 2017; Vanhoof et al., 2018) and activity duration when visiting the place; 2) repetition pattern: frequentation among distinct days of visits to the place (Vanhoof et al., 2018). Additionally, it is reasonable to assume that activities that occur at a place would strongly indicate the type of place (Liao et al., 2007). In this study, a two-level identification method was proposed, which analyzed the temporal circumstances at *level 1* to identify the activity type of each single stay and then identified the place type at *level 2* based on its activity composition and repetition among different days. The concept of the hierarchical framework is illustrated in **Figure 3.3**.

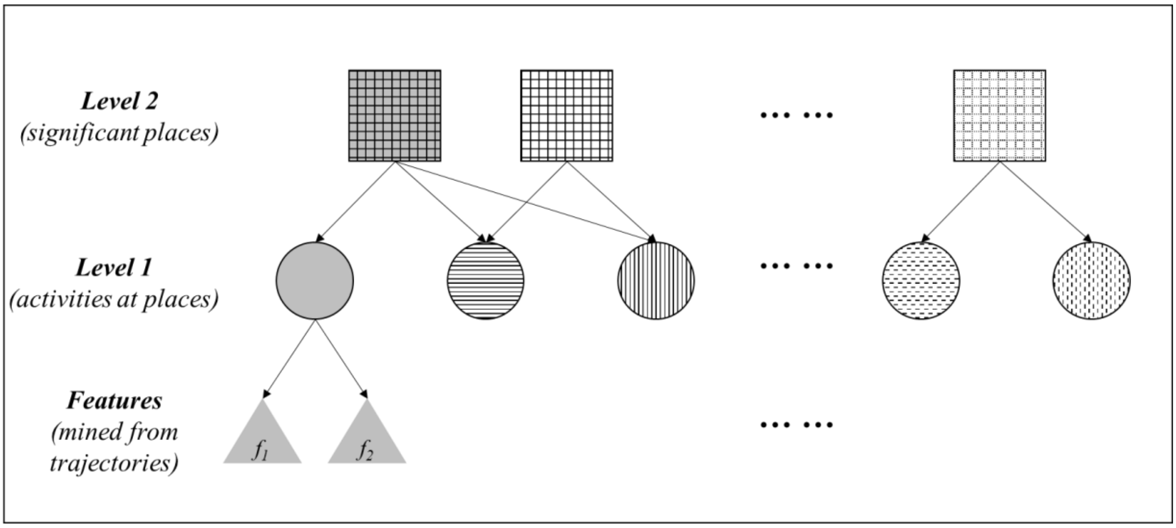


Figure 3.3 The concept of the 2-level identification of significant-place types

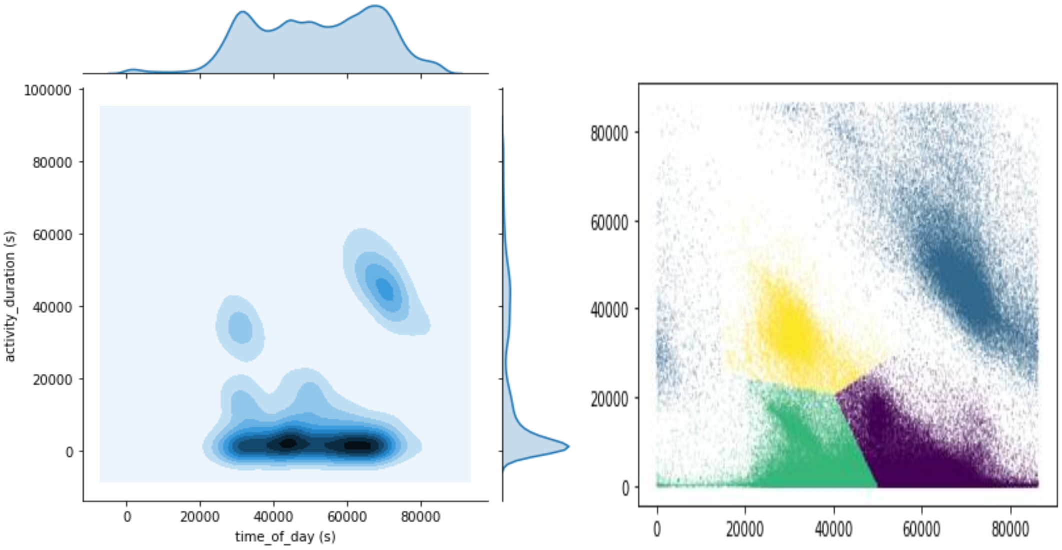
3.3.1.1. Stay-Activity Type Identification by Gaussian Mixture Model (GMM) Clustering

For each stay activity, the visit time of day (arrival time) and the activity duration were extracted from the raw trajectory to feature their characteristics. Weekday and weekend scenarios were treated separately in view of the different contexts. The distributions of such activities against the features are plotted in **Figure 3.4(a)** and **Figure 3.5(a)**. As can be seen, there are multiple density peaks implying a mixture of different distributions, which indicates different activity types. Gaussian Mixture Model (GMM) clustering was employed to identify these different distributions. GMM is a probabilistic model which assumes that observations (activities) are generated from a mixture of a certain number of gaussian distributions with unknown parameters. The parameters of each Gaussian can be estimated through an expectation maximization (EM) algorithm by fitting them to the observations via an iterative process. After the parameters are known, the probabilities of each observation belonging to different Gaussian distributions can be derived so as to cluster them into different groups, namely activity type $j_{n=0,\dots,N}$. Such a model fits well in this case due to its advantages in detecting overlapping clusters and oblong shapes. The number of components can be determined according to prior knowledge of the data distribution or by assessing relevant metrics such as the Bayesian information criterion, distance between GMM distributions, and so on. More details can be found from (*2.1. Gaussian Mixture Models — Scikit-Learn 0.23.1 Documentation*, n.d.; Press et al., 2007).

3.3.1.2. *Significant-Place Type Identification Based on Activity Composition and Frequentation*

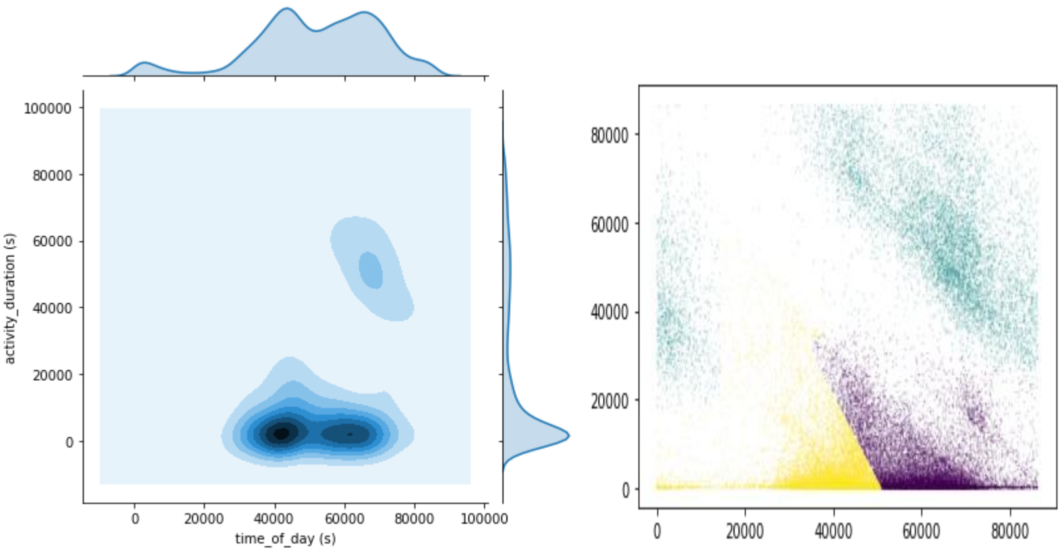
The functional types of places are characterized by their related activities. For a place i of individual k, an activity profile can be built by counting the frequency in different activity types j over distinct days, which can be formed as a vector of counts $\boldsymbol{F}_{ik} = \left(f_{ikj_0}, \ldots, f_{ikj_N}\right)$. To exclude the effects of vehicles not necessarily being used every day or being used equally frequently, each count f_{ikj} was further adjusted to a conditional relative frequency by dividing it by the total usage day numbers of the corresponding vehicle. Such a conditional relative frequency f_{ikj} reflects the likelihood of the activity being carried out in the place as a daily average. This adjusting process also helps to weigh each feature dimension equally in the following clustering process.

The K-means clustering algorithm was employed to partition these profiles to detect different place types. The optimal number k of clusters is commonly recognized as a subjective issue and depends on the essence of the data used for partitioning. In this study, the two widely used methods, the elbow method and the average silhouette method, were employed to determine the optimum cluster number. By comparing the suggested good candidates from the two analyses, a final choice of the optimal number of k can be determined.



(a) (b)

Figure 3.4 Plots of activities by time of day and duration on weekdays



(a) (b)

Figure 3.5 Plots of activities by time of day and duration on weekends

3.3.2. Results of Activity Type Identification

GMM clustering was run using the Scikit-learn implementation. The number of mixture components was determined as 4 on weekdays and 3 on weekends, as visually suggested from the density distribution plots (shown in **Figure 3.4(a) and Figure 3.5(a))**. The stays were therefore partitioned into 4 clusters for weekday scenarios and 3 for weekend scenarios, as shown in different colors in the scatter plots in **Figure 3.4(b)** and **Figure 3.5(b).** Considering the cyclical nature of time, activities that were after midnight but before dawn were considered as belonging to the previous day. By comparing the attribute characteristics, the activities can be characterized into 4 and 3 types respectively, along with descriptive statistics summarized in **Table 3.1**. Both the weekday and weekend scenarios were found with an activity type with a long duration and visiting time in the late part of a day, implying those trips of returning home journeys at the end of a day. Compared to weekends, there was one more type detected on weekdays, that with a long duration and visiting time in the early part of day. Such a type can be interpreted as being related to work, which is consistent with the finding that they were only majorly detected on weekdays. The other two types, early-day short and late-short activities, were found for both weekdays and weekends, but with the duration slighter longer and the visiting time a bit later on weekdays.

Table 3.1 Identified activity types and corresponding statistics

	Identified activity types	Visiting time of day (seconds*)		Activity duration (seconds)	
		Mean	*Std..*	*Mean*	*Std..*
Weekday (wkd)	Early-day short	36436	7948	4249	5187
	Early-day long	31526	6036	33738	6608
	Late-day short	60287	8853	4842	5864
	Late-day long	65981	15084	49481	13438
Weekend (wkn)	Early-day short	37467	9268	5218	8267
	Late-day short	60613	9476	6215	7490
	Late-day long	56883	23357	54878	15029

* Seconds for time of day start as 0 from midnight.

3.3.3. Results of Significant-Place Type Identification

Seven place clusters (c0-c6) were drawn from the K-means clustering on their activity profiles. The occurrence likelihoods of activities $\left(f_{ikj_0}, \dots, f_{ikj_N}\right)$ are displayed for each place cluster by heatmaps in **Figure 3.6**. The color gradient in the heatmap indicates the mean value of the likelihoods in each cluster. There was a total of 7 activity types $(j_{n=0,1,\dots,6})$ according to the activity type identification. The qualitive definition of "frequent" is a subjective issue and we assumed that a likelihood of over 0.33 could be regarded as a frequent pattern as it implies at least a visit every 3 days. By comparing the corresponding prominent characteristics, the places were characterized into 5 types, as described below:

- *Home places (c0 & c5)*: These two clusters shared similar patterns by showing a significantly high chance of late-day long activities on both weekdays and weekends, which is a typical pattern for home places or residence places. Compared to c0, places in cluster c5 also included more of the other types of short activities with a greater emphasis during the second half-day. This might indicate that vehicles pertaining to those home places were more frequently used between the home and other places rather than simply commuting behaviors.

- *Work places (c2 & c6):* Places in cluster c6 showed a comparatively high likelihood of early-day long activities during weekdays, which implies that their major place occupations are related to long working stays along the day. Cluster c2 showed frequent visits for short activities in the morning and afternoon only on weekdays. This corresponds to the pattern of work places for those who leave their work places during the noon break for another purpose such as lunch at home or at restaurants. It should be noted that these work-related activity likelihoods were not found to be as high as close to 1. However, such a finding may be consistent with the fact that people may not drive cars to work every day.

- *Secondary places I -Weekend frequently visited places with late-day short stay (c4)*: This cluster showed a pattern of frequent short visits in the second half-day on weekends, which may represent secondary places for typical weekend activities, such as frequently visited shops, recreation spaces or favorite meet-up places.

- *Secondary places II - Weekend moderately frequently visited places with early-day short stay (c3)*: Places in this cluster encountered a few more visits during the first half-day periods on weekends, which can be interpreted as secondary places such as markets or bakeries.

- *Secondary places III - Weekday less frequently visited places (c1):* Places in this cluster did not show any significantly frequent visits, which may indicate places that individuals come to and visit from time to time (at least twice) but with no significant regularity. Overall, the activity patterns were more distributed among weekdays.

Consequently, 8,111 out of the 10,000 vehicles studied were identified with home places and 4,295 vehicles were identified with work places. Such a portion of recognition is not high, especially with work place detection, but is consistent with the fact that home place normally display more regular patterns on a vehicle usage basis while patterns of work places do not, among individuals. The analysis timeframe of 14 days was also a limitation restricting the detection of frequent patterns.

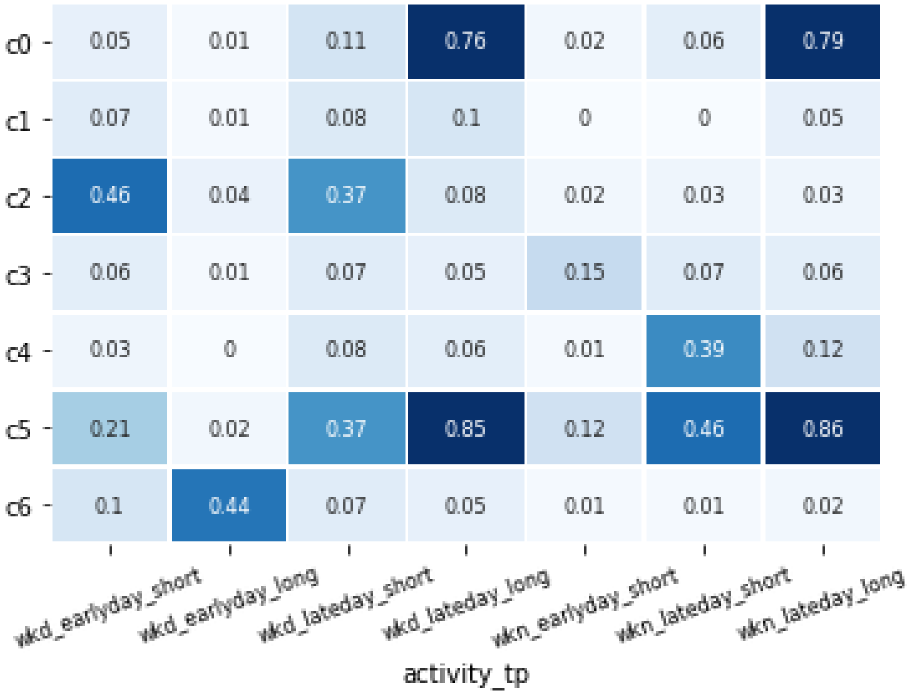


Figure 3.6 Identified place clusters and corresponding activity frequentation patterns

3.3.4. Comparison with Rule-Based Identification

Validation of such explorations is a common problem in current research into mobility pattern mining due to the lack of ground truth for the evaluation. Within this restriction, the results using the proposed significant place identification method were compared with those from a benchmark method: rule-based extraction, which was widely used in previous studies (Ahas et al., 2010; Chakirov & Erath, 2012; Vanhoof et al., 2018). The comparison was mainly on home and work places due to their ability to be characterized with explicit criteria. The criteria for the decision rules were set for the features directly on the place level, described as below:

- Home place: 1) High probability (p>0.5) as the first/end place visited during a day sequence. 2) High attendance among distinct days (>80%) when the vehicle is used.

- Work place: 1) More daytime (8h-18h) activities at the places; 2) Hight attendance during weekdays (>60%); 3) Long average activity duration (>2 hours)

Through the rule-based extraction, 5,822 vehicles were found with home places and only 1,375 vehicles among them were detected with work places. The use of cut-off rules is limited in detecting diverse patterns, but we can assume that the extracted parts are representative of home and work places. The comparison between our method and the rule-based method was based on the 5,822 vehicles by considering the type labels of the rule-based method as the base reference. The comparison results are summarized in **Table 3.2**, with each cell representing the number of places labeled in the two methods. All numbers were counted on the unit of significant places identified for each individual. As a result, our method showed a high consistency ratio against the rule-based method, with 96% on home places, 90% on work places and 89% on the secondary places.

Table 3.2 Comparison between our method and rule based identification

By our method	**By rule-based identification (considered as base reference)**			**Consistency ratio**
	Home	Work	Secondary	
Home	5601	107	116	96.2%
Work	26	1254	108	90.3%
Secondary	188	2063	18282	89.0%

3.4. Conclusion and Discussion

This study proposed a methodological approach to identify individual significant places by mining vehicle trajectories from FCD. A case study was conducted in the Paris region for an experiment. Meaningful trips were segmented from raw trajectories to recover stay points. Customized density-based clustering by DBSCAN was implemented to cluster stay points into places and extract the geo-locations of significant places for each individual vehicle. Next, a two-level hierarchy method was developed to identify the types of these significant places. On the 1st level of the process, it identified the types of each stay activity by Gaussian Mixture Model clustering based on the visit time of day and the activity duration. On the second level, it found the types of places by clustering their profiles in terms of occurrence likelihoods with different activities. As a result, 5 types of significant places were derived, including home place, work place, and 3 other types of secondary places. Most of the vehicles analyzed were detected with home places while around half of them were found with work places based on frequent vehicle usage. The results of the proposed method were also compared with those from the commonly used rule-based extraction method, and showed a highly consistent matching. Moreover, based on unsupervised mining, our proposed method was less dependent on the expertise of rules and more applicable for detecting places displaying more diverse situations. Although the findings of this paper were based on the vehicle mobility due to data essence, the proposed methodology is transferrable to other kinds of GPS based trajectory data (by cell phone, GPS tracker, etc.) in a straightforward way.

Overall, this research provides a large-scale instance of identification of significant places in terms of both geo-locations and types without prior knowledge, which shows its applicability to human geography in a cost-effective way by leveraging digital trajectories. The information mined can be used as bases for further studies of human mobility regularity and predictability, and thus be of benefit to future city planning. The results were restricted due to a limited 14-day period of available data. Future work could be done on a broader timespan to explore the place visiting frequency on a larger scale, such as monthly regularities. Data from other modes of mobility may be explored together to obtain a more complete view of individual mobilities. Complementary sources of data, such as land use and geographical reference data may also be incorporated to facilitate the place identification or confront with the results for validation.

CHAPTER 4. ESTIMATING ROADWAY TRAVEL TIME USING FLOATING CAR DATA INTERVALS

Adaptation of the article presented at:

Leurent, F., Sun, D., & Xie, X. (2021). Roadway Travel Times: Maximum Likelihood Estimation Based on Floating Car Data Intervals. *hEART 2020: 9th Symposium of the European Association for Research in Transportation*, Lyon, France.

The candidate contributed to processing the data, performing the study and writing the manuscript.

Abstract

Massive Floating Car Data (FCD) datasets have become available for roadway networks, which contain travel time information on short spatial intervals between pairs of successive observations along individual trips. This paper brings about a stochastic model of travel times with a Maximum Likelihood estimation method to exploit FCD material. Probabilistic specifications are put forward for link travel times as Gaussian random variables along with standard error of each estimator. This allows for simple estimation of link attributes based on "Link FCD intervals" and their confidence intervals. An application instance was dealt with for one motorway and one urban avenue in the Grand Paris area with results showing better accuracy than automotive methods based on pointwise average speed.

Keywords: Roadway travel time; Gaussian link time; Maximum likelihood estimation; Floating Car Data Intervals

4.1. Introduction

Roadway networks are purported to be traveled along by different kinds of vehicles. On any usage occurrence, the individual user makes his or her trip along a selected path. The path travel time is a major characteristic as the path costs time to its individual user: it is usually the main basis for path choice and also for departure time choice and travel mode choice (Ortuzar & Willumsen, 2004). Automated personal travel assistants such as Google Maps, Waze and so on provide path advice on the basis of local travel time as the first and foremost criterion. Thus, local travel times determine path choice, hence the formation of trip flows and in turn the local traffic conditions. At the same time, with the great diffusion of GPS technology, massive Floating Car Data datasets have become available for roadway networks. They contain travel time information on short spatial intervals between pairs of points that are successively recorded along individual trips, which enables to recover local traffic characteristics.

Travel time estimation as a key factor in understanding traffic patterns has been a recurrent research issue in the last few decades. Many well-established technologies have been developed based on loop detectors, vehicle diaries and video cameras (Mori et al., 2015). However, those traditional data collection methods, are inherently limited for wide application concerning its spatial-temporal coverage (D. Sun et al., 2014). Most recently, with the increasing diffusion of GPS technologies, Floating Car Data is emerging as massive available for collecting traces of a wide range of network all day long, which shows a great potential to resolve the data concern in travel time estimation (Jenelius & Koutsopoulos, 2013). Although much more attention has been aroused recently to studying FCD for traffic analysis, the literature for it on travel time estimation is still limited, in particular on the use of low-frequency floating car data, a more practical trend for the data source nowadays (Jenelius & Koutsopoulos, 2013; Mori et al., 2015).

This paper focuses on the local travel time estimation on the link level. In the literature, the current studies can be generally divided into two streams: data-based approaches and model-based approaches. For the prior ones, taking the average/median of all observed points to recover space-mean speed for link travel time estimation is widely adopted in many FCD based studies (Cheng et al., 2015; Ehmke et al., 2012; Fusco et al., 2016; Long Cheu et al., 2002; Ran et al., 2016; Shen & Ban, 2016; X. Wang et al., 2015). To be simplistic, we define this way as “pointwise average speed”. This method has the advantage of being straightforward but is limited to the case when there are sufficient observations over the targeted areas. As for the model-based approaches, a few studies built probabilistic graphical models from observed probe traces to obtain the travel time probability distributions in terms of a series of spatial and temporal traffic variables. A Bayesian Network was proposed by Hunter et al. (2009) for structuring the probabilistic model using low-frequency sparse taxi probe data to estimate the historical link travel time distributions. Development of such an approach was conducted by Hofleitner, Herring, Abbeel, et al. (2012) to focus on travel time forecasting, which proposed a dynamic Bayesian network model to model the state transition between neighboring segments. In another study by Hofleitner, Herring, & Bayen (2012), the authors did a further development by incorporating the traffic physics, the flow theory and state variables considering the number of querying vehicles and turning fractions at intersections. In addition to the Bayesian network, Ramezani & Geroliminis (2012) proposed a Markov chains model to estimate the arterial route travel time distribution. Other than probabilistic models, a study by Jenelius & Koutsopoulos (2013) developed a statistical regression model based on taxi probe vehicle data to estimate the travel time on urban road network as well as analyzing the impacts of the corresponding influencing variables.

Although such models have the advantage to take many comprehensive factors into account, they also require more external data in terms of the physical and spatial parameters to express the functions more precisely, which limits the large-scale applicability. In the meantime, the complexity of model structuring also restricts the transition to other cities for the variance of network structure and huge computation workload. Another persistent issue in most of the existing studies is the lack of a measure of reliability in the travel time estimation (Mori et al., 2015). Confidence intervals rather than just a unique value of average time would be especially helpful to provide a more complete information to the road users.

Acknowledging the need to address the above-mentioned problems, this paper aims to build a stochastic model of local travel times together with a Maximum Likelihood estimation method to exploit FCD material. Probabilistic specifications are put forward for link travel times as Gaussian random variables. This allows for simple estimations of link attributes based on “Link FCD intervals”. Analytical properties are to be obtained specifically at sub-links along with variance models dealt with postulates. An application study is dealt with for a major motorway segment as well as an urban link for comparison in the Grand Paris area.

4.2. Methodology

In the stochastic model, the travel time is analyzed as a random variable that adds up local random variables that involve local characteristics: we introduce a set of assumptions and derive some theoretical properties, including a Probability Density Function (PDF) for the travel time. At the link level, the local characteristics include the mean and standard deviation of local speed. This is for homogenous sections excluding link endpoints.

The estimation method takes the travel time PDF as a likelihood function for field observations of individual travel times. The network framework enables us to gather large samples of individual trips and extract the associated information by using an ad-hoc method of Maximum Likelihood Estimation. As for application instance, we have availed ourselves of an FCD dataset provided by the Coyote firm: car trajectories are monitored with one geolocation time stamp per half minute. Every pair of two successive individual timestamps contains information on the network conditions in-between. Our method to exploit such information is complementary to the link time estimation methods based on instant speeds monitored at timestamp points (Cheng et al., 2015; Long Cheu et al., 2002).

Let us consider travel times $\Delta h \equiv h' - h$ between point pairs $(\mathrm{M}, \mathrm{M}')$ along link a, separated by spatial length $\Delta s \equiv s' - s$. We model any Δh as a random variable, with stochastic characteristics that depend on the link conditions and the associated parameters. Our modelling assumptions are:

- (L1) that the average time $\mathrm{E}[\Delta h]$ is proportional to the spatial length Δs, with factor coefficient τ_a:

$$\mathrm{E}[\Delta h] = \tau_a \Delta s \quad (4\text{-}1)$$

- (L2) That the variations of the travel time come from a stochastic process along space with autocorrelation function $\chi_a(s, s')$: then,

$$\mathrm{V}[\Delta h] = \chi_a(s, s') \qquad (4\text{-}2)$$

Special instances will be considered to make the model simpler. Our basic specification is that local variations are mutually independent and identically distributed per unit of distance: then, denoting by σ_a the standard deviation of local variations (per length unit), it holds that

$$\mathrm{V}[\Delta h] = \sigma_a^2\,\Delta s \qquad (4\text{-}3)$$

The reason is that the variance of the sum of independent local variables is the sum of their respective variances, therefore leading to linear dependence according to length Δs under the assumption of homogenous distribution. The product form relies upon the hypothesis that successive intervals are statistically independent. Under the Gaussian assumption, the likelihood function of a link interval is simply:

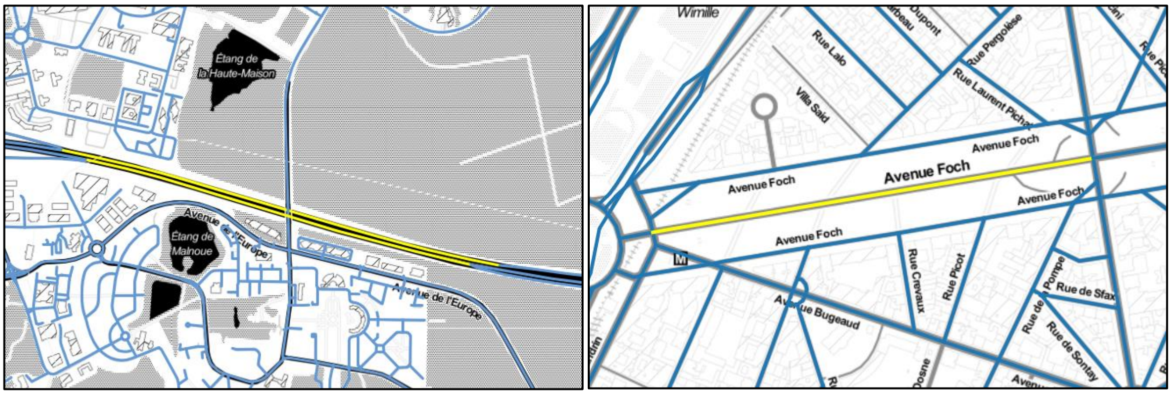
$$L_{ui}(\Theta_a) = \mathrm{f}(h_i^+, h_i^-, s_i^+, s_i^-, \Theta_a) = \frac{\exp(-\frac{1}{2}\frac{(\Delta h_i - \tau_a \Delta s_i)^2}{\sigma_a^2 \Delta s_i})}{\sigma_a \sqrt{\Delta s_i}\sqrt{2\pi}} \qquad (4\text{-}4)$$

We obtained analytical formulas joint the Maximum Likelihood estimation of the average time and variance parameters, as well as the standard error of estimation associated to each estimator. All formulas are easy to calculate so that the estimation method is straightforwardly applicable, and its accuracy can be controlled.

4.3. Application and Results

4.3.1. Study Location

The proposed method was applied on two different roadway segments, with the aim to compare the experimental results between highway setting and urban setting. The highway segment was selected from a major link along the motorway A4 in Great Paris region, which performs as a main arterial serving the traffic between the center and eastern sub-regions. The urban segment is chosen from on the Avenue Foch, which a major avenue in Paris. Geographical layouts are shown as in **Figure 4.1**. Travel time of the two-directional movement was studied separately. No ramp access or intersection was included in this application as the model focuses on the link travel time. The segment length is 1445m and 1635m for the eastbound and the westbound direction respectively on the A4 motorway segment, and 607m for both directions on the urban segment.



(a) Segment along A4 motorway (b) Segment along Avenue Foch

Figure 4.1 Studied roadway segments

4.3.2. Dataset

FCD over two normal weekdays (February 05 and February 06, 2019) on the selected segments were analyzed. The network roadway data were extracted from OpenStreetMap. Due to imperfect recording of GPS coordinates, the deviation between FCD points and road network is quite common. Numerous effective map-matching algorithms were developed by previous studies (X. Liu et al., 2017; Newson & Krumm, 2009). In this study, the FCD was map-matched to the nearest roadways according to the travelling directions and re-projected the locations to the nearest foot-points on the segment.

4.3.3. Link Interval Extraction

Link intervals along the segments were extracted for the two directions in two days respectively. Each interval consists of a pair of two successive FCD timestamps. Distance travelled along the road to the starting node was also calculated based on geo-coded coordinates using geo-packages in Python. Anonymized vehicle ID was used to track different vehicles. Invalid pairs were excluded if the trajectory time span was abnormal, setting the rule as less than 300s considering consecutive sampling frequency is around 30s. Tolerance was made for in-stable signal condition. A descriptive summary of extracted intervals and all the point-wise observations is given in **Table 4.1**.

Table 4.1 Descriptive summary of extracted intervals

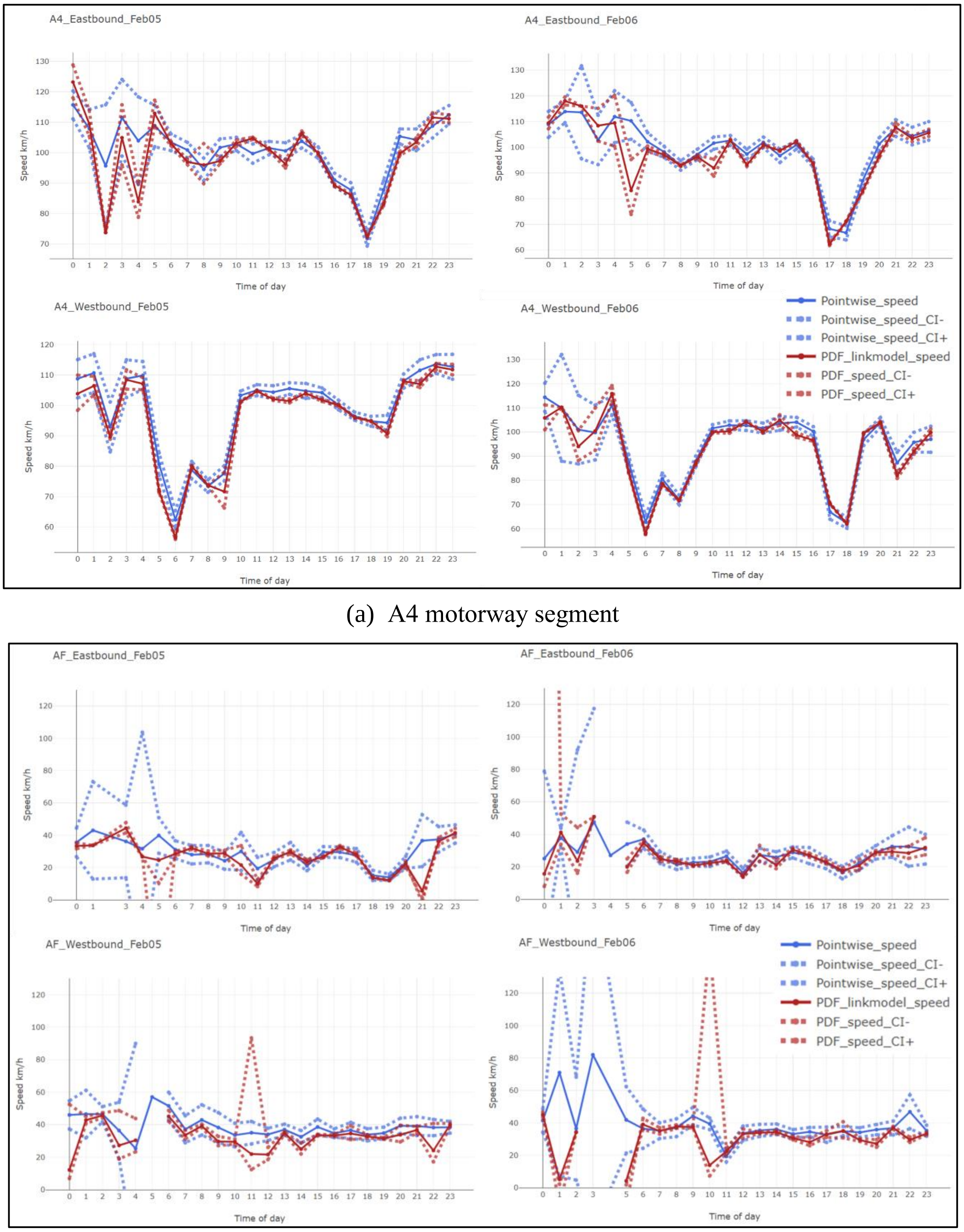
Setting		Eastbound 05	Eastbound 06	Westbound 05	Westbound 06
A4	**Link intervals** (count)	1248	1327	2014	2073
	Pointwise observations	3472	3605	4401	4416
	Average speed overall	96.7 km/h	93.0 km/h	91.5 km/h	86.5 km/h
AF*	**Link intervals** (count)	787	983	452	464
	Pointwise observations	1364	1692	1019	1117
	Average speed overall	23.7 km/h	24.1 km/h	36.7 km/h	34.5 km/h

*AF stands for Avenue Foch

4.3.4. Link Analysis Results

The stochastic parameters of the link model were estimated based on the extracted data by different time of the day. Besides, the corresponding pointwise average speed was also computed. To measure the reliability of the estimation, confidence intervals were computed stemming from those estimated parameters. As a result, line-charts were plotted to shows a detailed comparison between the interval estimation of PDF link model and the point-wise estimation for both the two segments, shown in **Figure 4.2**. Space mean speed was used for the comparison in the plot, as for a given length, modeling space mean speed is essentially equivalent as modeling the link travel time (Hall, 1996; Mori et al., 2015).

As can be seen from the two plots, the space mean speed estimated by the interval estimation is generally consistent with pointwise average speed with a similar fluctuating trend. Significant speed reductions were observed on the motorway segment during peak hours along the tidy movement to and from the city center. The urban segment was observed with less fluctuation but overall with relatively low speed. However, the interval estimation was found more likely to estimate a lower speed than pointwise average especially on the urban segment which involves more congested scenarios with higher variation in vehicular motion. Moreover, the confidence intervals were significantly narrower than those of pointwise average speed for most of the situations, which indicates that the interval estimation could provide a more reliable estimation of the travel time. It was also found that the more data available, the more precise results on the estimation. Nevertheless, the interval estimation would require less data to reach a higher precision level.



(a) A4 motorway segment

(b) Avenue Foch urban segment

Figure 4.2 Result comparison between the interval estimation and the pintwise average estimation

4.4. Conclusion

This paper puts forward a stochastic model of local travel time estimation along roadway links on the network. Basic modeling assumptions were postulated to model link travel times as random variables. Building upon the stochastic model, we have devised a Maximum Likelihood estimation method that can be applied to FCD trajectories along the network. Intervals in time and space between two successive timestamps monitored along the trajectories constitute the basic data. The practicality of the estimation was demonstrated in a case study. Estimations were computed and compared between our stochastic model and straightforward conventional pointwise average method along with confidence intervals. Results indicate that the stochastic model is able to deliver a more reliable estimation and require fewer observations to reach a higher precision. Moreover, the pointwise average estimation was found tending to provide a higher speed than the stochastic model with less certainty. This may lead to an underestimation of the travel time, implicating the limitations of the current applications adopting such a straightforward estimation.

This research is restricted to the link level. However, it could be saved as a modular section. Further research may be invested to build the probabilistic model for node or intersections between different links so as to develop more reliable estimation of path travel time.

CHAPTER 5. DISCOVERING FUNCTIONAL OCCUPATIONS OF TERRITORY ZONES BY MOBILITY ACTIVITIES

Adaptation of the paper published in:

Sun, D., Leurent, F., & Xie, X. (2021). Uncovering mobility typologies of territorial zones based on Floating Car Data mining. *Transportation Research Procedia*, 52, 708-715.

The candidate contributed to designing the method, performing the study and writing the manuscript.

Abstract

This paper describes a data exploration study using Floating Car Data to analyze mobility patterns of geographical spaces. The objective is to build a mobility-related typology of territorial zones by investigating related vehicle movements. Mobility features at the level of trips and stay places are recovered from daily vehicle trajectories. Place visiting frequentation is further analyzed at each vehicle level to identify significant places and corresponding activity regularity. Based on these mined patterns, a multi-view cooperative clustering method is developed to feature out the zonal mobility typology in terms of the composition of local stays, temporal flows of trip generation and attraction, and spatial connections in trip distance distribution. The proposed framework was applied to the Great Paris region for an experiment using 14 days data. Consequentially, 5 mobility types of zones were obtained, with each holding a different orientation of mobility usage. Discovered areas were also compared with the common recognition of their social functions, which showed a consistent matching. Overall, this study provides a data-driven approach to study mobility interactions with territorial spaces, by spatial segmentation, characterization, and differentiation.

Keywords: Territorial mobility pattern; Space clustering; Mobility typology; Floating Car Data;

5.1. Introduction

Understanding mobility patterns has been widely acknowledged as a critical role in urban planning, traffic management, and many other place-based applications. Generally, the spatial configuration may induce mobility generation, while reversely, human movement would further re-impact the space development (Wegener & Fürst, 2004). The evolution between them is reciprocal. Therefore, investigating the mobility pattern of territorial spaces will help to reveal such interaction and provide valuable guidance for future development. With the growing diffusion of location tracking technologies, digital traces are becoming more and more available nowadays. Among them, Floating Car Data (FCD) has emerged as a new essential data source on roadway traffic for a high spatial and temporal coverage, which thus offers a great potential to investigate mobility patterns of territorial spaces.

In the existing literature, FCD was mainly used to determine traffic states, including speed detection, travel time estimation, and congestion prediction (Altintasi et al., 2017; Fusco et al., 2016; Mori et al., 2015). Mobility perspective analysis using FCD was relatively limited. Among those works, some studies analyzed the city structure by deriving hotspots from taxi pick-ups and drop-offs (Jahnke et al., 2017; X. Liu et al., 2015). Major hubs such as airports were also analyzed to further investigate its specific mobility role interacted with the city (Ding et al., 2016). Some other studies analyzed mobility origin-destination patterns aiming to extract major spatial movement and their temporal variations (Ciscal-Terry et al., 2016; Lian et al., 2018). Yet the related studies hardly dealt with the detailed mobility pattern on each intra-homogenous geographical space unit, namely territorial zone at a regional level.

In fact, human mobility dynamics are largely correlated with the spatial configuration. Investigating people's mobility activities could contribute to revealing the social functions of territory zones. Yuan et al. (2014) conducted a study do discover urban functional zones based on human trajectories and POIs by developing a topic-modeling-based approach. Another study by Qi et al. (2011) measured social functions of regions according to the temporal variation of get-on/off amounts from taxi GPS data. These studies showed the potential of capturing spatial functions by mining mobility trajectories. However, an issue of them was found as such region functions were mostly interpreted according to limited perspectives of mobility information, which was either by the time of flows or by the origin-destination transitions. A more comprehensive exploration would therefore be prompted by combining different perspectives together. Another persistent issue is the lack of focus on the activity stay following the movement, which leads to the role of space occupation.

Therefore, the objective of this study is to investigate the mobility pattern of territorial zones by mining multiple aspects of information from Floating Car Data. By building a zonal mobility profile in terms of trip departures and arrivals, and related vehicles' activity context as well as their hour and weekday variation, this research aims to cluster mobility patterns of territorial zones so as to build a functional typology.

5.2. Methodology

Territories are generally formed with different areas to meet the various needs of social dynamics. Functional zones are behavioral-based areas with intra-homogeneity oriented to undertake certain social activities (Dubrova et al., 2015; N. J. Yuan et al., 2014). These functional zones correspond to the idea that a geographical space can be characterized by spatially related human occupations (Tomaney, 2009). These zones can be either developed artificially by urban planners or progressively formulated by human's activities. Land use, administrative boundaries and physical characteristics can contribute to the delimitation of zones (Dubrova et al., 2015). In this study, we propose a 3-stage analysis framework to categorize and differentiate the space uses by human activities based on observed FCD trajectories. The overall processing flows can be illustrated as in **Figure 5.1**.

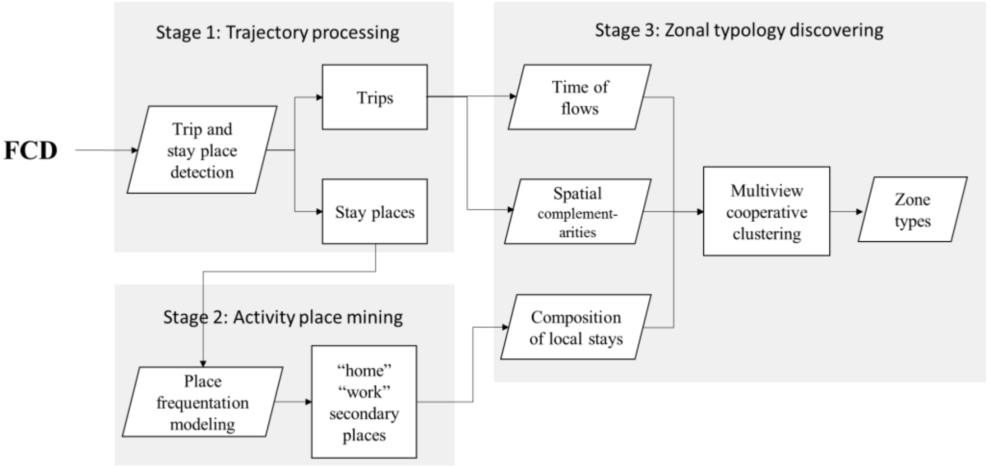

Stage 1: Trajectory processing
FCD
Trip and stay place detection
Trips
Stay places
Stage 2: Activity place mining
Place frequentation modeling
"home" "work" secondary places
Stage 3: Zonal typology discovering
Time of flows
Spatial complement-arities
Composition of local stays
Multiview cooperative clustering
Zone types

Figure 5.1 The 3-stage analysis of space functional type discovering

5.2.1. Trip and Stay Place Mining

Trajectory information by FCD is collected as a sequence of logs (trace points) with no direct semantic representations. The first step of mining is to sequence the trajectories for each vehicle and segment them into meaningful trips. Then the stay points at the end of trips can be concatenated to represent the location visiting history for each vehicle. However, due to the fussiness of locations, stay points with slightly different geo-coordinates may correspond to the same place. The second step of mining is therefore conducted to group up those that are spatially adjacent to recover the meaningful places. Detailed descriptions of the processing algorithms for the two steps can be referred to in *Section 3.2.2* and *Section 3.2.3* respectively.

5.2.2. Place Frequentation at the Individual Level

Places may hold different roles in people's mobility activities, with some visited more frequently and others less primary. Investigating presence frequentation at different places can help to understand individual mobility regularities and further contribute to analyzing the spatial functions. The recovered meaningful places are examined for each vehicle with the aim to identify their significant activity places such as home place, work place and other regularly visited places. It should be noted that the method used in this study is an earlier version than the one developed in the study (D. Sun et al., 2021b), described in *Chapter 3*, due to the insights gained over time during the thesis. The performance comparison between the 2 methods is provided in *Section 3.3.4.* However, for clarifying how the results were obtained in the case study, which was conducted at an earlier time point, the detailed method of this earlier version is described as below.

The place frequentation modeling considers three major features, which are 1) time of the day in visiting the place (Tongsinoot & Muangsin, 2017; Vanhoof et al., 2018); 2) attendance among distinct days in visiting the place (Vanhoof et al., 2018); 3) activity duration at the place. Places are treated separately with different vehicles by bundling place identifiers with vehicle identifiers. This is because one same place may mean differently to different visitors, e.g a place of a restaurant may represent the work place for vehicle user 1, while just playing as an entertainment stop for the others.

A decision rule and clustering combined method is developed to investigate the place frequentation and identify significant activity places. The approaches are conducted at two levels. The first level is to identify primary places including home and work places, which are commonly acknowledged in deriving fundamental mobility activities. As the two kinds of places can be characterized with explicit criteria, a series of decision rules are designed for the identification according to the common sense and criteria proposed in previous literature (Ahas et al., 2010; Vanhoof et al., 2018), which are described as below:

- Home place: 1) High probability (p>0.5) as the first/end place visited during a day sequence. 2) High attendance among distinct days (>80%) when the vehicle is used.
- Work place: 1) More daytime (8h-18h) activities at the places; 2) Hight attendance during weekdays (>60%); 3) Long average activity duration (>2 hours)

As for the other places not detected in primary ones, the second level of identification is to explore their frequentation difference by conducting a K-means clustering analysis. Three features are used for distinguishing, which are 1) Attendance among distinct days; 2) Average activity duration on site and 3) Ratio of visits during daytime. The number of clusters can be determined according to the average silhouette value or the elbow method.

5.2.3. Zonal Mobility Typology Discovery

5.2.3.1. Usage-Based Zonal Attributes

Mobility usage of a territory zone can be naturally described from different perspectives. By exploiting the information presented in multiple views, a more precise data structure can be disclosed (Mitra & Saha, 2019). The mobility patterns of zones are modeled from three views to form the attributes.

- View 1: Composition of local stays. Significant places identified in section 4 are counted in each type j for a zone and used as the attribute in reflecting activity frequentation and duration. Place counts are then converted to percentage values r by zone subtotals and normalized to explore relative difference in composition ratios. Such an attribute is formed as a vector $\{r_{i1}, \dots r_{ij}\}$ for the zone i.

- View 2: Time of flows. Trips leaving from and arriving to the zone are counted by time bins to reflect the temporal mobility pattern. 8 time bins (k) are used by considering both of time of day and day of the week, which are pre-defined as: *Morning (5-10h), Mid-day(11h-15h), Evening(16-21h), Night(22h-4h)* for weekdays and weekends respectively. To weigh each dimension equally, each period count is converted to a period flow factor p by further dividing by the day average trip number of the corresponding zone. The attribute is finally formed as a vector of 16 factors $\{p_{i1O}, \dots, p_{ikO}, p_{i1D}, \dots, p_{ikD}\}$ for the zone i for both trip departures *O* and arrivals *D*.

- View 3: Spatial complementarities. This view looks at the distance ranges of trip generations and attractions. Distance bands by the power of 2 kilometers are used in counting the trips. Counts in different range bands are converted to percentage values d and normalized. The attribute is formed as a vector $\{d_{i2O}, \dots, d_{i2^nO}, d_{i2D}, \dots, d_{i2^nD}\}$ for the zone i for both trip departures *O* and arrivals *D.*

For each analytical zone, mobility patterns (feature vectors) of the above views are assembled and used as the input for base clustering. Results of these base clusterings are further integrated to get a more comprehensive understanding of a zone's mobility characteristics.

5.2.3.2. Multi-View Cooperative Clustering

A multi-view cooperative clustering method is employed to identify the final zone typology by integrating the explorations from different views. Many different terminologies have been used to refer to this kind of method in the literature, involving: cooperative clustering, multi-view clustering, clustering ensemble etc. (Topchy et al., 2004; Cornuéjols et al., 2018; Mitra & Saha, 2019). Overall, a similar aim of such methods is to improve the clustering quality by combining complementary contributions from different base clusterings on different omics of features with each holding its own bias. More hidden patterns are expected to be spotted, which are, otherwise, likely to be diluted if all features are concatenated into a single-view. More so, other benefits include its robustness to data variations, and a wider range of applications on real-world problems of multiple modalities. In our case, the routine approaches conducted are as below:

1) Perform the base clusterings (K-means) on different views of mobility patterns respectively.

2) Integrate base clustering results to build a co-association matrix (Mitra & Saha, 2019), which depicts the average frequency that a pair of objects is partitioned into the same cluster.

3) Build a connection graph of the zones based on the co-association matrix.

4) Perform the spectral clustering to final partition the graph and discover zone typology.

5.3. Application and Results

5.3.1. Settings

FCD of 14 days in the February 2019 over the Paris region was explored for an applicational case study. A total of 168,308 unique vehicles were found showing up within the Paris region during the time period. However, due to the inefficiency to do the data exploration on the whole large dataset with total logs over 15 million, the dataset was downsized by a pre-selection. As a result, 2,500 random vehicles were selected from the pool of those having frequent usage among the 14 days. Such a frequent usage was defined as showing up on at least 4 distinct days and holding at least one re-visiting location.

As for the territory divisions, the zoning of IRIS (Islands Grouped for Statistical Information) was used as the geographical units for zonal typology discovery due to its suitable granulation. Such IRIS zones were developed by INSEE (the national statistics bureau of France), which divides the country into units of equal size on the basis of population with respect to geographic and demographic limitations. The Paris region consists of 5260 IRIS zones.

5.3.2. Identified Trips and Stay Points

Upon the trajectory sequencing algorithm, a new dataset of detected trips was generated along with describing features including origin-destination locations, time window (departure and arrival time), and trip distance. A total of 63036 trips were identified for the 2500 vehicles during the 14 days. The stop points at the end of trips were then clustered into meaningful stay places, by using the DBSCAN clustering method. An overall demonstration of the outcomes of the clustering process can be found in **Figure 5.2**. The stay points were finally grouped into 8 554 clusters. By detecting these spatial clusters, places with the same representations were bundled and activity frequentation in visiting them could be further investigated.

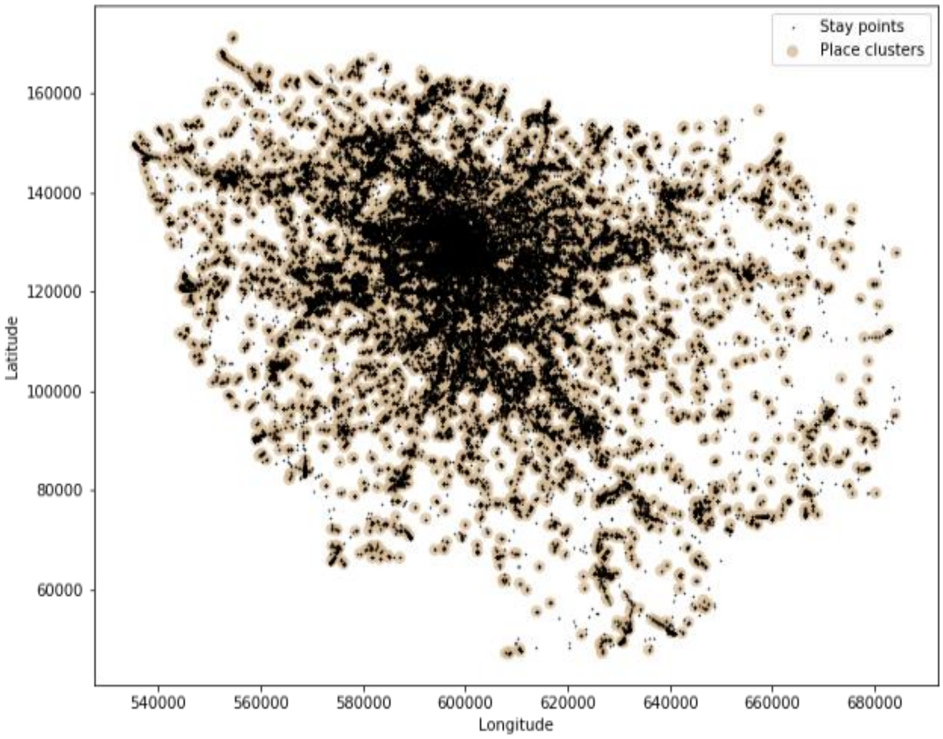
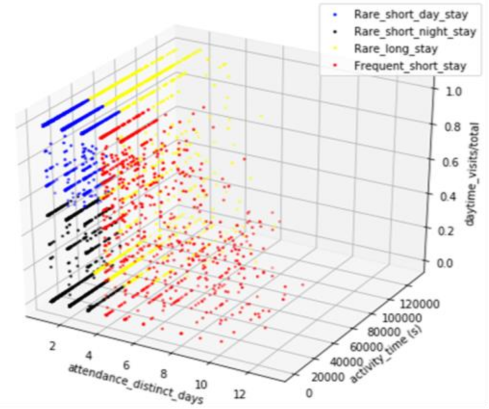
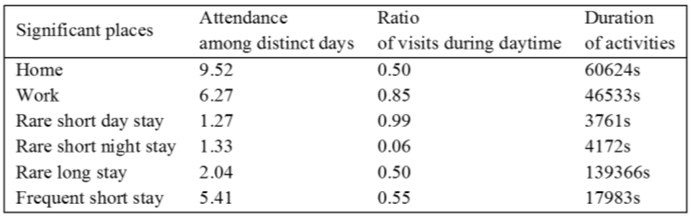


Figure 5.2 Stay points vs Place clusters

5.3.3. Identified Activity Places by Frequentation Modeling

Home and work places were identified according to the proposed rules and the other secondary activity places were further distinguished based on the frequentation difference in features by a K-means clustering analysis. As a result, the secondary places were partitioned into 4 clusters, as shown by a scatter plot in **Figure 5.3(a)**. Each cluster was also characterized with a label as displayed in the figure according to its relatively prominent characteristic. Overall, to sum up the two levels of identification, 6 types of significant places were drawn, the statistics of which were summarized in **Figure 5.3(b).**





Significant places	Attendance among distinct days	Ratio of visits during daytime	Duration of activities
Home	9.52	0.50	60624s
Work	6.27	0.85	46533s
Rare short day stay	1.27	0.99	3761s
Rare short night stay	1.33	0.06	4172s
Rare long stay	2.04	0.50	139366s
Frequent short stay	5.41	0.55	17983s

(a) Clustering results on secondary places (b) Statistics summary

Figure 5.3 Identified activity places by frequentation modeling

5.3.4. Discovered Zonal Typology and Functional Annotations

5 clusters (*C0-C4*) of zones were drawn from the zonal typology discovery by the Multi-View Cooperative Clustering analysis. Geographical distributions of them are geo-visualized in **Figure 5.4(a)** along with their attributes displayed by heatmaps in **Figure 5.4(b)**. The color gradient in the heatmap indicates the mean values of corresponding attribute vectors. By comparing the attribute characteristics, the zones were annotated into 5 types, described as below.

- *C0 Residential oriented areas:* These zones were constituted of higher departure rate in the morning and higher arrival rate in the evening on weekdays, which is a typical temporal pattern of residential areas. A relatively larger portion of home places was also shown in the composition of stays. Spatial travel patterns were mixed with more mid-range and local trips.

- *C1 Commercial-residential mixed areas (locally-accessible-based)*: These zones were composed by more of less-frequently-visited places and home places, which can be interpreted as a mix of residential dwellings and living quarters such as super-markets, cafés, restaurants etc. More activities showed up during mid-day and evening, with no significant flow drops on weekends. The trip distance was more local based.

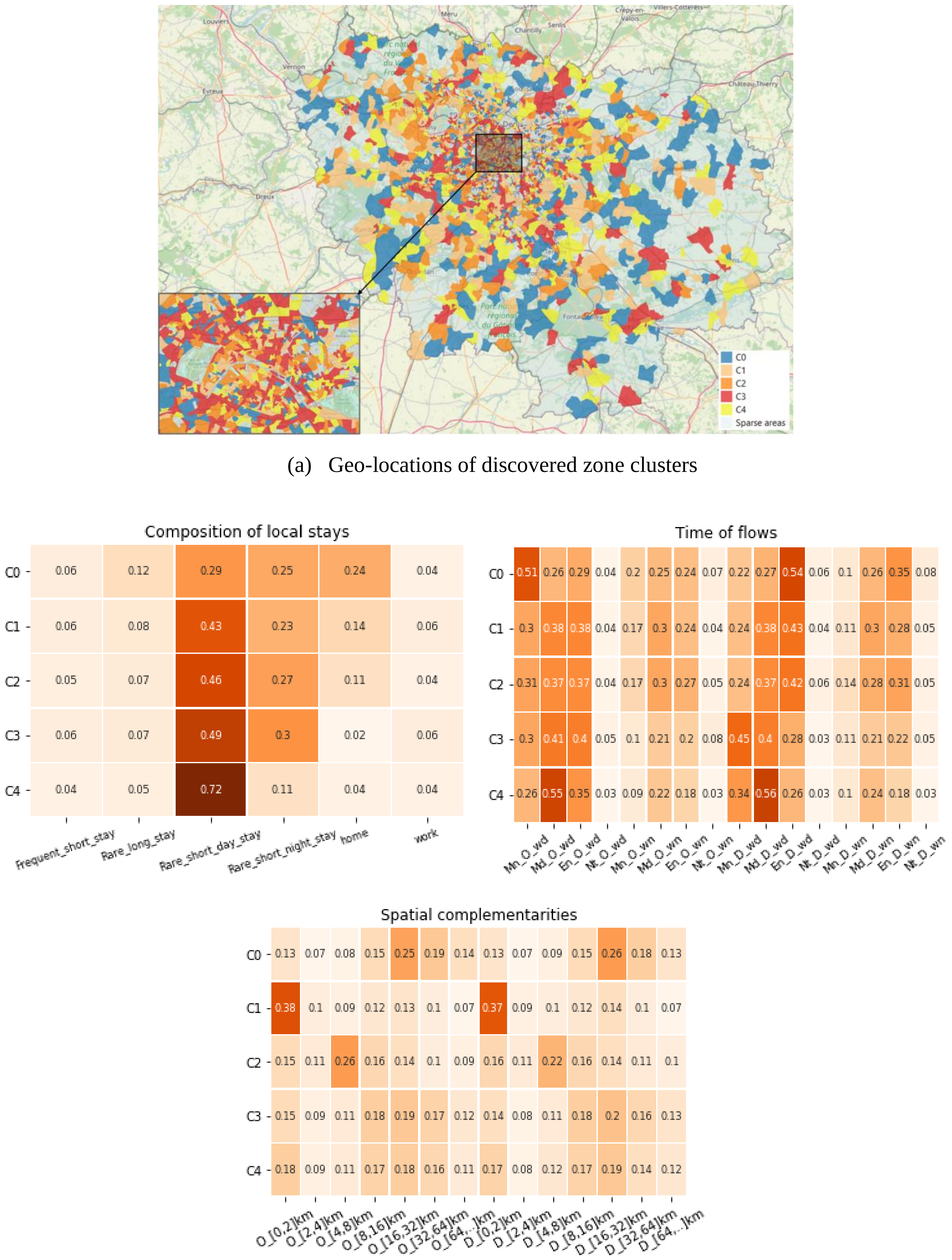
- *C2 Commercial-residential mixed areas (intermediate-range-accessible-based)*: This cluster had a similar pattern with C1, except the spatial patterns were composed more of longer trips.

- *C3 Business/employment oriented areas*: These zones showed up a significantly higher arrivals in the morning and higher departures in the evening on weekdays with the overall volume dropping on weekends. Such time of flows is consistent with employment areas with vehicles used for commuting. Home places were also found least in this cluster.

- *C4 Day-time mobility-oriented areas:* These zones had a maximum portion of less-frequently day-time visited places. Significant more vehicles usage was also presented during the mid-day time period. Such a pattern may correspond to retail shops or commons, town centers in the rural areas and the campuses of educational places, where the visiting time is flexible during the day.

Besides, there were zones without enough observations of trips (<15) for the total of departures and arrivals), so that they were excluded from the clustering analysis and grouped into a separate type, which was annotated as *Observation sparse areas*.

Validation of such explorations is a common problem in published research of mobility trace exploration due to the absence of ground truth for the evaluation. Within this restriction, the results of our zone annotations were inspected by geo-visualizing their locations and comparing the annotation to the common recognitions of their social functions. For example, La defense, Val de Fontenay and Charles De Gaulle airports surrounding areas are commonly known as employment areas consisting of business centers or company facilities, which matches with our identification with these areas clustered into *C3*. More so, the Paris center are highly urbanized areas, especially for the zones of 1e, 2e, 8e 9e arrondissements, most areas of which were detected as commercial or mixed areas. It also has to be noted that many zones were found with mixed place functions which is partially consistent to the fact that land uses are rarely monotonous within an urban area. However, future exploration may be done to with further segmented zone units of the territory.



(a) Geo-locations of discovered zone clusters

*Where *O* indicates departure trips; *D* indicates arrival trips; *Mn* indicates morning; *Md* indicates mid-day, *En* indicates evening, *Nt* indicates, night, *wd* indicates weekday, and *wn* indicates weekend.

(b) Patterns of zonal attribute vectors

Figure 5.4 Discovered mobility types of zones and patterns of their usage attributes

5.4. Conclusion and Discussion

This research proposed a methodological approach to discover mobility typology of territorial space by mining FCD. An application study was dealt with the Great Paris region. Mobility features at the level of trips and stays were firstly recovered from vehicle trajectories. Meaningful places were then detected based on stay points. Place frequentation at different places was modeled at each individual level to identify their anchorage places and other regular presences. Upon those, 3 views of zonal attributes were assembled for each zone to describe the pertaining vehicle usage, including the composition of local stays by frequentation, time of trip flows and spatial complementarities in trip distance distribution. Consequentially, 5 mobility types of zones were obtained through the analysis. The discovered areas were visualized on the map and showed a consistent matching with the common recognition of their social functions.

The explorative approach by multi-view cooperative clustering enabled to explore more patterns from different perspectives and combine them to get a more comprehensive discovery. Practically, this research provides a big data driven instance to analyze mobility interaction with geographical spaces, the knowledge of which may provide useful insights in understanding their heterogeneity and similarity and thus benefit many place-based applications. Our current results were restricted to the vehicle mobility level due to the essence of the data used. However, the proposed framework can be scaled with data of other modes of transport, which may contribute to a more complete picture of the zone mobility typology in the future.

CHAPTER 6. DISCOVERING SPATIAL RELATIONS BY MOBILITY CONNECTIONS

Adaptation of the paper published in:
Sun, D., Leurent, F., & Xie, X. Exploring jobs-housing spatial relations from vehicle trajectory data: A case study of the Paris Region. *Transportation Research Procedia* (in press).

The candidate contributed to designing the method, performing the study and writing the manuscript.

Abstract

This study explores the spatial relations between job and housing locations by mining trajectory data. The objective is to establish a data-driven method to recognize employment core areas and identify their residential catchment areas. More specifically, mobility traces are firstly mined to detect the home and work places of each vehicle according to temporal patterns of the day and repetition patterns over distinct days. A spatial density distribution analysis is then conducted to identify the employment cores and sub-cores. The core-periphery patterns between the cores and other areas are further investigated by building a connection graph based on the home and work locations over spatial zones. The graph-based clustering algorithm is employed to partition the graph so as to identify bonded zones as communities and interpret the catchment areas pertaining to different employment cores. A case study has been applied with the field-collected Floating Car Data in the Paris Region to showcase the method applicability. Overall, this study offers a referential framework for capturing urban spatial dynamics by digital traces, with the advantages of being data-driven, scalable for large-scale, and less dependent on prior expertise. The results may contribute to the planning guidance corresponding to up-to-date changes.

Keywords: Trajectory data mining; Home-work commuting; Employment cores; Catchment areas; Community detection;

6.1. Introduction

Home-work commuting has long been studied in transportation planning and modeling as it is deemed to shape fundamental human mobility (Kung et al., 2014). The jobs-housing spatial relation is also widely recognized as a critical issue in understanding the evolution of urban structure (Han et al., 2015), the insights of which may help to resolve many underlying problems of traffic congestions, demand-supply mismatching, etc. Such an issue was conventionally approached by survey or census data, which are however inherently limited due to belated updates and high collection costs. With the growing diffusion of location-based technologies, digital traces are collected from more and more mobility entities at a large spatial-temporal scale, which brings the potentials to analyze up-to-date mobility patterns. Thus, the increasing pervasiveness of trajectory data prompts a new cost-effective way to study urban spatial dynamics corresponding to the rapid environment changing (Long & Thill, 2015).

Related works have studied the issue of jobs-housing spatial relations using three major kinds of methods. The first kind is to set up empirical rules to deconstruct the city structure such as extracting typical catchment areas of urban cores by a certain threshold of worker/residence rates (Bellefon et al., 2020). This may be limited to the local expertise in determining the cut-off threshold and not adaptive to the dynamic variation. The second kind employs model-based approaches to quantity the relationships between sub-areas, which involves regression modeling by determinants (S. Li & Liu, 2016) and interaction modeling by gravity models (H. Wang et al., 2014) or field strength models (Ferrari et al., 2011; Wu et al., 2020). Such models provide explanations for spatial relations but are still largely dependent on empirical measurements of many external factors such as economic strength, market size, population density, etc. The last kind develops data-driven methods based on various kinds of trajectory data including taxi GPS data (Rinzivillo et al., 2012), smart card data (Han et al., 2015), geotagged tweets (M. Chen et al., 2019; Hollenstein & Purves, 2010), etc. These studies showed the potential of capturing urban structures by finding human mobility regularities from their trajectories. However, most of them were restricted to the detection of hotspots when regarding spatial patterns, and very few of them explored the spatial relations between job and housing places.

Therefore, the objective of this study is to establish a method to explore jobs-housing spatial relations by identifying employment cores and their catchment areas. Unlike the conventional approach based on field surveys, this study builds up a data-driven pipeline to uncover spatial structures based on observed trajectories. The proposed framework can identify the jobs-housing spatial communities directly from the traces without relying on external information or prior expertise. The analysis can be easily replicated with updated data inputs, thanks to the ease of modern data collection, showing a practical potential of capturing up-to-date evolutions, benefiting quick responses to the mobility changes. We utilized the Floating Car Data to showcase the method's applicability with a case study in the Paris Region. Although the empirical findings of the case study were on the basis of vehicle mobility only, they hold specific implications for roadway traffic planning. The proposed methodological approach is expected to be transferrable to other sources of GPS trajectories in a much referential way when further data are available.

6.2. Methodology

6.2.1. Home and Work Place Identification

Mobility contexts of vehicles can be recognized from their daily trajectories, by patterns of their trip-making and activity characteristics, etc. Such patterns can be examined over a sufficiently long period to understand individual mobility regularities and then further contribute to the interpretation of anchor places such as home and work places. In this study, we employed the method developed in our previous paper (D. Sun et al., 2021b), described in *Chapter 3*, to detect the home and work places of each vehicle. Comparing to conventional rule-based methods, this method is exempt from setting specific thresholds and more applicable over diverse situations. The main procedures can be summarized as follows: Vehicle trajectories are firstly segmented into valid single trips to extract corresponding stay points. A density-based clustering approach is then employed to cluster stay points into places considering that adjacent points may correspond to the same place. For each vehicle, the functional types of its frequently visiting places are identified by analyzing the places’ activity profiles (set of activities carried out at the place) based on temporal circumstances (time of day and duration) and repetition patterns (frequentation over weekdays/weekends). Home and work places are finally detected from the clusters with significant patterns of overnight activities and commuting characteristics respectively.

6.2.2. Spatial Distribution Analysis

To understand the structure of a city, one crucial problem is to reveal the spatial arrangement of the city centers and investigate how these centers interact with the rest of the territory (Y. Sun et al., 2016). Different terms have been used to describe such centers in the literature, including city centers, urban cores, and urban areas of interest. With knowing the work places of a large amount of individuals from the above data mining, this analysis aims to reveal the spatial distribution of those places and identify the employment cores and sub-cores.

First, a Kernel Density Estimation (KDE) method is used to transform discrete geo-locations into contours which constructed the density distribution of the work places. The density peaks can suggest the core areas for employments. The optimal bandwidth is determined via the cross-validation method (grid-search experiments implemented by Scikit-learn package) using candidates by every 50 meters ranging from 100 to 2000 meters. It should be noted that although these density contours perform well to overall describe the spatial aggregations, the method itself is limited to delineate the specific boundaries of the cores (M. Chen et al., 2019). Therefore, the second step is to extract the cores with boundaries, for which the family of Density Based Clustering Methods is commonly used following recent advances in the literature (M. Chen et al., 2019). We adopted the hierarchical density-based clustering method (HDBSCAN) by (Campello et al., 2013), as it overperforms the other ones due to its robustness to parameter selection and capability in detecting clusters with various densities. This especially fits well the circumstance that we need a finer-grained clustering in dense areas whereas the extent can be larger in rural areas. Technically, instead of setting up a definite value for the core distance (*epsilon),* HDBSCAN uses a “mutual reachability” to measure the point distances in a relative way. The cluster hierarchy is constructed based on these relative distances so that it enables a detection with various *epsilons* in different areas. The only parameter of HDBSCAN, the minimum number of points to form a cluster (*min_cluster_size*), can be carefully tuned with experiments of sensitivity analysis in trying with different candidate numbers. An appropriate *min_cluster_size* can be determined by comparing the experiment results to the layout of KDE contours. Lastly, a convex-hull geometric method: the alpha shape algorithm (Edelsbrunner et al., 1983), is employed to delineate the boundaries of the employment core areas from the work places in the dense clusters. To be consistent with the zoning analysis in the next step, we projected the core areas to pre-defined spatial zones to find out the core zones by finding the best match by overlaying. An example of such spatial zones is provided in *Section 6.4.1* with the context in Paris Region and can be illustrated as in **Figure 6.1**.

6.2.3. Spatial Relation Analysis

The jobs-housing spatial relations are analyzed by the core-periphery patterns pertaining to the employment cores. The geographical zones are used as the base geographical units. For each zone, the residents' job locations over different employment cores are structured as the core-periphery profiles. A weighted topological graph is generated based on these profiles to represent the jobs-housing connections between the cores and other zones, which can be illustrated as **Figure 6.1** (such a graph is denoted as *jobs-housing network* hereafter). In the network, each zone is treated as a node and the home-to-work flow between zones is used to construct the edges. The edges are weighted by the flow "volume" that is the number of individuals having home and work places at the two ends of the edges. The zones within the same employment cores are joined together.

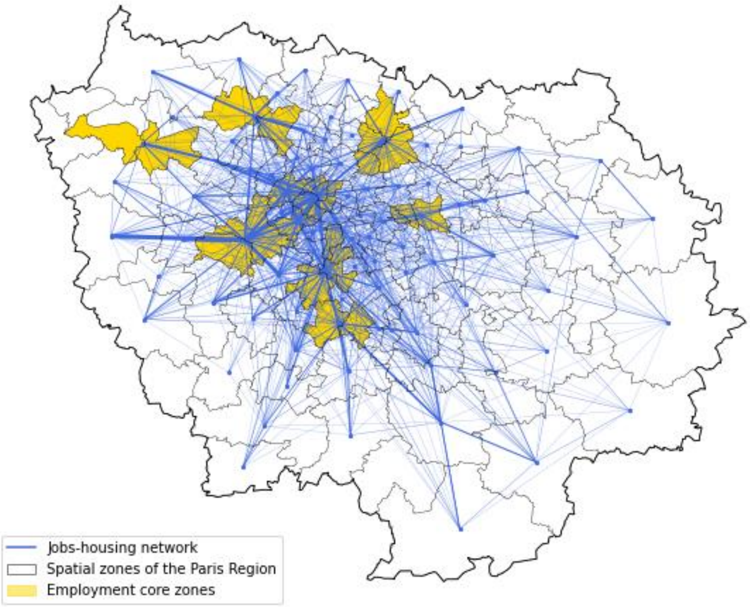


Figure 6.1 Jobs-housing network over different zones

Graph-based clustering methods are then applied to partition the jobs-housing network into densely connected sub-networks so as to find sub-regions with stronger inner-interactions. Such methods are termed as Community Detection Methods (CDM) in network science, where the detected sub-networks are called communities. By doing so, we can detect bonded zones as commuting communities and interpret the catchment areas of each employment core. Many algorithms were developed to achieve such community detections, among which the most widely used for large networks can be grouped into 2 major fashions. The 1st group is modularity-based algorithms that divide the network into communities by maximizing the modularity score, which measures the intra-communities connectivity vs inter-communities connectivity. The 2nd group is dynamic processes based on random walks, which measure the distances between nodes and finds densely connected parts as communities that a random walker tends to get trapped in. Other categories of algorithms also exist with their various searching schemes, including edge-removal, statistical inference etc. More details of method comparison can be referred to from Dao et al. (2018). In this study, the modularity-based algorithms were finally adopted due to the random walk algorithms required directed edges, which led to too many isolated nodes with no out-links in our jobs-housing network. We evaluated the most commonly used modularity optimization algorithms on our jobs-housing network and chose the results from the one achieving the optimist modularity score. Here is the list of algorithms we experimented: Fastgreedy, Multilevel, Louvain, Leiden and Leading Eigenvectors. The modularity score for evaluation can be calculated based on the formula proposed by (Newman & Girvan, 2004), defined as

$$modularity\ Q = \frac{1}{2m} * \sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\,\delta(i,j) \quad (6\text{-}1)$$

where m is the number of edges; A_{ij} is the element of the adjacency matrix A in its row *i* and column j; k_i, $k_j$denote the degree of node *i* and *j* respectively that is the number of edges connected to the node; and $\delta(i,j)$ is equal to 1 if node *i* and *j* belong to the same cluster, otherwise 0. The formula compares the observed intra-community edges number with an expected number obtained under a null model if we assume no priori correlation exists.

6.3. Application and Results

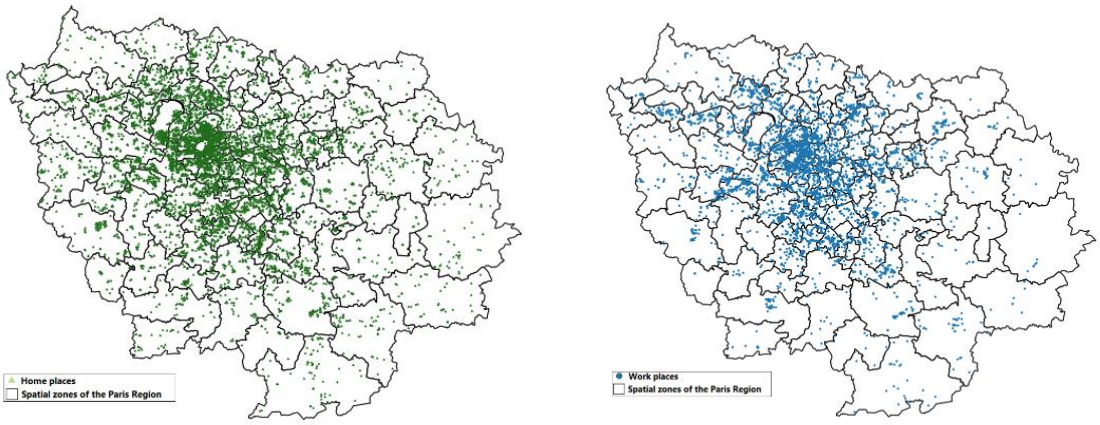
6.3.1. Settings

A case study was conducted using 14 days of FCD in the Paris Region (Île-de-France), which is the most populous region of France made up of the Paris City and 7 other departments. Trajectories with total logs of 15 million were collected in total. However, with the computation efficiency in mind, a sampling was performed for the analysis by selecting 10,000 vehicles with frequent usage out of the total of 42,596 qualified vehicles. The criterion of frequent usage was defined as showing up at least 7 distinct days during the period. Vehicles with erroneous records, such as sparse traces and abnormal recording frequency, were excluded from the data selection.

To conduct the spatial analysis, the Paris Region was divided into a set of zones, based on which the mobility features can be aggregated. Such spatial zones can be obtained by many principles, such as uniform grids, census blocks, and administrative boundaries, etc. In this study, we adopted the spatial tessellation data from a local authority, IAU (*L'Institut Paris Région*), which divides the Paris Region into 118 zones (as in Figure 1) by considering both of the administrative boundaries of French municipalities and the territory development coherence. The reasons for using such zones are twofold: the zoning system takes into account the population distribution by making finer division in dense areas and coarser division in outskirts; census data with other information are also available at each municipality level which enables direct comparison or data fusion.

6.3.2. Identified Home and Work Places

By trajectory sequencing and processing, a total of 44,882 frequently visiting places was recovered for the 10,000 vehicles with an average of 4.48 per vehicle. By analyzing the activity-holding profiles, home places and work places were extracted out from the pool of the places. The detailed patterns of identified clusters of home places and work places can be found in (D. Sun et al., 2021b) as well as an evaluation of their accuracy comparing to conventional rule-based detection. Consequently, 8,111 home places corresponding to 7,632 vehicles were identified as well as 4,295 work places corresponding to 4,156 vehicles. Among them, a total of 3,803 jobs-housing pairs were extracted with corresponding identical vehicle identifiers, based on which the jobs-housing network can be generated. The geo-locations of these places are visualized by scatter plots in **Figure 6.2**. It should be noted that such a portion of home and work place recognition was not high, especially with work place detection, but it is consistent with the fact that home place normally display more regular patterns on a vehicle usage basis while patterns of work places do not. A few vehicles were also found with more than just one home places or work places.


Home places
Spatial zones of the Paris Region
Work places
Spatial zones of the Paris Region

(a) Geo-locations of identified home places

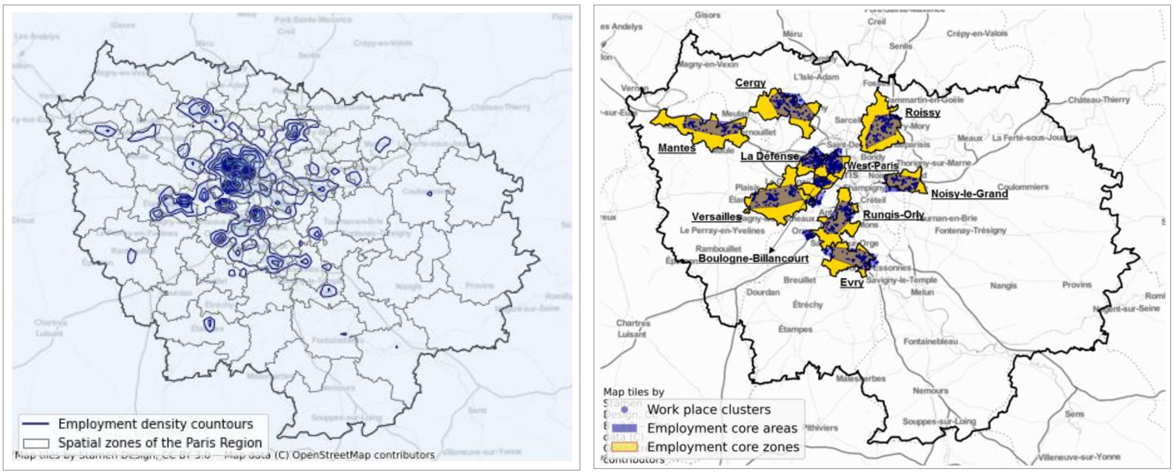
(b) Geo-locations of identified work places

Figure 6.2 Scatter plots of identified home and work places

6.3.3. Identified Employment Cores by Spatial Densities

With knowing the geo-locations of the work places, the KDE method was run to generate a density surface to describe the employment density distribution. Gaussian functions were chosen for the basic kernels. The optimal bandwidth was determined as 1000 meters via the grid-searching experiments. The estimated employment density contours are shown in **Figure 6.3(a)**, in which the darker fills within the contour lines the denser the distribution. It was found that the job locations were more densely distributed around the west areas in the Paris City, as well as other commonly acknowledged sub-centers in the Paris Region, such as La-Defense, Boulogne-Billancourt, Versailles, Roissy, Rungis-Orly, etc.

The HDBSCAN was then applied to further identify the employment cores. The most appropriate value for the parameter of *min_cluster_size* was determined as 50 after multiple experiments with the visual reference of the contours. As a result, 12 densely aggregated clusters were identified with 5 as major cores and the other 6 as sub-ones according to the clusters' point size. The core boundaries were delineated by applying the convex-hull algorithm and the core zones were projected according to the overlays, the results of which are illustrated in **Figure 6.3(b)**. Consequently, the 12 clusters were projected into 10 employment cores zones due to certain zones may cover two clusters. Likewise, zones corresponding to the same cluster were joined into one shape. The core zones were named and recalled after the major municipalities within the cores in the following text.



(a) Spatial density contours of work places

(b) Employment core areas and core zones

Figure 6.3 Spatial distribution of employment

6.3.4. Detected Spatial Communities by Jobs-Housing Relations

Recall that the communities were clusters of zones with denser intra-connections detected from the jobs-housing network. Two experimental settings were made when building the jobs-housing network and the results were compared for a discussion. The first setting (Setting 1) was to use the original jobs-housing network generated from the sample data, which were the 3,803 pairs of home and work places. However, considering the sample data may not be representative across different areas, the second setting (Setting 2) was based on a calibrated jobs-housing network that incorporated the census data to correct the biases. We did the calibration by making an expansion of the sample to the full population depending on the home locations of each zone. More specifically, the upscaling was according to the local auto-mobilist population of each zone recorded in the census data of *Recensement 2017* by Insee (the French Statistic Bureau). The distribution share of job locations (zones) was assumed to be the same as observed from the sample data. It should be borne in mind that certain population groups are still likely to be underrepresented and such inherent biases are expected to be further refined if better source data are available, such as mobile phone data with a high penetration rate or knowing the population breakdowns. Nevertheless, we utilized the two settings to demonstrate the methodological feasibility and compared the results between them for the potential differences before and after the expansion.

6 communities were detected under Setting 1 through the Multilevel CDM algorithm, which achieved the best modularity score in the experiments. In fact, the 5 evaluated algorithms all resulted in quite similar community partitions no matter what the modularity optimizing scheme is, which implicates that there is a robust community structure of jobs-housing network existing. The spatial organizations of the communities are shown in **Figure 6.4**, together with employment core zones displayed with highlighted emborders. The number of zones composing each community is also shown by bar plots in **Figure 6.4**. Zones with too few observations, whose node degree was less than 15, were excluded from the analysis. As can be seen from the results, the detected communities were generally formed of adjacent zones and the identified employment cores were embedded within the communities. Such results can be a confirmation of the effectiveness of the method as it is consistent with the nature of spatial cohesion that adjacent areas normally hold denser communication (Rinzivillo et al., 2012). The community range can also yield the major catchment areas of each employment core or a core agglomeration. When taking a closer look, we found that the core zones of West-Paris (City), La Défense, Boulogne-Billancourt, and Mantes were found bonded together, while the core zones of Roissy and Noisy-le-Grands were joined into the same community. The whole Paris City was actually divided into 3 major parts which joined to the corresponding communities in accordance with road network radiations.

The results of Setting 2 were shown in **Figure 6.5** for the map visualization and the community size respectively. The Multilevel algorithm was chosen for the detection, to make the comparison consistent with Setting 1. Setting 2 also resulted in 6 communities after the sample expansion. It turned out that the overall spatial distribution of the communities of the 2 settings was similar, but some differences can be spotted in the specific community constitutions. We can summarize the major changes as follows. Comparing to Setting 1, the employment cores of West-Paris, La-Défense, and Boulogne-Billancourt were no more bonded together (originally joined in the Community 1 (CM1) in Setting 1). Instead, the Boulogne-Billancourt Core joined with Versailles Core to form the community situated in the west of the Paris Region, while the West-Paris Core was connected with northeast areas to the Paris City to form a new community. Besides, the core zone of Roissy was separated with Noisy-le-Grands while the two cores in the northwest areas, Mantes and Cergy, were joined together. All these differences can be attributed to the weight change of different population groups in the sample expansion process. The result differences showed that the biases due to under-reported commuters did exist and might cause a different conclusion if without calibration.

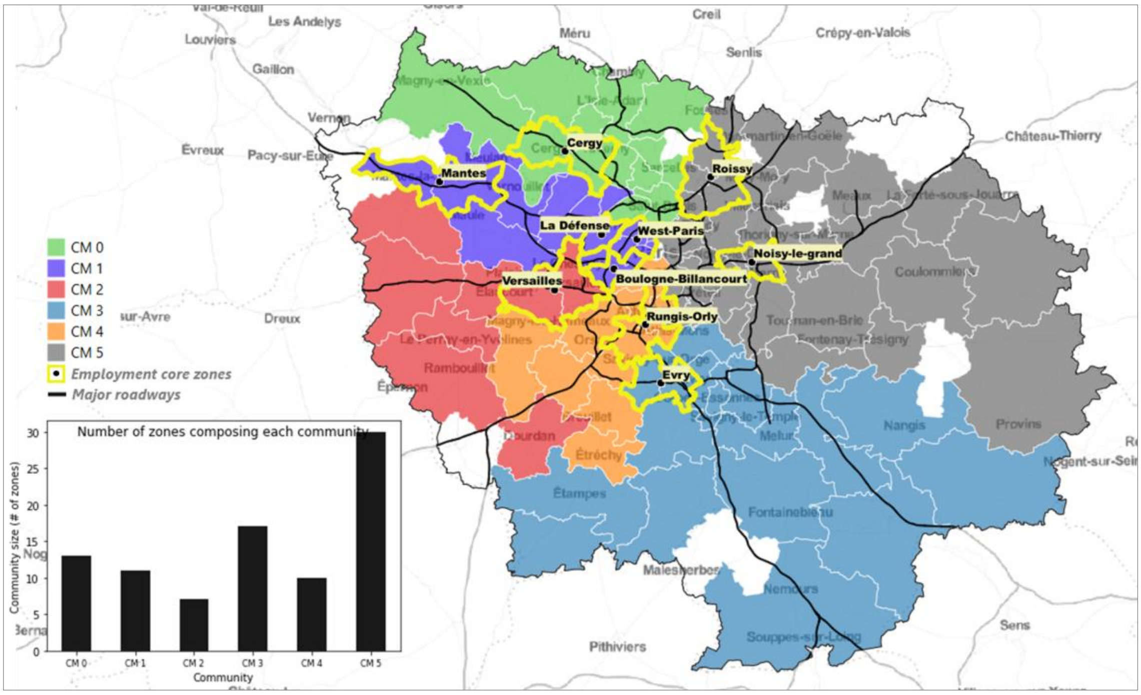


Figure 6.4 Community detection results of Setting 1 (original jobs-housing network based on FCD observations)

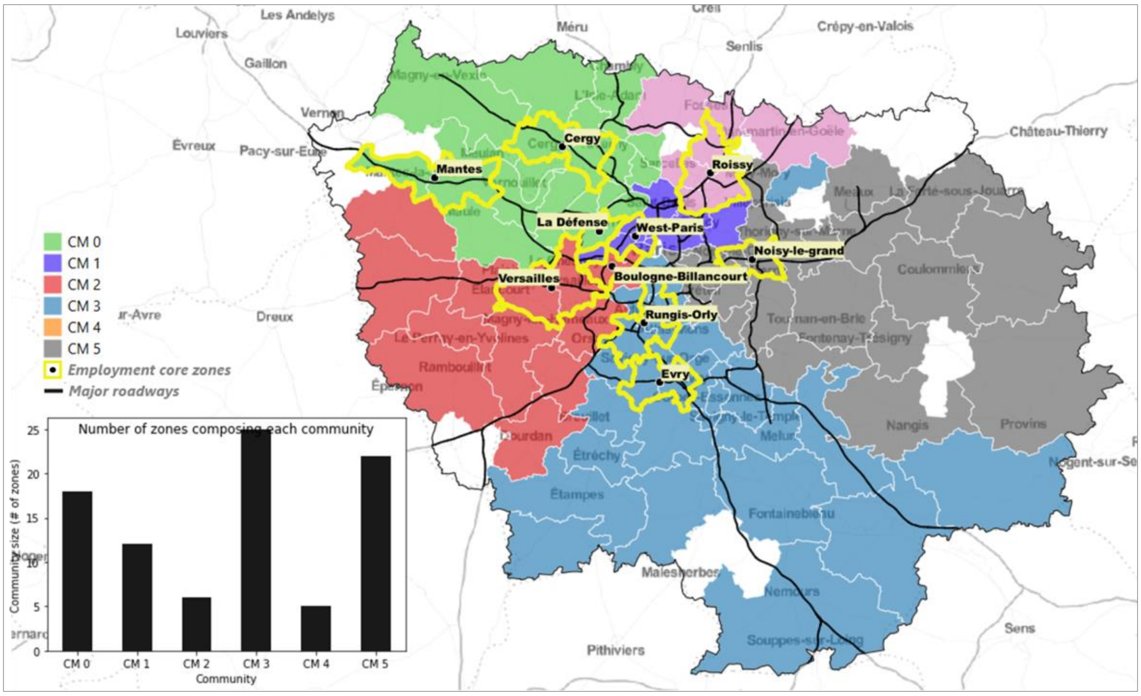


Figure 6.5 Community detection results of Setting 2 (calibrated jobs-housing network based on sample expansion)

6.4. Conclusion and Discussion

This paper presents a study on exploring the jobs-housing spatial relations by identifying the employment cores and their catchment areas based on vehicle trajectory data. The catchment areas were interpreted from the jobs-housing network by communities which were detected by finding sub-networks with a denser exchange internally and a sparser communication with the external areas. A case study was conducted to evaluate the methodology effectiveness using 14 days Floating Car Data in the Paris Region. 10 employment core zones were identified according to the spatial density. The jobs-housing network was built between these cores and other spatial zones. The community results were obtained under 2 different settings: the original network and the calibrated network, to compare for potential sampling biases across different areas. The 2 results were consistent on the overall spatial distribution of the detected communities, although the specific aggregations of some areas were different. The differences before and after the calibration can reflect that certain commuter groups were underrepresented in the sample data, which implicates further consideration in future data collection. However, both of the two results showed a good spatial adjacency of zones when forming communities as well as with employment cores embedded inside, although this is not a necessary property of the community detection method. Such a pattern can be proof of the effectiveness of the proposed method for its good capability in discovering densely connected sub-regions while maintaining spatial cohesion.

This study offers a referential practice for using modern trajectory data to analyze urban dynamics. The methodological framework holds the advantages of being data-driven, scalable for large-scale, and less dependent on prior expertise. The delineated jobs-housing communities can provide insights into the fundamental mobility structure of a territory and benefit future transportation planning and regional development cooperation etc. It also holds the potential to monitor up-to-date evolutions thanks to the ease of digital data collection. However, the findings of this work were based on vehicle commuting mobility only and a limited sample size. Future work may extend this study by incorporating other modes of transport and other purposes of traveling to obtain a more complete picture of human geography. The spatial relations can also be quantified to better understand the intensity as well as determine the influencing factors.

CHAPTER 7. ESTIMATING THE SPATIAL INTERACTIONS OF MOBILITY FLOWS BETWEEN ORIGINS AND DESTINATIONS

Adaptation of the article submitted for:
Leurent, F., Sun, D., & Xie, X. (2021). On Heterogenous Sampling Rates in Origin-Destination Matrix Estimation based on Trajectory Data and Link Counts. *Transportation Research Record.*

The candidate contributed to processing the data, performing the study and writing the manuscript.

Abstract

Being a fundamental metric of the transportation network, the origin-destination flow matrix is a critical input for various transportation models and studies. This paper deals with the estimation of an OD matrix of trip flows based on two kinds of data: probe trajectory data and local traffic counts. A Bayesian assignment framework is developed for demonstrating the relationship between the link probe sampling rates and the fractional contributions from the sampling rates on different OD pairs. The unknown OD matrix is estimated by applying cross entropy minimization using a prior matrix from the probe trajectories, along with the Bayesian assignment rules on link sample rates as the constraints. The methodology was applied using Floating Car Data and camera link flow counts for a numerical experiment. The results show that the method can achieve in a robust estimation of OD matrices, even using different prior matrices. The issue of the heterogenous sampling rates can be well addressed with link count constraints, effectively correcting the unknown bias in the probe sampling. The case study using real data also proves the feasibility of mining observed trajectory data to obtain the assignment fractions and estimate the OD matrix inversely, avoiding the conventional sophisticated process of traffic assignment modeling.

Keywords: Origin-Destination matrix estimation; Bayesian assignment; Heterogenous sampling rates; Probe vehicle trajectories; Link flow counts; Cross entropy minimization**.**

7.1. Introduction

As the primary purpose of a transportation network is to carry out mobile entities, the fundamental metrics of its activity is the matrix of entity flows between the places of origin and destination, in short, the OD trip flow matrix, per time period such as a day or a year.

The OD flow matrix is a prominent input in network traffic assignment models that simulate path choice and traffic conditions. It is also the outcome of travel demand models such as the trip generation and spatial distribution models in the classical 4-step model that derives local traffic flows from the spatial variables of land-use in the different places making up the territory under study. Among the models of trip distribution, let us mention the gravity model (Bhat et al., 1998; Bouchard & Pyers, 1965), the intervening opportunities models (Heanue & Pyers, 1966; Nazem et al., 2013), and discrete choice models (Uncles et al., 1987) of destination places that emphasize travel behavior in an economic perspective. There are as well statistical models purported to estimate or infer OD flows from different data sources, such as local traffic counts, en-route interview surveys to reveal the ODs composing some local flows, along with trajectory data at the trip level.

This paper deals with the estimation of an OD matrix of trip flows on the basis of two kinds of data: local traffic counts, on the one hand, and a trajectory dataset, on the other hand. The underlying traffic variables consist in link flows (observed by local counts) and path flows (aggregating trajectory data). Between the two types of flow variables, there is a fundamental relation: on any network link, the local link flow adds up the path flows of all the paths that use that link. When the path flows are based on trajectory data collected with some sampling rate, then the numbers of trajectories are multiplied by an expansion factor inversely proportional to the sample rate for comparison to link flows which are exhaustive (presumably). Yang et al., (2017) addressed the heterogeneity of such sampling rate by allowing for OD & time specific rates. They formulated a second kind of relationship involving the path and link flows with the heterogeneous sampling rates in original formula, linking the sample rate in the link flow to those in the OD flows multiplied by the ratio of sampled OD flows and the unknown OD flow (which thus appears in the denominator). Let us call this specific relationship “the composition of the local sampling rate” – in short LSR composition. Yet the LSR relation is presented in Yang et al., (2017) in a very concise way and it lacks a step-by-step demonstration, thereby making the resulting formula questionable. Our paper is aimed to provide a detailed demonstration, yielding outcomes that differ from Yang's formula, while keeping the spirit behind it.

In sum, the objective of this research is to estimate the OD flow matrix by using the probe trajectory data and the link flow counts. It aims to leverage the modern data sources in digital trajectory forms to ease the estimation of the fundamental traffic metrics while accounting for the problem of sampling biases of such data by using complementary information from flow counts at several links. Comparing to previous efforts, our methodology involves formulating clear statements of probability dependences on the relations of the sampling rates between the two data sources, contributing to providing the methodological reference for drawing the traffic equations. We use conditional probabilities to analyze the traffic variables and the observation protocols in an explicit, rigorous way. In this probabilistic setting, the LSR equation is obtained by reversing the conditionality between probabilistic events, in a typically Bayesian way. More so, we aim to provide a data-driven instance in mining observed trajectories to obtain OD matrix and flow assignment. A numerical experiment using real-world data is involved in this paper for demonstrating methodological feasibility and application simplicity.

7.2. Review of Related Work

The OD flow matrix estimation problem on the road network is the inverse problem of the Traffic Assignment (TA) problem (Bierlaire, 2002). Traditionally, an OD matrix is estimated thanks to travel surveys, flow observations on selected links, and socio-economic models, since lack of network-scale vehicle trip measurement (Cascetta et al., 2013). Recent technological advances permit to rethink the estimation problem. From 2000, the emerging sensing technologies that recorded large-scale vehicle trips on the road network promoted the development of new approaches to construct directly OD matrices from sampled trajectories to avoid the sophisticated process of trip generation and distribution (Bachir et al., 2019; Y. Cao et al., 2021; Gómez et al., 2015; Michau et al., 2019; Yang et al., 2017).

The application of emerging sensing technologies in the field of transport makes more and more large-scale trip geo-location data available. The most used sensing data that given sampled trajectories on the network included: Connected Vehicle (CV) trajectories and/or

Automatic Vehicle Identification (AVI) data (Y. Cao et al., 2021; Dixon & Rilett, 2002; J. Sun & Feng, 2011; X. Zhou & Mahmassani, 2006); GPS probe vehicle data (Yang et al., 2017; J. Yao & Chen, 2014); Floating Car Data (FCD) (Ásmundsdóttir, 2008; Gómez et al., 2015); and cell phone data (Bachir et al., 2019; Calabrese et al., 2011; Iqbal et al., 2014; Sohn & Kim, 2008). In recent years, CVs, such as vehicles of Uber and DiDi that are equipped with GPS units or drivers that use navigation services on their mobile phones, have emerged as a promising mobile data source because they can provide detailed and accurate vehicle geo-locations (Y. Cao et al., 2021). The AVI detectors that could uniquely identify each vehicle include Radio Frequency Identification device (RFID)-based detectors (Guo et al., 2019), Bluetooth detectors (Carpenter et al., 2012; Hainen et al., 2011; Michau et al., 2017, 2019), and license plate recognition (LPR) devices/cameras (Castillo et al., 2008; Rao et al., 2018). GPS probe vehicles are equipped with GPS for collecting data in real-time (Yang et al., 2017). There are two main types of FCD in operation today: cellular FCD data, derived from cellular networks, and global positioning system (GPS)-based FCD, derived from different types of devices that are equipped with a GPS receiver (Gómez et al., 2015). Mobile network data consist of two types (Bachir et al., 2019): Call Detail Records (CDR) including calls, text messages and sometimes Internet usage, named active events; and network records generated from an interaction between a device and the mobile network, named passive events.

For those typical datasets, in **Table 7.1**, we realize a comprehensive summary for the most recent and representative studies, in which the authors introduced their newest and the most complete studies based on previous studies in the literature to their knowledge. Different types of networks are revealed from the intersection scale to the metropolitan scale. A key concept in those studies was to generate the seed (prior) matrix directly from sampled trajectory data (Yang et al., 2017). In detail, the sophisticated process of trip generation and distribution was replaced by the map-matching and geo-location data processing algorithms, in which the parameter calibration and driver behavior assumptions were avoided.

An important issue, the heterogeneity of OD probe ratios among different OD pairs, was overlooked by most of the above studies, except in Yang et al., (2017) and Y. Cao et al., (2021). The authors (Yang et al., 2017) proposed a systematical way, not only to correct the potential bias caused by the three mains issues (1) the underspecified nature of OD estimation problem;

(2) reliability of the seed OD matrix; and (3) accuracy of assignment matrix estimation, but also to consider explicitly the heterogeneity of OD probe ratios. Later, the authors (Y. Cao et al., 2021) proposed more sophisticated penetration rates with heterogeneous sampling, whereas those of heterogeneous sampling were sampled from Gaussian distribution N(0.1, 0.022).

Although these estimators provided more accurate results than previous estimators, the optimization problems were sometimes no-convex or hard to solve numerically, e.g. given only local optimization. The application of Maximum Entropy (ME) approach seems necessary. Although the estimation formula of ME model are more complex, but the bias (the squared error) of ME estimator is much smaller than those of Generalized Least Squares, Maximum Likelihood or Bayesian Inference estimators and the problem of entropy maximization is convex and easy to solve numerically (Golan et al., 1996; Leurent, 1997).To this end, a ME estimator is proposed in this study based on our previous studies (Leurent, 1997), inspired the idea of OD correction models using trajectory data (Yang et al., 2017) and of EM model for sub-network (Xie et al., 2010). More so, this paper aims to provide a step-to-step derivation for the probabilistic formulation of LSR composition, along with a case study for a numerical experiment using real data instead of simulated data.

Table 7.1 A Comprehensive Summary about OD Matrix Estimation based on Typical Trajectory Data

Reference Elements	***Cao et al. (2021)***	***Bachir et al.(2019)***	***Michau et al. (2019)***	***Yang et al. (2017)***	***Gómez et al. (2015)***
Data	CV data + AVI data	Cell phone data + census data	AVI data +loop traffic counts	GPS probe data +link flow	FCD +loop traffic counts
Prior matrix	Robustly reconstruct prior OD flows via self-supervised learning.	-	A prior OD flows from Bluetooth data.	A prior OD flows from probe data.	-
Scaling	Temporal and spatial variation of the CV penetration rate is proposed and linear projection is extended to deal with variations in CV penetration rate.	Mobile phone flows are rescaled up to the total population with expansion factors using census data:	Homogeneous Bluetooth OD penetration ratio = 0.21.	Probe OD penetration ratio: Homogeneous = 0.15; Heterogeneous in [0.05, 0.3]	FCD penetration rates (FCD-PR): a priori, different levels [5%, 100%].
Routing systems	*Case (a):* Arterial network draft. *Case (b):* Simulation network by VISSIM. *Case (c):* Simulation network.	OSM road network, and rail network retried from the STIF Open Data platform.	OSM road network.	Simulated network.	Simulated network.
Network assignment	Data processing and Map matching.	Data processing and Transport mode inference (a two-steps semi-supervised model)..	Simulated.	Simulated	Simulated.
OD matrix estimation: approach	A self-supervised learning model called the Latency-Constrained Auto-Encoder (LCAE) is established to search for the optimal solution based on the estimated prior.	Three state-of-the-art steps: segmentation, origin-destination identification and rescaling.	Inverse problem of Argmin(*).	Generalized Least Squares (GLS)	A bi-level optimization using fuzzy logic theory
Final estimated OD matrix	*Case (a-c):* Node-to-node day-to Day dynamic OD matrices.	Ring-scale zone-to-zone day-to day dynamic OD matrix by transport mode (road and rail)	Node-to-node 3D network Link-OD matrix, giving information on the traffic assignment.	Zone-to-zone dynamic OD matrix	Static intersection OD matrix, and node-to-node static network OD matrix.
Purposes of OD matrix estimation	Dynamic origin–destination flow reveals the time-dependent travel demand on road networks. It serves as the fundamental input for dynamic traffic assignment (DTA) models as well as for network optimization programs.	Estimate, analyze and understand *daily urban mobility* to developpe supportive policies for decision making.	Origin-Destination matrix estimation is a *keystone for traffic representation and analysis of* traffic demand.	The origin-destination flow matrix is one *essential input to many dynamic traffic assignment and traffic simulation systems*.	Origin–destination matrices are of vital importance for transportation systems planning and design, as well as analysis, modelling, and simulation.

7.3. Bayesian Assignment: Theoretical Framework

This section presents a step-by-step formulation of the Bayesian assignment equations on the local sampling rate compositions. It starts with the fundamental principle of flow conservation and generalizes the application to any sub-population (i.e., FCD marked samples). By using Bayesian relations, the assignment equations are then derived to decompose the sampling rate at the link level with respect to the demand segment on the OD pairs. Such a relationship on the local sampling rates will then be used as the constraints to adjust the heterogeneous sample rate issues, with the adjustment approach stated in the next section.

7.3.1. *Notations*

a : index of network link, a unidirectional arc; A the set of network links.

q_a = number of trips going through a during some time period H.

r : index of network path, or route; R the set of network paths.

q_r = number of trips going through r during some time period H.

s : index of demand segment such as OD pair or another partition of the set of trips (i.e. by adding dimensions such as trip length range, or the time length of the activity at the destination, or the trip departure time).

q_s = trip flow inferred on s during H.

M_s = number of trajectories observed on s during H.

M_a = number of trajectories observed on a during H.

$M_{a,s}$ = number of trajectories observed on $a \cap s$ during H.

i : index marking a subpopulation of trips.

P$(a|i)$: Probability of using arc a conditionally to belonging to set i.

7.3.2. *Flow Conservation Principle*

For all networks of any kind, flow conservation at all nodes is a fundamental principle clearly stated by Kirchhoff's law in a local way between the link flows of the arcs headed to, or tailed at, any node. For transportation networks, the flow conservation principle is stated in a multi-scale fashion as a relation between local flow q_a on any arc a and the path flows q_r of all paths r at any geographic scale:

$$q_a = \sum_{r \in a} q_r. \qquad (7\text{-}1)$$

Condition $\{r \in a\}$ means the selection of paths r that go through a. Presumably, only elementary paths are considered, i.e. using any link at most once.

7.3.3. *Demand Segmentation*

The decomposition of network trips with respect to elementary network paths amounts to a segmentation of travel demands at the trip level. It was first derived by Beckmann et al. (1956). These authors also emphasized the segmentation of trip demand according to origin-destination pairs of places in the territory, i.e. another segmentation based on space but primarily oriented to places and pairs of places. Let us denote s an OD pair and $q_{a\cap s}$ the trip flow passing by a and belonging to s. It holds that:

$$q_a = \sum_{s\in S} q_{a\cap s}. \quad \text{(7-2)}$$

Denoting also $q_{r\cap s}$ the trip flow along path r and belonging to s, we have that:

$$q_r = \sum_{s\in S} q_{r\cap s}. \quad \text{(7-3)}$$

Notation $\{r \cap s\}$ may look superfluous as the OD pair associated to path r is unique and unambiguous. Yet hereafter we shall consider a generic segmentation of demand, still with index $s \in S$: for instance according to not only OD pair but also time of departure, or vehicle type: additional dimensions of analysis give rise to so-called “multiclass” network flow models.

We can also decompose the segment flow with respect to the paths that serve it:

$$q_s = \sum_r q_{r\cap s}. \quad \text{(7-4)}$$

When s only denotes an OD pair, formula (7-4) simplifies into,

$$q_s = \sum_{r\in s} q_r. \quad \text{(7-4')}$$

7.3.4. Probabilistic Setting

We are now ready to cast the previous relations of flow composition and demand segmentation in a probabilistic framework. The statistical population of interest is that of trips on the network during a certain period, say H. Let Q denote its overall size: by assumption, $Q > 0$.

For every subset $z \in \mathrm{A} \cup \mathrm{R} \cup \mathrm{S}$, the probability of belonging to that subset in the population of trips amounts to

$$\mathrm{P}(z) = q_z / Q. \quad \text{(7-5)}$$

For all $z \in \mathrm{A} \cup \mathrm{R}$ and $s \in \mathrm{S}$ it holds that

$$\mathrm{P}(z \cap s) = \mathrm{P}(z|s)\mathrm{P}(s), \qquad (7\text{-}6)$$

So that

$$q_{a\cap s} = \mathrm{P}(a|s)q_s, \ (7\text{-}6a)$$

$$q_{r\cap s} = \mathrm{P}(r|s)q_s. \ (7\text{-}6b)$$

As the demand segments make up a partition of the trip set, any subset z decomposes as the disjoint union of $\{z \cap s\}$ over all $s \in$ S. We can also restate (7-2) and (7-3) as

$$\mathrm{P}(a) = \sum_{s\in S} \mathrm{P}(a|s)\mathrm{P}(s). \quad (7\text{-}7a)$$

$$\mathrm{P}(r) = \sum_{s\in S} \mathrm{P}(r|s)\mathrm{P}(s). \quad (7\text{-}7b)$$

These relations have the flavor of stochastic assignment, yet at a more fundamental stage prior to any economic interpretation of $\mathrm{P}(r|s)$ or $\mathrm{P}(a|s)$ as the outcome of a discrete choice model of route choice for the trip-makers in segment s.

7.3.5. Sub-population

The above relations pertain to a population of trips on the network. They hold as well for any sub-population of trips: let us use an index i to denote the belonging to such sub-population. Let also M denote its size, $M_a = \mathrm{P}(a|i)M$ the trip flow of that category on link a, $M_r = \mathrm{P}(r|i)M$ that on route r, $M_s = \mathrm{P}(s|i)M$ that on demand segment s. As the set S of segments constitutes a partition of set i as well as of the full population of trips, it holds that

$$\mathrm{P}(a|i) = \sum_{s\in S} \mathrm{P}(a \cap s|i), \quad (7\text{-}8a)$$

$$\mathrm{P}(r|i) = \sum_{s\in S} \mathrm{P}(r \cap s|i). \quad (7\text{-}8b)$$

The sub-population can be thought of as extracted from the general trip population and marked in some specific way – e.g. “marked trips” made by marked vehicles.

7.3.6. Conditionality and Bayesian Relations

For $z \in$ A ∪ R and $s \in$ S we may consider the interplay of z, s and i in conditional probabilities:

$$\mathrm{P}(z \cap s \cap i) = \mathrm{P}(s, i|z)\mathrm{P}(z), \quad \text{(7-10a)}$$

$$= \mathrm{P}(z, i|s)\mathrm{P}(s), \quad \text{(7-10b)}$$

$$= \mathrm{P}(z, s|i)\mathrm{P}(i), \quad \text{(7-10c)}$$

$$= \mathrm{P}(z|i, s)\mathrm{P}(i, s), \text{(7-10d)}$$

$$= \mathrm{P}(s|i, z)\mathrm{P}(i, z), \text{(7-10e)}$$

$$= \mathrm{P}(i|z, s)\mathrm{P}(z, s). \text{(7-10f)}$$

From (7-10d) and (7-10a) we get that

$$\mathrm{P}(z|i, s) = \mathrm{P}(s, i|z)\frac{\mathrm{P}(z)}{\mathrm{P}(i,s)}. \quad \text{(7-11a)}$$

Then, from (7-10b) and (7-10d), it holds that $\mathrm{P}(z, i|s)\mathrm{P}(s) = \mathrm{P}(z|i, s)\mathrm{P}(i, s)$, hence that

$$\mathrm{P}(z, i|s) = \mathrm{P}(z|i, s)\mathrm{P}(i|s). \text{ (7-11b)}$$

Next, from (7-10f) and (7-10a), it holds that $\mathrm{P}(i|z, s)\mathrm{P}(z|s)\mathrm{P}(s) = \mathrm{P}(s, i|z)\mathrm{P}(z)$, hence that

$$\mathrm{P}(z|s) = \frac{\mathrm{P}(s, i|z).\mathrm{P}(z)}{\mathrm{P}(i|z, s).\mathrm{P}(s)} = \frac{\mathrm{P}(i, z|s)}{\mathrm{P}(i|z, s)}. \text{(7-11c)}$$

The different combinations in system (7-10a – 7-10f) enable us to derive a number of such relationships which are typical of Bayesian analysis.

7.3.7. Orientation-Marking Independence

When trip marks come from on-board boxes including GPS geolocation with the provision of route planning and guidance, then marking would seem to be correlated to orientation. However, as nowadays most people have smartphones and can avail themselves of dynamic traffic information (DTI) together with route planning services, all the flow would be influenced by DTI so that subscription to the marked service could still be considered as independent of the orientation. Moreover, at the local level, the use of spefic roadways is less likely to be affected by the socio-demographic background, as all users are assumed to opt the routes with better conditions.

Under this condition, we propose the Postulate "OMI" (orientation-marking independence): if marking i (not necessarily to be random sampled) and orientation z are statistically independent, then for any demand segment s it holds that

$$\mathrm{P}(z|i,s) = \mathrm{P}(z|s). \qquad \text{(7-12)}$$

Thus, the orientation coefficients at the link or route level are the same in the marked sub-population as in the full population. Combining (7-11a) and (7-12), we get that

$$\mathrm{P}(z|s) = \frac{\mathrm{P}(s,i|z).\mathrm{P}(z)}{\mathrm{P}(i|s).\mathrm{P}(s)}, \text{ and in turn}$$

$$\mathrm{P}(s,i|z) = \frac{\mathrm{P}(z|s).\mathrm{P}(i|s).\mathrm{P}(s)}{\mathrm{P}(z)}, \text{ so that}$$

$$\mathrm{P}(s,i|z) = \mathrm{P}(i|s).\mathrm{P}(s|z). \quad \text{(7-13)}$$

7.3.8. Flow Observation

7.3.8.1. Probe Data

From now on we shall postulate that the marked trips are observed along all of their paths: i.e. the trip flows M, M_r, M_a, M_s, $M_{s\cap a}$ etc. are given. From these, it stems the probabilities conditionally to i. Notably, in the marked population the flow composition at the link level is

$$\mathrm{P}(s|a,i) = \frac{M_{s\cap a}}{M_a}, \ \forall s \in S, \ \forall a \in \mathrm{A}. \quad \text{(7-14a)}$$

Conversely, at the segment level the link assignment proportions are obtained as

$$\mathrm{P}(a|s,i) = \frac{M_{s\cap a}}{M_s}, \ \forall s \in S, \ \forall a \in \mathrm{A}. \quad \text{(7-14b)}$$

Under the OMI postulate, (7-14b) yields that

$$\mathrm{P}(a|s) = \frac{M_{s\cap a}}{M_s}, \ \forall s \in S, \ \forall a \in \mathrm{A}. \qquad \text{(7-14c)}$$

Combining (7-14b) to (7-11b) applied to $z = a$, we recover that

$$\mathrm{P}(a,i|s) = \frac{M_{s\cap a}}{M_s}\frac{M_s}{q_s} = \frac{M_{s\cap a}}{q_s}. \quad (7\text{-}14\text{d})$$

7.3.8.2. Link Counts

Let us also assume that the local flow is observed (counted) on a selection A_i of links a, with outcomes x_a, so that

$$q_a = x_a, \ \forall a \in \mathrm{A}_i. \qquad (7\text{-}15)$$

From the joint observation of marked trips and link counts, we can estimate

$$\mathrm{P}(i|a) = \frac{M_a}{x_a}, \ \forall a \in \mathrm{A}_i. (7\text{-}16\text{a})$$

$$\mathrm{P}(i,s|a) = \frac{M_{s\cap a}}{x_a}, \ \forall s \in S, \ \forall a \in \mathrm{A}_i. \ (7\text{-}16\text{b})$$

More on assignment proportions. On applying (7-11c) to $z = a$ and using (7-16b) we obtain

$$\mathrm{P}(a|s) = \frac{\mathrm{P}(s,i|a).\mathrm{P}(a)}{\mathrm{P}(i|a,s).\mathrm{P}(s)} = \frac{M_{s\cap a}}{x_a}.\frac{1}{\mathrm{P}(i|a,s)}.\frac{q_a}{q_s}. \qquad (7\text{-}17\text{a})$$

Now, comparing (7-17a) to (7-14b),

$$\frac{\mathrm{P}(a|s,i)}{\mathrm{P}(a|s)} = \frac{x_a}{M_s}.\frac{q_s}{q_a}.\mathrm{P}(i|a,s). \quad (7\text{-}17\text{b})$$

Thus, the OMI postulate implies that $\mathrm{P}(i|a,s) = \frac{M_s}{q_s}\frac{q_a}{x_a}$: under (7-15) it reduces to $\mathrm{P}(i|s)$.

7.3.8.3. *On Flow Composition*

Furthermore, under OMI formula (7-13) applied to $z = a$ enables us to recover from (7-16b) that

$$\mathrm{P}(s|a) = \frac{\mathrm{P}(s,i|a)}{\mathrm{P}(i|s)} = \frac{M_{s\cap a}}{x_a}\frac{q_s}{M_s}. \ (7\text{-}17\text{c})$$

Comparing (7-17c) to (7-14a), we get that

$$\frac{\mathrm{P}(s|a,i)}{\mathrm{P}(s|a)} = \frac{x_a}{M_a}\frac{M_s}{q_s}. \qquad (7\text{-}17\text{d})$$

7.3.8.4. Exogenous Sampling Rates.

A further specification would be to take the marking rate as given for every demand segment:

$$\mathrm{P}(i|s) = \frac{M_s}{q_s} = \theta_s. \text{ (7-18a)}$$

Then it would be easy to recover the segment trip flow as

$$q_s = \frac{M_s}{\theta_s}. \text{ (7-18b)}$$

7.3.9. Bayesian Assignment Rule

A key element in Yang et al (2017) is to decompose the sampling rate at the link level with respect to the demand segments. Under our notations,

$$\mathrm{P}(i|a) = \textstyle\sum_{s\in S} \mathrm{P}(i \cap s|a), \text{ so that}$$

$$\mathrm{P}(i|a) = \textstyle\sum_{s\in S} \mathrm{P}(i|s,a).\mathrm{P}(s|a). \qquad \text{(7-19)}$$

Under the OMI postulate, $\mathrm{P}(i|s,a) = \mathrm{P}(i|s)$ so that (7-19) becomes

$$\mathrm{P}(i|a) = \textstyle\sum_{s\in S} \mathrm{P}(i|s).\mathrm{P}(s|a). \qquad \text{(7-20)}$$

We may estimate $\mathrm{P}(s|a)$ by $\mathrm{P}(s|a,i)$, so that

$$\mathrm{P}(i|a) \approx \textstyle\sum_{s\in S} \mathrm{P}(i|s).\mathrm{P}(s|a,i). \qquad \text{(7-21)}$$

Then, replacing $\mathrm{P}(s|a,i)$ with $M_{s\cap a}/M_a$, $\mathrm{P}(i|s)$ with M_s/q_s and $\mathrm{P}(i|a)$ with M_a/x_a, it comes out that

$$\frac{M_a}{x_a} \approx \textstyle\sum_{s\in S} \frac{M_{s\cap a}}{M_a} \frac{M_s}{q_s}. \qquad \text{(7-22)}$$

The outreach of equation (7-22) is to make matrix demand estimation sensitive to the heterogeneity of marking rates. Yet the sensitivity was achieved by replacing $\mathrm{P}(s|a,i)$ with $\mathrm{P}(s|a)$. The substitution is questionable because without making it, it is straightforward to see that (7-22) could be reduced to $1 = 1$ i.e. carrying no additional information.

Let us focus on expansion factors instead of sampling rates. The heterogeneity of sampling rates induces that of expansion coefficients: $F_a := q_a/M_a$ at the link level and $F_s := q_s/M_s$ at the segment level.

For links that are subjected to traffic counts, we may require the average link expansion factor, $F_a = q_a/M_a$, to be equal to the average of the respective expansion factors of the constitutive segments: i.e.

$$F_a = \sum_{s\in S} \mathrm{P}(s|a,i).F_s. \qquad (7\text{-}23)$$

After substitution, it comes out that

$$\frac{q_a}{M_a} = \sum_{s\in S} \frac{M_{s\cap a}}{M_a}\frac{q_s}{M_s}. \qquad (7\text{-}24)$$

This relation is essentially a formula to predict link flow q_a on the basis of the segment flows q_s together with the marked flows $M_{s\cap a}$ and M_s.

7.4. Application Methodology

7.4.1. Practical Conditions

A case study was conducted to experiment the proposed theories in the area of Saint-Cyr-l'École, a French commune (municipality) situated in the southwest of the Great Paris Region. Three main datasets were used for the numerical analysis, including 1) 14-days Floating Car Data covering all trajectories concerning the studied area; 2) a dataset of camera link flow counts at 11 link locations, which counts the traffic flow passing the observing links (different driving directions were differentiated as different links); and 3) the road network of the area with each link indexed. The studied time period was from November 31to December 13th in the year of 2020. The study area was divided into 6 traffic analysis zones for the Origin-Destination (OD) flow analysis, as shown in **Figure 7.1**. Such a zonal division was actually formed to correspond to a particular interest to the local authority in monitoring the traffic on the arterial road of the area (as highlighted in **Figure 7.1**), which was prone to congestion and needed a shift of its demand allocation to other roads. The case study was set to account for that specific traffic problem, while maintaining the generality of the methodology. This also explains why the cameras were installed mainly along the specific itinerary, restricted by the practical condition and affordability of a municipality. It should be noted the major aim of this case study was to serve as a pilot experiment for demonstrating the methodological feasibility using real-world data. One should bear in mind that the location of link counts collection and the segmentation of time of day would notably have an influence on the performance of the OD estimation. Such investigations on those practical settings or the findings of optimal link count locations would therefore be limitations of this study and prompt further works.

FCD dataset was structured as a sequential set of GPS trace points with vehicle geo-locations, time stamps and vehicle identifiers (anonymized). Such trajectory data were firstly transferred into trip datasets by segmenting the trajectories into parts. As the FCD collect the entire path information, after the segmentation, the OD flow, their assignment paths and the partial flow pertaining to the marked vehicles by the FCD can be easily recovered. More specifically, the segmentation was done by identifying activity stops (D. Sun et al., 2021b), whose patterns can be featured as staying at a place over a certain time threshold. To identify the path, a map-matching approach (Newson & Krumm, 2009) was then employed to match the trajectory to the road network. Extending trips were clipped at the area border with its origin/destination replaced by the intersecting point at the border for determining their belonging zones. Therefore, only the trips traversing the studied area and relevant to the use of internal roads were considered.

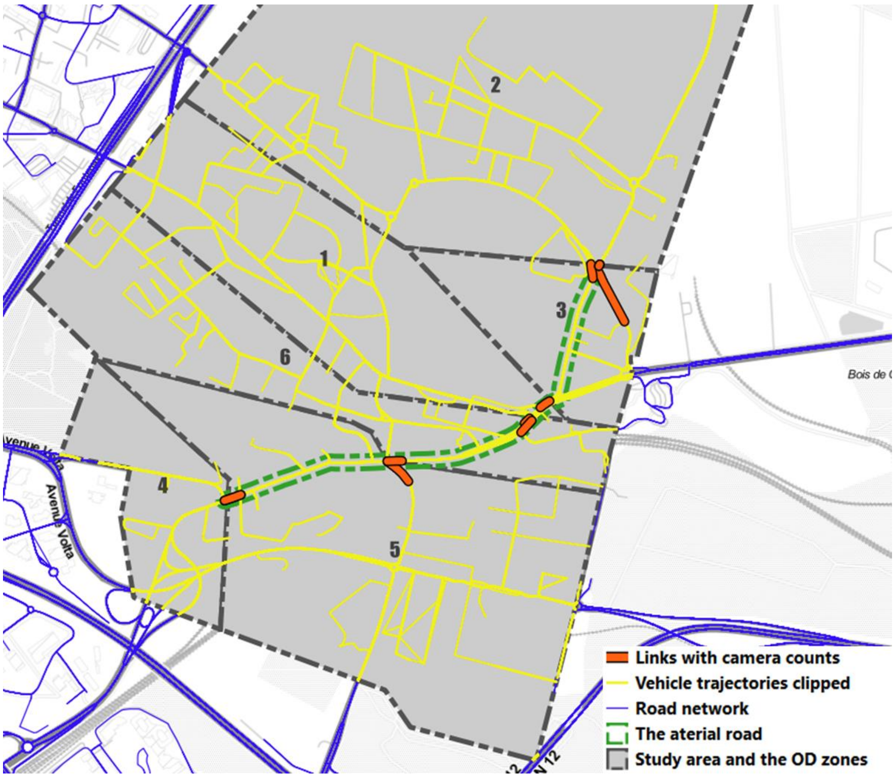


Figure 7.1 Studied location, OD zoning, FCD trajectories, and links with traffic counts

7.4.2. Typical Application Scheme

The main application scheme is to derive an ex-ante OD matrix from the probe trajectories and then use it to estimate the ex-post OD matrix by adjustment based on the relations of expansion factors between link counts and OD segment flow (OD pair flow in this case) derived from the above Bayesian rules. The adjustment process follows a cross entropy minimization problem, which can be formulized as equation (7-25). The optimal solutions can be obtained by introducing a dual variable with the Lagrangian function and then solving the dual function with the Newton-Raphson method. Thus, if given ex-ante flows, we can easily estimate the ex-post flows by solving the minimizing problem with the constraints.

$$\min \mathrm{C}(\boldsymbol{q}) := \sum_{s \in S} q_s \left(\ln \left(\frac{q_s}{v_s} \right) - 1 \right) \qquad (7\text{-}25)$$

Subject to

$$q_s \geq 0 \ : \ \forall s \in S, \ (7\text{-}26)$$

$$x_a = \sum_{s\in S} m'_{a,s} q_s \quad : \ \forall a \in \mathrm{A}_i \qquad (7\text{-}27)$$

Where q_s denotes the ex-post flow, v_s denotes the ex-ante flow and $m'_{a,s} := M_{s\cap a}/M_s$. Here, v_s, $M_{s\cap a}$, and M_s can be directly computed from the trajectory observations. Different specifications of v_s can be used, as proposed in the next section.

7.4.3. Ex-ante Segment Flows and Sampling Rates

In the problem of cross entropy minimization under constraints, the objective function is a metric of the difference between the current distribution and the prior one (ex-ante flow). The optimization consists in finding the smallest adaptation of the prior distribution (i.e., to change it in the smallest way) so as to meet the constraints. Thus, the prior vector is a major factor of the final outcome. We introduce 3 alternative specifications (*Sp*) for a prior set of segment trip flows for the estimation

1) Probe-based (*Sp1*): From the probe data, we can get the number M_s of trips made in OD segment s. It may constitute a prior estimate of the flow – let us denote it as:

$$v_s^{(1)} := M_s \qquad (7\text{-}28)$$

2) Using probes and average sample rate in link counts (*Sp2*): Let us consider an average sample rate (θ_i) over the different link counts:

$$\theta_i := \frac{1}{|\mathrm{A}_i|} \sum_{a\in \mathrm{A}_i} \frac{M_a}{x_a} \qquad (7\text{-}29)$$

Combining this rate to the number M_s of trips made in segment s, we obtain a second prior estimate of the flow – let us denote it as

$$v_s^{(2)} := M_s/\theta_i \quad (7\text{-}30)$$

3) Using probes and segment specific sample rate from link counts (*Sp3*): Using link counts together with probe data, we may infer a segment specific sample rate (θ_s) by considering the level of involvement of the segment in each link count, then:

$$\theta_s := \frac{1}{\sum_{a\in \mathrm{A}_i} M_{s\cap a}} \sum_{a\in \mathrm{A}_i} M_{s\cap a} \frac{M_a}{x_a} \qquad (7\text{-}31)$$

Combining this rate to the number M_s of trips made in segment s, we obtain a third prior estimate of the flow – let us denote it as:

$$v_s^{(3)} := M_s/\theta_s \quad (7\text{-}32)$$

7.4.4. Comparison Criteria

Due to the lack of ground truth of the OD flows, which is a common issue of using field collected probe data rather simulated data, we propose the following 3 criteria to measure the distribution of OD flows and make the comparison between the results of different specifications

- *Information Entropy*: This metric is used to measure the disorder of the flow distribution of an OD matrix, equivalently, measuring the lack of informative structure of an OD matrix**.**

$$H(p) = -\sum_{s\in S} p_s \ln(p_s) \qquad (7\text{-}33)$$

where $p_s = q_s/N_q$

- *Mean Absolute Gap (MAG):*This metric is to measure the difference between two OD matrices (e.g. ex-ante flow q_s^f vs ex-post flow $q_s^{f'}$).

$$MAG = \sum_{s\in S} \frac{|q_s^f - q_s^{f'}|}{N_s} \qquad (7\text{-}34)$$

- *Mean Relative Gap (MRG):* This metric is to measure the relative difference between two OD matrices (e.g. ex-ante flow q_s^f vs ex-post flow $q_s^{f'}$).

$$MRG = \sum_{s\in S} \frac{|q_s^f - q_s^{f'}|/\, q_s^f}{N_s} \qquad (7\text{-}35)$$

7.5. Numerical Results

7.5.1. Case Study: Ex-Ante Analysis

After processing, 54,728 trips in total were recovered from the raw trajectories, in which 5,089 trips were observed passing the camera links, corresponding to 1,675 vehicles. We only used this subset for the OD flow estimation as the objective was to estimate the local circulating OD flow concerning the links with camera installed, as they compose the major arterial of the commune with main accesses and exits to the external region. The ex-ante OD flows $v_s^{(1)}$, $v_s^{(2)}, v_s^{(3)}$ were then derived under three alternative specifications. The distribution of them on each OD pair is shown in **Figure 7.2**. As $v_s^{(2)}$ was upscaled from $v_s^{(1)}$ using an average sampling rate, the flows of them were under a same distribution. However, the difference between $v_s^{(2)}$ and $v_s^{(3)}$ showed that the upscaling from observations was uneven among different OD pairs, if using link counts information on calculating segment sampling rates, indicating the heterogeneity of probe penetration rates (sampling rates).

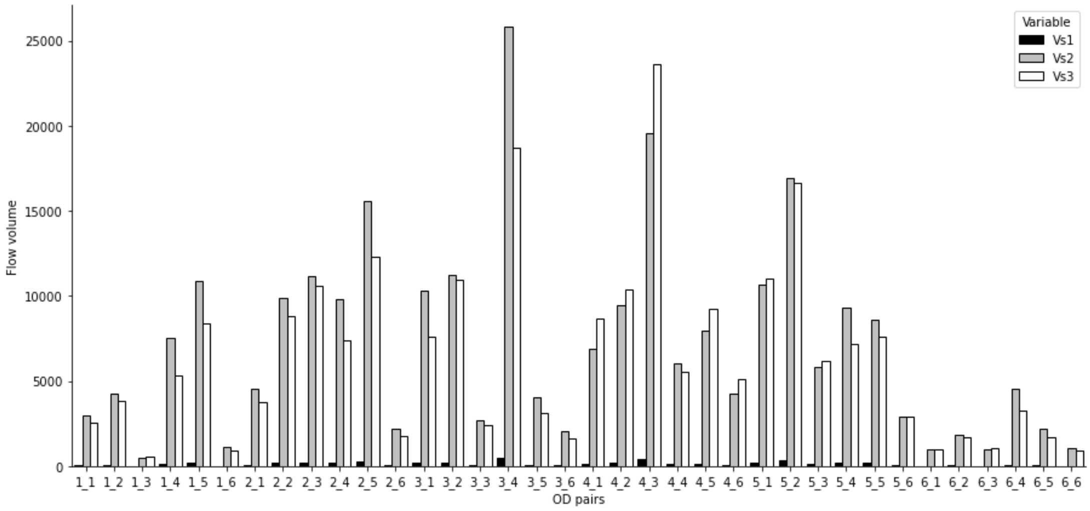


Figure 7.2 Ex-ante OD flows under 3 different specifications

7.5.2. Case Study: Ex-Post Analysis

The ex-post OD flows $q_s^{(1)}, q_s^{(2)}, q_s^{(3)}$ were estimated under the 3 different specifications of ex-ante flows of $v_s^{(1)}, v_s^{(2)}, v_s^{(3)}$ respectively. **Figure 7.3** compares the results between them on each OD pair. Comparing to ex-ante distributions, the ex-post distributions were found more consistent with each other with less difference between the 3 specifications, which can be an indication of the robustness of the adjustment approach by cross entropy minimization with link constraints. **Figure 7.4** shows the ratios of the ex-post flows to the observed probe flows on each OD pair, which actually compares the inverse of the sampling rates, but based on adjusted ex-post flows. Such ratios were found varying significantly among different OD pairs, which implicates that the penetration of probe vehicles data exists in a quite heterogenous way among different OD pairs. Therefore, to obtain a representative OD estimation from probe trajectory data, careful adjustment accounting for such a sampling bias issue is required.

The ex-post OD flows $q_s^{(1)}$, $q_s^{(2)}, q_s^{(3)}$ were estimated under the 3 different specifications of ex-ante flows of $v_s^{(1)}$, $v_s^{(2)}, v_s^{(3)}$ respectively. **Figure 7.3** shows the estimated OD flows ($q_s^{(1)}$, $q_s^{(2)}, q_s^{(3)}$) with respect to each OD pair. Comparing to ex-ante flows ($v_s^{(1)}$, $v_s^{(2)}, v_s^{(3)}$) displayed in **Figure 7.2**, the ex-post flows were more consistent among the 3 specifications, which can be an indication of the robustness of the adjustment approach. **Figure 7.4** shows the ratios of the ex-post flows to the observed probe flows $(\frac{q_s^{(1)}}{ms}, \frac{q_s^{(2)}}{ms}, \frac{q_s^{(3)}}{ms})$ on each OD pair. These ratios actually reflect the inverse of the sampling rates, but are based on adjusted ex-post flows. As can be seen in **Figure 7.4**, such ratios were found varying significantly among different OD pairs, which implicates that the penetration of probe vehicles data exists in a quite heterogeneous way among OD pairs. Therefore, to obtain a representative OD estimation from probe trajectory data, careful adjustment accounting for such a sampling bias issue is required.

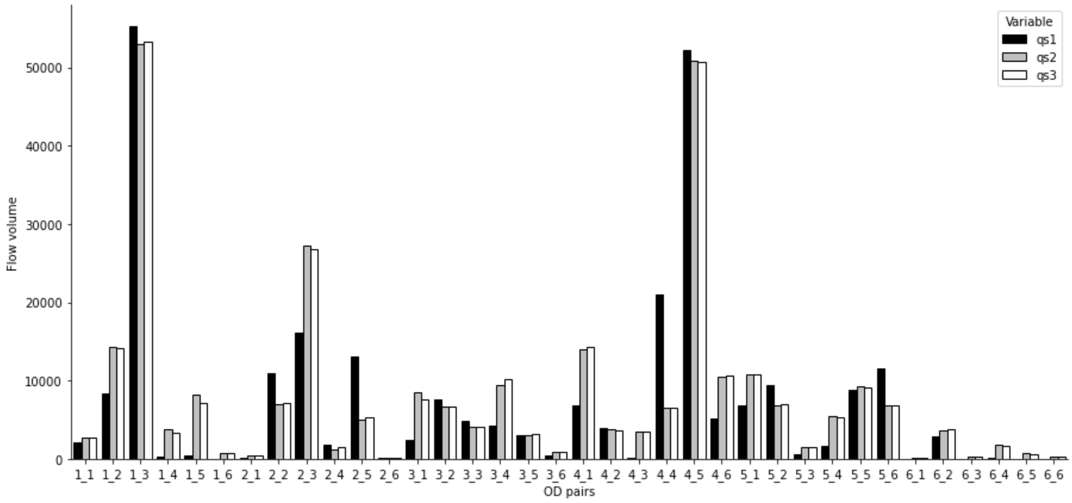


Figure 7.3 Estimated ex-post OD flows under 3 different specifications

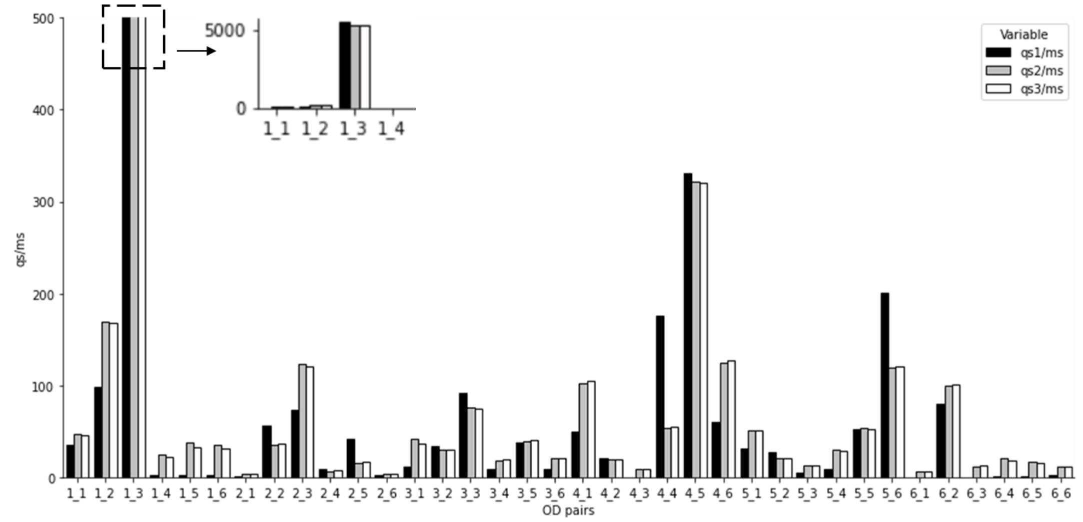


Figure 7.4 Ratios of ex-post flows to probe flows

The flow distribution entropy (H(p)) was measured for each ex-ante (prior) flow matrix and ex-post flow matrix of the 3 specifications (Sp1, Sp2, Sp3), the results of which are summarized in **Table 7.2**. It shows an entropy relation of $H(Sp1\ prior)$ = $H(Sp2\ prior)$ > $H(Sp3\ prior)$, implicating that upscaling with the specific sample rate can bring up more information in the structure of the prior flows. For all specifications, the $H(piror) > H(ex - poste)$, indicating an effective disorder reduction (information gain) by the proposed adjustment method. The entropy reduction was also calculated as shown and was found that $\Delta H(sp1) > \Delta H(sp2) > \Delta H(sp3)$, suggesting that using the link constraints can bring more information gain to a non-informative prior than to an informative prior that is already adjusted with sampling rates based on information from link counts.

Table 7.2 Entropy Measurement of Flow Distribution

Prior matrix setting	Ex-ante matrix entropy	Ex-post matrix entropy	Entropy difference (prior – ex-post)
Specification 1 (Sp1)	3.286	2.671	0.615
Specification 2 (Sp2)	3.286	2.898	0.388
Specification 3 (Sp3)	3.273	2.892	0.381

Table 7.3 compares the MAG between the OD matrices of each paired ex-ante matrix and ex-post matrix, where rows and columns are the 2 matrices compared and the cell is the metric value. While **Table 7.4** displays the MRG for a same kind of comparison. The discrepancy was found large between the ex-ante flows and the ex-post flows, especially for the specification 1 ex-ante flow (probe observations only) to the others, indicating again the adjustment was quite

crucial for estimating a representative OD matrix with heterogenous probe sampling rates. Although the choice of prior matrix would have effects on the ex-post estimation, the matrices after adjustment were still more reliable than simply using the any of the priors as the OD matrices. Comparing the ex-post flows, the 3 matrices were close to each other, but especially between Sp2 and Sp3, suggesting that using informative priors would contribute to a more robust estimation of the ex-post flow.

Table 7.3 Mean Absolute Gap of OD flows between Different Ex-ante and Ex-post Matrices

MAG	Sp1 prior-	Sp2 prior-	Sp3 prior-	Sp1 post-	Sp2 post-	Sp3 post-
Sp1 prior-	0.0	6994.2	6370.0	7168.0	8009.2	7963.3
Sp2 prior-	6994.2	0.0	1161.9	7452.0	6944.8	6977.5
Sp3 prior-	6370.0	1161.9	0.0	6971.9	6322.5	6308.1
Sp1 post-	7168.0	7452.0	6971.9	0.0	3069.3	2983.3
Sp2 post-	8009.2	6944.8	6322.5	3069.3	0.0	196.1
Sp3 post-	7963.3	6977.5	6308.1	2983.3	196.1	0.0

Table 7.4 Mean Relative Gap of OD Flows between Different Ex-ante and Ex-post Matrices

MRG	Sp1 prior-	Sp2 prior-	Sp3 prior-	Sp1 post-	Sp2 post-	Sp3 post-
Sp1 prior-	0.000	49.477	45.054	196.069	196.874	196.985
Sp2 prior-	0.980	0.000	0.151	3.904	3.680	3.700
Sp3 prior-	0.978	0.179	0.000	3.703	3.432	3.448
Sp1 post-	0.970	22.585	21.522	0.000	5.692	5.460
Sp2 post-	0.946	2.050	1.731	0.595	0.000	0.043
Sp3 post-	0.946	2.069	1.744	0.582	0.044	0.000

7.6. Conclusion and Discussion

This paper presents a study dealing with OD matrix estimation using probe trajectory data and link flow counts. A step-to-step Bayesian LSR formulation was derived and demonstrated for the relationship between the link sampling rates and the assignment fractions from different OD pairs. Using such a relationship, a cross entropy minimization adjustment was applied to estimate the unknown OD matrix using a prior matrix. Different specifications were proposed for the prior matrix.

A case study using real-world FCD and camera link flow counts was conducted for a numerical experiment. It is shown that the proposed framework can achieve in a robust estimation of the OD flows from sampled trajectory data. The issue of the heterogeneous sampling rates can be well addressed with link count constraints, effectively correcting the unknown bias in the probe sampling. It was also found that using an informative prior matrix using link counts to calculate OD pair specific sample rates would contribute to a more reliable estimation.

The major contributions of this study are twofold. Methodologically, we derive the sampling rate relation between trajectory data and link counts with a step-by-step statement of the probability dependences, offering rigorous evidence for formulating the constraints in estimating the unbiased OD flows. More so, the formulation of local sample rate compositions actually expresses the assignment relation in a more generalizable form than the conventional equations on the flows. The OMI postulate has also been stated in a clear formulation by conditional probability formulation with derived consequences. The estimation approach overall is able to achieve the estimation in a reliable yet economical way. Practically, this study offers a data-driven instance to estimate OD matrix from trajectory observations, avoiding the conventional complex process of traffic assignment modeling.

Future work can extend the analysis with large spatial scales and more links with traffic counts, to better interpret the heterogeneity issue on the basis of a more complete network path instead of specific itineraries. A finer differentiation of the OD demand segmentation such as by different time periods, types of vehicles and others would yield much significance in practical applications. Different adjustment approaches can also be compared and evaluated to determine the state-of-practice in terms of accuracy and efficiency.

CHAPTER 8. CONCLUSIONS AND FUTURE PERSPECTIVE

8.1. Research Synthesis and Contributions

With the increasing pace of modern life, mobility is becoming a keyword to characterize the current society. Understanding human mobility has been a major contributor to the investigation of many important issues in terms of urban forms, land uses, and transportation systems. Meanwhile, with the advancement of information technologies, the wide availability of trajectory data is emerging as a new powerful tool, enabling a comprehensive discovery of human mobility patterns. This thesis takes advantage of the availability of massive Floating Car Data (FCD) to study vehicle-related human mobility patterns regarding the physical movements and the associated social impacts. To discover the associated mobility phenomena in a more complete way, the mobility patterns are studied at two levels: the individual level of personal behaviors and the spatial level of aggregated phenomena. Specific research questions addressed at the individual level pertain to the vehicle usage patterns in terms of daily trip-making, the traveling regularities about “anchoring” places where the mobility activities are centered around, and the travel time estimation based on observed trajectories. While at the spatial level, the questions solved include revealing functional occupation of urban areas from human activities, identifying spatial relations between places to find out core areas and bonded communities, and quantifying the spatial interaction intensity by estimating the traffic flow between places.

The first contribution of the thesis comes from its methodological meanings. For the scientific aspect, this thesis develops a general knowledge framework in characterizing mobility patterns from a semantic view for capturing the individual regularities and spatial evolutions. These patterns could assist in a more comprehensive diagnosing of the territory mobility dynamics by obtaining the current traffic situation, the service mismatches, and the prevision of future demand derivation. In addition, the easy-updatable essence of modern data also enables the possibility of quick responses to the mobility changes at a large scale. For the technical aspect, this thesis proposes a series of data-driven methods to enhance mobility analysis by leveraging the massive trajectory data and the great potentials of machine learning methods. Various techniques have been integrated with their own applicability into mobility pattern exploration, such as clustering, topic modeling, kernel density estimation, ensemble learning, graph partitioning, distribution modeling, and many other possible ones. Common advantages of using these techniques include the independence from human intervention, the ability to learn things from up-to-date observations, the feasibility of handling multidimensional features lying in mobility, and the potential of discovering new patterns of diverse situations that are free from pre-set rules. Compared to the traditional modeling process, this research extends the mobility analytics from a data exploration standing point, which can be automated to ease the process for future studies and help in decision makings.

The second part of contribution lies in its practical implications. Many use cases in the mobility landscape have been proved with good applicability by using the big trajectory data. For demand issues, the discovered vehicle usage patterns and regularities are useable for an investigation of different groups of roadway users, the preferences of which can be integrated into further analyses such as building models for agent-based simulation, etc. For planning issues, the explored functional areas and spatial communities provide insights of the current territory morphology, which could be used as hints for future calibration of territory planning and regional cooperation guidance. For traffic management, the Floating Car Data shows its significant prospects in measuring roadway network conditions at a low cost for wide spatial-temporal coverage. The traffic flow matrices can also be obtained in a much simpler way by utilizing the rich path information to replace the conventional complicated assignment modeling.

Empirical findings were also obtained through case studies using real-world data in the Paris Region. At the individual level, most vehicles were found associated with local usage within specific areas through the vehicle usage analysis. The prevailing pattern was found on short-medium trips around pericenter and suburb areas. Besides, based on investigating the trajectory over time, most of the vehicles in the sampling could be detected with home places while around half of them could be found with work places while maintaining good accuracy. Other secondary places could also be effectively extracted based on the diverse frequentation of visiting. While looking at the spatial structures, the core activity areas could be well discovered, evaluated against census surveys, despite the low sampling rate of the trajectory data. The graph-partitioning method also showed its good capability in detecting spatial community with dense intra-connections based on the spatial relational graph constructed from trajectory data.

More reliable yet simpler estimations of traffic metrics, such as travel time and origin-destination flow were also proved obtainable using the rich and massive observations from trajectories. Through these applications, this thesis offers real-world instances contributing to concluding the potential and feasibility of using trajectory data to obtain extensive mobility measurements.

8.2. Limitations and Outlook for Future Research

This thesis concludes with a discussion on the limitations and outlook for future perspectives.

First of all, the thesis pays particular attention to mining Floating Car Data for territory and mobility analytics. Although the developed algorithms and methods are much transferrable to other GPS-based trajectory data, the empirical findings are limited to roadway-related mobility, which accounts for only a partial segment of total travel demand. The obtained patterns and trends are subject to the traits of vehicle driving. Future work may incorporate other modes of transport data to have a more complete picture of human mobility investigation.

Second, the massive data bring us rich information but in turn draws the concern of many other problems, such as the complexity of massive data processing, erroneous information, and so on. As for now, the representativeness of such data is still a persistent issue that is commonly existed despite the sources. Due to the concern of privacy protection, such a sampling bias problem is inherent and cannot be substantially solved from the collection phase in the upstream, while the users are anonymized without the inclusion of social-demographic information allowed. In this thesis, we experimented with two ways to calibrate the representativeness problem in the post phases using census data and local traffic counts data respectively. However, it is difficult to maintain precision and large-scale applicability at the same time. Thus, future work may extend to systematically explore the method for addressing such an issue.

Third, the thesis focuses on the discovery of mobility patterns and trends based on data mining and machine learning algorithms. The emphasis lies on recovering the information conveyed in the ubiquitous trajectories. However, the relations against influencing variables are not fully investigated, yet uncovering the underlying mechanism with them. This aspect can be improved with the further development of explanatory models for examining the contributing factors and quantifying the relations.

Last, due to the lack of ground truth information, a high level of validation of such mobility pattern exploration is restricted, which is, however, a common problem, in the current research of this field. Nevertheless, future work is expected to fuse multi-source of information to make them confront each other for the evaluation, thus improving the robustness of the discovered knowledge.

REFERENCES

2.1. Gaussian mixture models—Scikit-learn 0.23.1 documentation. (n.d.). Retrieved July 31, 2020, from https://scikit-learn.org/stable/modules/mixture.html#gmm

Aguilera, A., & Boutueil, V. (2018). *Urban mobility and the smartphone: Transportation, travel behavior and public policy*. Elsevier.

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, *17*(1), 3–27.

Altintasi, O., Tuydes-Yaman, H., & Tuncay, K. (2017). Detection of urban traffic patterns from Floating Car Data (FCD). *Transportation Research Procedia*, *22*, 382–391.

Andrade, T., Cancela, B., & Gama, J. (2019). Mining Human Mobility Data to Discover Locations and Habits. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 390–401.

Angkhawey, U., & Muangsin, V. (2018). Detecting points of interest in a city from taxi gps with adaptive dbscan. *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*, 1–6.

Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, *7*(5), 275–286.

Ásmundsdóttir, R. (2008). Dynamic ODmatrix estimation using floating car data. *MSc Thesis – Civil Engineering, Delft University of Technology*, 1–148.

Astarita, V., Giofrè, V. P., Guido, G., & Vitale, A. (2017). *The Use of Adaptive Traffic Signal Systems Based on Floating Car Data* [Research article]. Wireless Communications and Mobile Computing. https://doi.org/10.1155/2017/4617451

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., & Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data.

Transportation Research Part C: Emerging Technologies, *101*, 254–275. https://doi.org/10.1016/j.trc.2019.02.013

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., & Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports, 734,* 1–74.

Beckmann, M., McGuire, C. B., & Winsten, C. B. (1956). *Studies in the Economics of Transportation*.

Bekhor, S., Ben-Akiva, M. E., & Ramming, M. S. (2006). Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research*, *144*(1), 235–247.

Bellefon, M.-P. de, Eusebio, P., Forest, J., Pégaz-Blanc, O., & Warnod, R. (2020, October 21). *In France, nine out of ten people live in the catchment area of a city—Insee Focus—211*. https://www.insee.fr/fr/statistiques/4806694?pk_campaign=avis-parution#consulter

Ben-Akiva, M. E., Lerman, S. R., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand* (Vol. 9). MIT press.

Bhat, C., Govindarajan, A., & Pulugurta, V. (1998). Disaggregate attraction-end choice modeling formulation and empirical analysis. *Transportation Research Record*, *1645*, 60–68. https://doi.org/10.3141/1645-08

Bierlaire, M. (2002). The total demand scale: A new measure of quality for static and dynamic origin-destination trip tables. *Transportation Research Part B: Methodological*, *36*(9), 837–850. https://doi.org/10.1016/S0191-2615(01)00036-4

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boarnet, M. G. (1994). The monocentric model and employment location. *Journal of Urban Economics*, *36*(1), 79–97.

Bouchard, R. J., & Pyers, C. E. (1965). Use of Gravity Model for Describing Urban Travel. *Highway Research Record*, *88,* 1–43.

Brockfeld, E., Lorkowski, S., Mieth, P., & Wagner, P. (2007). Benefits and limits of recent floating car data technology–an evaluation study. *11th WCTR Conference, Berkeley, USA.*

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*(7075), 462–465.

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, *10*(4), 36–44. https://doi.org/10.1109/MPRV.2011.41

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, *26*, 301–313.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172.

Cao, X., Cong, G., & Jensen, C. S. (2010). Mining significant semantic locations from GPS data. *Proceedings of the VLDB Endowment*, *3*(1–2), 1009–1020.

Cao, Y., Tang, K., Sun, J., & Ji, Y. (2021). Day-to-day dynamic origin–destination flow estimation using connected vehicle trajectories and automatic vehicle identification data. *Transportation Research Part C: Emerging Technologies*, *129*. https://doi.org/10.1016/j.trc.2021.103241

Carpenter, C., Fowler, M., & Adler, T. (2012). Generating route-specific origin-destination tables using bluetooth technology. *Transportation Research Record*, *2308*, 96–102. https://doi.org/10.3141/2308-10

Cascetta, E., Papola, A., Marzano, V., Simonelli, F., & Vitiello, I. (2013). Quasi-dynamic estimation of o-d flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transportation Research Part B: Methodological*, *55*, 171–187. https://doi.org/10.1016/j.trb.2013.06.007

Castillo, E., Menéndez, J. M., & Jiménez, P. (2008). Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. *Transportation Research Part B: Methodological*, *42*(5), 455–481. https://doi.org/10.1016/j.trb.2007.09.004

Chakirov, A., & Erath, A. (2012). Activity identification and primary location modelling based on smart card payment data for public transport. *13th International Conference on Travel Behaviour Research (IATBR 2012)*.

Chen, H., Yang, C., & Xu, X. (2017). Clustering vehicle temporal and spatial travel behavior using license plate recognition data. *Journal of Advanced Transportation*, *2017*.

Chen, M., Arribas-Bel, D., & Singleton, A. (2019). Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*, *21*(1), 89–109.

Cheng, X., Zhao, X., Xu, Z., Zhou, J., & Yang, N. (2015). Prediction of the shortest travel time based on intersection delay. *2015 IEEE First International Smart Cities Conference (ISC2)*, 1–7.

Ciscal-Terry, W., Dell'Amico, M., Hadjidimitriou, N. S., & Iori, M. (2016). An analysis of drivers route choice behaviour using GPS data and optimal alternatives. *Journal of Transport Geography*, *51*, 119–129.

Cornuéjols, A., Wemmert, C., Gançarski, P., & Bennani, Y. (2018). Collaborative clustering: Why, when, what and how. *Information Fusion*, *39*, 81–95.

Dabbas, H., Fourati, W., & Friedrich, B. (2021). Using Floating Car Data in Route Choice Modelling-Field Study. *Transportation Research Procedia*, *52*, 700–707.

Dao, V.-L., Bothorel, C., & Lenca, P. (2018). Community structure: A comparative evaluation of community detection methods. *ArXiv Preprint ArXiv:1812.06598*.

Découpage Morphologique d'Île-de-France. (2017). https://data-iau-idf.opendata.arcgis.com/datasets/d%C3%A9coupage-morphologique-d%C3%AEle-de-france

Dewulf, B., Neutens, T., Vanlommel, M., Logghe, S., De Maeyer, P., Witlox, F., De Weerdt, Y., & Van de Weghe, N. (2015). Examining commuting patterns using Floating Car Data and circular statistics: Exploring the use of new methods and visualizations to study travel times. *Journal of Transport Geography*, *48*, 41–51.

Ding, L., Jahnke, M., Wang, S., & Karja, K. (2016). Understanding spatiotemporal mobility patterns related to transport hubs from floating car data. *Proc. Int. Conf. Location-Based Services*, 175–185.

Dixon, M. P., & Rilett, L. R. (2002). Real-time OD estimation using automatic vehicle identification and traffic count data. *Computer-Aided Civil and Infrastructure Engineering*, *17*(1), 7–21. https://doi.org/10.1111/1467-8667.00248

Dubrova, S. V., Podlipskiy, I. I., Kurilenko, V. V., & Siabato, W. (2015). Functional city zoning. Environmental assessment of eco-geological substance migration flows. *Environmental Pollution*, *197*, 165–172.

Edelsbrunner, H., Kirkpatrick, D., & Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, *29*(4), 551–559.

Ehmke, J. F., Meisel, S., & Mattfeld, D. C. (2012). Floating car based travel times for city logistics. *Transportation Research Part C: Emerging Technologies*, *21*(1), 338–352.

Einstein, A. (1956). *Investigations on the Theory of the Brownian Movement*. Courier Corporation.

Fabritiis, C. de, Ragona, R., & Valenti, G. (2008). Traffic Estimation And Prediction Based On Real Time Floating Car Data. *2008 11th International IEEE Conference on Intelligent Transportation Systems*, 197–203. https://doi.org/10.1109/ITSC.2008.4732534

Ferrari, C., Parola, F., & Gattorna, E. (2011). Measuring the quality of port hinterland accessibility: The Ligurian case. *Transport Policy*, *18*(2), 382–391.

Furletti, B., Cintia, P., Renso, C., & Spinsanti, L. (2013). Inferring human activities from GPS tracks. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 1–8.

Fusco, G., Colombaroni, C., & Isaenko, N. (2016). Short-term speed predictions exploiting big data on large urban road networks. *Transportation Research Part C: Emerging Technologies*, *73*, 183–201.

Giannotti, F., & Pedreschi, D. (2008). *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer Science & Business Media.

Golan, A., Judge, G., & Miller, D. (1996). Maximum Entropy Econometrics. *Wiley, Chichester, England*.

Gómez, P., Menéndez, M., & Mérida-Casermeiro, E. (2015). Evaluation of trade-offs between two data sources for the accurate estimation of origin-destination matrices. *Transportmetrica B*, *3*(3), 225–245. https://doi.org/10.1080/21680566.2015.1025892

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences*, *138*(Supplement C), 557–565.

Gong, L., Sato, H., Yamamoto, T., Miwa, T., & Morikawa, T. (2015). Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation*, *23*(3), 202–213.

Guo, J., Liu, Y., Li, X., Huang, W., Cao, J., & Wei, Y. (2019). Enhanced least square based dynamic OD matrix estimation using Radio Frequency Identification data. *Mathematics and Computers in Simulation*, *155*, 27–40. https://doi.org/10.1016/j.matcom.2017.10.014

Hainen, A. M., Wasson, J. S., Hubbard, S. M. L., Remias, S. M., Farnsworth, G. D., & Bullock, D. M. (2011). Estimating route choice and travel time reliability with field observations of bluetooth probe vehicles. *Transportation Research Record*, *2256*, 43–50. https://doi.org/10.3141/2256-06

Hall, F. L. (1996). Traffic stream characteristics. *Traffic Flow Theory. US Federal Highway Administration*, *36*.

Han, H., Yang, C., Wang, E., Song, J., & Zhang, M. (2015). Evolution of jobs-housing spatial relationship in Beijing Metropolitan Area: A job accessibility perspective. *Chinese Geographical Science*, *25*(3), 375–388.

Heanue, K. E., & Pyers, C. E. (1966). A comparative evaluation of trip distribution procedures. *Highway Research Record*, *114*, 20–50.

Herder, E., & Siehndel, P. (2012). Daily and weekly patterns in human mobility. *UMAP Workshops*, 338–340.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 856–864.

Hofleitner, A., Herring, R., Abbeel, P., & Bayen, A. (2012). Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. *IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1679–1693.

Hofleitner, A., Herring, R., & Bayen, A. (2012). Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological*, *46*(9), 1097–1122.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, *2010*(1), 21–48.

Houbraken, M., Logghe, S., Schreuder, M., Audenaert, P., Colle, D., & Pickavet, M. (2017). Automated Incident Detection Using Real-Time Floating Car Data. *Journal of Advanced Transportation*, *2017*. https://doaj.org

Huang, L., Li, Q., & Yue, Y. (2010). Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, 27–30.

Hunter, T., Herring, R., Abbeel, P., & Bayen, A. (2009). Path and travel time inference from GPS probe vehicle data. *NIPS Analyzing Networks and Learning with Graphs*, *12*(1), 1–8.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Jahnke, M., Ding, L., Karja, K., & Wang, S. (2017). Identifying Origin/Destination Hotspots in Floating Car Data for Visual Analysis of Traveling Behavior. In G. Gartner & H. Huang (Eds.), *Progress in Location-Based Services 2016* (pp. 253–269). Springer International Publishing. https://doi.org/10.1007/978-3-319-47289-8_13

Jenelius, E., & Koutsopoulos, H. N. (2013). Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, *53*, 64–81.

Kieć, M., Ambros, J., Bąk, R., & Gogolín, O. (2018). Evaluation of safety effect of turbo-roundabout lane dividers using floating car data and video observation. *Accident Analysis & Prevention*. https://doi.org/10.1016/j.aap.2018.05.009

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS One*, *9*(6), e96180.

Leurent, F. (1997). Generalized maximum entropy methods in transportation planning. *Work Note Interets*, 1–13.

Li, D., Miwa, T., Morikawa, T., & Liu, P. (2016). Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets. *Transportation Research Part C: Emerging Technologies*, *67*, 31–46.

Li, Q., Zhang, T., Wang, H., & Zeng, Z. (2011). Dynamic accessibility mapping using floating car data: A network-constrained density estimation approach. *Journal of Transport Geography*, *19*(3), 379–393.

Li, S., & Liu, Y. (2016). The jobs-housing relationship and commuting in Guangzhou, China: Hukou and dual structure. *Journal of Transport Geography, 54*, 286–294.

Lian, J., Li, Y., Gu, W., Huang, S.-L., & Zhang, L. (2018). Joint mobility pattern mining with urban region partitions. *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 362–371.

Liao, L., Fox, D., & Kautz, H. (2007). Hierarchical conditional random fields for GPS-based activity recognition. In *Robotics Research* (pp. 487–506). Springer.

Lin, K., Xu, Z., Qiu, M., Wang, X., & Han, T. (2016). Noise filtering, trajectory compression and trajectory segmentation on GPS data. *2016 11th International Conference on Computer Science & Education (ICCSE)*, 490–495.

Liu, X., & Ban, Y. (2013). Uncovering Spatio-Temporal Cluster Patterns Using Massive Floating Car Data. *ISPRS International Journal of Geo-Information*, *2*(2), 371–384. https://doi.org/10.3390/ijgi2020371

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, *43*, 78–90.

Liu, X., Liu, K., Li, M., & Lu, F. (2017). A ST-CRF Map-Matching Method for Low-Frequency Floating Car Data. *IEEE Transactions on Intelligent Transportation Systems*, *18*(5), 1241–1254. https://doi.org/10.1109/TITS.2016.2604484

Liu, Y., Singleton, A., Arribas-Bel, D., & Chen, M. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: A case study in New York City. *Computers, Environment and Urban Systems*, *86*, 101592.

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic ‘source-sink areas’: Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning*, *106*(1), 73–87.

Long Cheu, R., Xie, C., & Lee, D.-H. (2002). Probe vehicle population and sample size for arterial speed estimation. *Computer-Aided Civil and Infrastructure Engineering*, *17*(1), 53–60.

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, *53*, 19–35.

Mannini, L., Cipriani, E., Crisalli, U., Gemma, A., & Vaccaro, G. (2017). On-Street Parking Search Time Estimation Using FCD Data. *Transportation Research Procedia*, *27*, 929–936. https://doi.org/10.1016/j.trpro.2017.12.149

Michau, G., Pustelnik, N., Borgnat, P., Abry, P., Bhaskar, A., & Chung, E. (2019). Combining traffic counts and Bluetooth data for link-origin-destination matrix estimation in large urban networks: The Brisbane case study. *ArXiv:1907.07495v1 [Eess.SP] 17 Jul 2019*, 1–15.

Michau, G., Pustelnik, N., Borgnat, P., Abry, P., Nantes, A., Bhaskar, A., & Chung, E. (2017). A Primal-Dual Algorithm for Link Dependent Origin Destination Matrix Estimation. *IEEE Transactions on Signal and Information Processing over Networks*, *3*(1), 104–113. https://doi.org/10.1109/TSIPN.2016.2623094

Mitra, S., & Saha, S. (2019). A multiobjective multi-view cluster ensemble technique: Application in patient subclassification. *Plos One*, *14*(5), e0216904.

Mori, U., Mendiburu, A., Álvarez, M., & Lozano, J. A. (2015). A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science*, *11*(2), 119–157.

Nazem, M., Trépanier, M., & Morency, C. (2013). Integrated intervening opportunities model for public transit trip generation-distribution. *Transportation Research Record*, *2350*, 47–57. https://doi.org/10.3141/2350-06

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113.

Newson, P., & Krumm, J. (2009). Hidden Markov map matching through noise and sparseness. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 336–343.

Nguyen, T. T., Armoogum, J., Madre, J. L., & Pham, T. H. T. (2017). GPS and travel diary: Two recordings of the same mobility. *ISCTSC, 11th International Conference on Transport Survey Methods*, 13p.

Nigro, M., Cipriani, E., & del Giudice, A. (2018). Exploiting floating car data for time-dependent Origin–Destination matrices estimation. *Journal of Intelligent Transportation Systems*, *22*(2), 159–174.

Ortuzar, J. de D., & Willumsen, L. G. (2004). *Introduction in modeling transport*. West Sussex, England: John Wiley & Sons, LTD.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., & Pelekis, N. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, *45*(4), 1–32.

Peuportier, B., Leurent, F., & Roger-Estrade, J. (2019). *Éco-conception des ensembles bâtis et des infrastructures, tome 2*.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Proulhac, L. (2019). Qui se cache derrière la baisse de la mobilité automobile en Île-de-France? Une analyse typologique des pratiques modales des actifs occupés franciliens. *Cybergeo: European Journal of Geography*.

Qi, G., Li, X., Li, S., Pan, G., Wang, Z., & Zhang, D. (2011). Measuring social functions of city regions from large-scale taxi behaviors. *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 384–388.

Rahmani, M., Jenelius, E., & Koutsopoulos, H. N. (2015). Non-parametric estimation of route travel time distributions from low-frequency floating car data. *Transportation Research Part C: Emerging Technologies*, *58*, 343–362. https://doi.org/10.1016/j.trc.2015.01.015

Rahmani, M., Koutsopoulos, H. N., & Jenelius, E. (2017). Travel time estimation from sparse floating car data with consistent path inference: A fixed point approach. *Transportation Research Part C: Emerging Technologies*, *85*, 628–643. https://doi.org/10.1016/j.trc.2017.10.012

Ramezani, M., & Geroliminis, N. (2012). On the estimation of arterial route travel time distribution with Markov chains. *Transportation Research Part B: Methodological*, *46*(10), 1576–1590. https://doi.org/10.1016/j.trb.2012.08.004

Ran, B., Song, L., Zhang, J., Cheng, Y., & Tan, H. (2016). Using tensor completion method to achieving better coverage of traffic state estimation from sparse floating car data. *PloS One*, *11*(7), e0157420.

Rao, W., Wu, Y. J., Xia, J., Ou, J., & Kluger, R. (2018). Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transportation Research Part C: Emerging Technologies*, *95*, 29–46. https://doi.org/10.1016/j.trc.2018.07.002

Rempe, F., Franeck, P., Fastenrath, U., & Bogenberger, K. (2017). A phase-based smoothing method for accurate traffic speed estimation with floating car data. *Transportation Research Part C: Emerging Technologies*, *85*, 644–663.

Ren, K., Kim, A. M., & Kuhn, K. (2018). Exploration of the Evolution of Airport Ground Delay Programs. *Transportation Research Record*, *2672*(23), 71–81.

Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D., & Giannotti, F. (2012). Discovering the geographical borders of human mobility. *KI-Künstliche Intelligenz*, *26*(3), 253–260.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Russo, F., Villani, M. G., D’Elia, I., D’Isidoro, M., Liberto, C., Piersanti, A., Tinarelli, G., Valenti, G., & Ciancarella, L. (2021). A Study of Traffic Emissions Based on Floating Car Data for Urban Scale Air Quality Applications. *Atmosphere*, *12*(8), 1064.

Sarti, L., Bravi, L., Sambo, F., Taccari, L., Simoncini, M., Salti, S., & Lori, A. (2017). Stop purpose classification from gps data of commercial vehicle fleets. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 280–287.

Seo, T., & Kusakabe, T. (2015). Probe Vehicle-based Traffic Flow Estimation Method without Fundamental Diagram. *Transportation Research Procedia*, *9*, 149–163. https://doi.org/10.1016/j.trpro.2015.07.009

Shen, J., & Ban, Y. (2016). Route Choice of the Shortest Travel Time Based on Floating Car Data. *Journal of Sensors*, *2016*. https://doaj.org

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*(7392), 96–100.

Simoncini, M., Sambo, F., Taccari, L., Bravi, L., Salti, S., & Lori, A. (2016). Vehicle Classification from Low Frequency GPS Data. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 1159–1166. https://doi.org/10.1109/ICDMW.2016.0167

Simoncini, M., Taccari, L., Sambo, F., Bravi, L., Salti, S., & Lori, A. (2018). Vehicle classification from low-frequency GPS data with recurrent neural networks. *Transportation Research Part C: Emerging Technologies*, *91*, 176–191. https://doi.org/10.1016/j.trc.2018.03.024

Sohn, K., & Kim, D. (2008). Dynamic origin-destination flow estimation using cellular communication system. *IEEE Transactions on Vehicular Technology*, *57*(5), 2703–2713. https://doi.org/10.1109/TVT.2007.912336

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021.

Spatial Join—GIS Wiki. (n.d.). Retrieved October 26, 2021, from http://wiki.gis.com/wiki/index.php/Spatial_Join

Stouffer, S. A. (1940). Intervening opportunities: A theory relating mobility and distance. *American Sociological Review*, *5*(6), 845–867.

Sun, D., Leurent, F., & Xie, X. (2020). Floating Car Data mining: Identifying vehicle types on the basis of daily usage patterns. *Transportation Research Procedia*, *47*, 147–154.

Sun, D., Leurent, F., & Xie, X. (2021a). Discovering vehicle usage patterns on the basis of daily mobility profiles derived from floating car data. *Transportation Letters*, *13*(3), 163–171.

Sun, D., Leurent, F., & Xie, X. (2021b). Mining Vehicle Trajectories to Discover Individual Significant Places: Case Study using Floating Car Data in the Paris Region. *Transportation Research Record*, 0361198121995500.

Sun, D., Leurent, F., & Xie, X. (in press). Floating Car Data mining: Identifying vehicle types on the basis of daily usage patterns. *Transportation Research Procedia*.

Sun, D., Zhang, C., Zhang, L., Chen, F., & Peng, Z.-R. (2014). Urban travel behavior analyses and route prediction based on floating car data. *Transportation Letters*, *6*(3), 118–125.

Sun, J., & Feng, Y. (2011). A Novel OD Estimation Method Based on Automatic Vehicle Identification Data. *Communications in Computer and Information Science*, *135*(PART 2), 461–470. https://doi.org/10.1007/978-3-642-18134-4_74

Sun, Y., Fan, H., Li, M., & Zipf, A. (2016). Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning B: Planning and Design*, *43*(3), 480–498.

Sun, Z., & Ban, X. (Jeff). (2013). Vehicle classification using GPS data. *Transportation Research Part C: Emerging Technologies*, *37*, 102–117. https://doi.org/10.1016/j.trc.2013.09.015

Sunderrajan, A., Viswanathan, V., Cai, W., & Knoll, A. (2016). Traffic State Estimation Using Floating Car Data. *Procedia Computer Science*, *80*, 2008–2018. https://doi.org/10.1016/j.procs.2016.05.521

Survey of the use of road freight vehicles (TRM). (2018, November 21). https://www.statistiques.developpement-durable.gouv.fr/enquete-sur-lutilisation-des-vehicules-de-transport-routier-de-marchandises-trm

Suzuki, J., Suhara, Y., Toda, H., & Nishida, K. (2019). Personalized visited-poi assignment to individual raw gps trajectories. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, *5*(3), 1–28.

Tang, J., Liu, F., Wang, Y., & Wang, H. (2015). Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and Its Applications*, *438*, 140–153.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276.

Tomaney, J. (2009). Region. In R. Kitchin & N. Thrift (Eds.), *International Encyclopedia of Human Geography* (pp. 136–150). Elsevier. https://doi.org/10.1016/B978-008044910-4.00859-2

Tongsinoot, L., & Muangsin, V. (2017). Exploring home and work locations in a city from mobile phone data. *2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 123–129.

Topchy, A., Jain, A. K., & Punch, W. (2004). A mixture model for clustering ensembles. *Proceedings of the 2004 SIAM International Conference on Data Mining*, 379–390.

Uncles, M. D., Ben-Akiva, M., & Lerman, S. R. (1987). Discrete Choice Analysis: Theory and Application to Travel Demand. *The Journal of the Operational Research Society*, *38*(4), 370. https://doi.org/10.2307/2582065

Valiquette, F., & Morency, C. (2010). Trip chaining and its impact on travel behaviour. *12th World Congress on Transportation Research*, 11–15.

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, *34*(4), 935–960.

Von Eye, A. (2003). *Configural frequency analysis: Methods, models, and applications*. Psychology Press.

Wang, A., Zhang, A., Chan, E. H., Shi, W., Zhou, X., & Liu, Z. (2021). A review of human mobility research based on big data and its implication for smart city development. *ISPRS International Journal of Geo-Information*, *10*(1), 13.

Wang, H., Deng, Y., Tian, E., & Wang, K. (2014). A comparative study of methods for delineating sphere of urban influence: A case study on central China. *Chinese Geographical Science*, *24*(6), 751–762.

Wang, J., Kong, X., Xia, F., & Sun, L. (2019). Urban human mobility: Data-driven modeling and prediction. *Acm Sigkdd Explorations Newsletter*, *21*(1), 1–19.

Wang, X., Peng, L., Chi, T., Li, M., Yao, X., & Shao, J. (2015). A hidden Markov model for urban-scale traffic estimation using floating car data. *PloS One*, *10*(12), e0145348.

Wegener, M., & Fürst, F. (2004). Land-use transport interaction: State of the art. *Available at SSRN 1434678*.

Witayangkurn, A., Horanont, T., Nagai, M., & Shibasaki, R. (2015). Large Scale Mobility Analysis: Extracting Significant Places Using Hadoop/Hive and Spatial Processing. *International Conference on Knowledge, Information, and Creativity Support Systems*, 205–219.

Wu, J., Feng, Z., Zhang, X., Xu, Y., & Peng, J. (2020). Delineating urban hinterland boundaries in the Pearl River Delta: An approach integrating toponym co-occurrence with field strength model. *Cities*, *96*, 102457.

Xie, C., Kockelman, K. M., & Waller, S. T. (2010). Maximum entropy method for subnetwork origin-destination trip matrix estimation. *Transportation Research Record*, *2196*, 111–119. https://doi.org/10.3141/2196-12

Xue, A. Y., Zhang, R., Zheng, Y., Xie, X., Huang, J., & Xu, Z. (2013). Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 254–265.

Yang, X., Lu, Y., & Hao, W. (2017). Origin-destination estimation using probe vehicle trajectory and link counts. *Journal of Advanced Transportation*, *2017*, 1–18. https://doi.org/10.1155/2017/4341532

Yao, J., & Chen, A. (2014). An analysis of logit and weibit route choices in stochastic assignment paradox. *Transportation Research Part B: Methodological*, *69*, 31–49. https://doi.org/10.1016/j.trb.2014.07.006

Yao, R., & Bekhor, S. (2020). Data-driven choice set generation and estimation of route choice models. *Transportation Research Part C: Emerging Technologies*, *121*, 102832.

Ye, Y., Zheng, Y., Chen, Y., Feng, J., & Xie, X. (2009). Mining individual life pattern based on location history. *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, 1–10.

Yin, B. (2019). *Car trips clustering based on EGT 2010*.

Yin, J., Chai, X., & Yang, Q. (2004). High-level goal recognition in a wireless LAN. *AAAI*, 578–584.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., & Huang, Y. (2010). T-drive: Driving directions based on taxi trajectories. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 99–108.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2014). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *27*(3), 712–725.

Yue, Y., Wang, H., Hu, B., Li, Q., Li, Y., & Yeh, A. G. O. (2012). Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Computers, Environment and Urban Systems*, *36*(2), 140–153. https://doi.org/10.1016/j.compenvurbsys.2011.09.002

Zhao, Y., Zhu, X., Guo, W., She, B., Yue, H., & Li, M. (2019). Exploring the Weekly Travel Patterns of Private Vehicles Using Automatic Vehicle Identification Data: A Case Study of Wuhan, China. *Sustainability*, *11*(21), 6152.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 1–41.

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(3), 1–55.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., & Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, *25*(3), 12-es.

Zhou, X., & Mahmassani, H. S. (2006). Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Transactions on Intelligent Transportation Systems*, *7*(1), 105–114. https://doi.org/10.1109/TITS.2006.869629

Zhu, S., & Levinson, D. (2015). Do people use the shortest path? An empirical test of Wardrop's first principle. *PloS One*, *10*(8), e0134322.

Zipf, G. K. (1946). The P 1 P 2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, *11*(6), 677–686.