Modeling Route Choice Behavior Using Smartphone Data

THÈSE N^{O} 5649 (2013)

PRÉSENTÉE LE 25 JANVIER 2013

À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT

LABORATOIRE TRANSPORT ET MOBILITÉ

PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Jingmin CHEN

acceptée sur proposition du jury:

Prof. K. Hess Bellwald, présidente du jury
Prof. M. Bierlaire, directeur de thèse
Dr F. Camara Pereira, rapporteur
Prof. E. Frejinger, rapporteur
Prof. P. Vandergheynst, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

To my parents, my sister, my grandma
&
to Shiqi

Acknowledgments

In this final stage of my thesis, I start to retrospect my PhD life from day one. In these 4 years, I have met many challenges, both in research and life. But thanks to many people, I have fulfilled this life adventure.

Foremost, I would like to express my deepest gratitude to my thesis supervisor Michel Bierlaire. His vision, ideas and immense knowledge have guided me throughout my PhD career. I am also appreciative of his constructive criticisms which helped me to identify and focus the problems. I would have been lost without his advices. He is also a fun friend, and he has educated me successfully in the field of (Belgian) beer.:)

Special thanks to my jury members, who provided constructive and thoughtful feedbacks. My work is a continuation of Emma Frejinger's PhD thesis on route choice modeling. I appreciate her insightful comments and discussions on this topic. I also want to thank Francisco Camara Pereira especially for his detailed comments, and inspirational discussions on inferring transport information from advanced data.

I love the exciting and friendly working atmosphere in TRANSP-OR lab. I am grateful to all the members of the lab. I will remember forever those wonderful moments that we share, in the lab, in SAT, in Zinal, in Ascona... Marianne Ruegge is the most helpful and efficient secretary. Without her help, my life in Switzerland would have been miserable. Jeffrey Newman mentored me in the first year of my PhD. Gunnar Flötteröd always provided me helpful ideas and answers whenever I had questions. My office mate Antonin Danalet shares a lot of information with me, and we often have interesting discussions. He also offered a lot of practical help to me, a foreigner who does not speak French.

I would also like to thank my collaborators in Nokia Research Center in Lausanne (NRCL), and IDIAP Switzerland. Niko Kiukonen helped in smartphone data collection. I benefited from discussions with Gian Paolo Perrucci on learning mobility information from smartphone data. Olivier Dousse and Olivier Bornet helped the smartphone data access. NRCL sponsored my first year of PhD, and also provided the necessary smartphone data for my research. Swiss National Science Foundation has financed my research for 3 years, grant 200021/131998 (October 2009 - January 2013). I appreciate their generous support.

I appreciate the friendship of my Chinese fellows in Switzerland. We have spent many great times together in skiing, cooking, hiking, football and so on. Having them in

Acknowledgments

Switzerland makes my life abroad joyful.
I can't achieve anything without the support from my parents, Gongfeng Chen and Meijiao Chen. Although my decisions might not always meet their expectations, they are always understanding and supportive. I would also like to thank my little sister Xinyi for her love. I left home 10 years ago when she was 6. I feel pity that I did not accompany her as she grows up. I wish to thank my girlfriend Shiqi Yang for her trust and unconditional support. I am very happy that we have managed a healthy relationship in long distance. I am really sorry that I am away for more than 4 years. I wish that we would live together soon. Finally, I dedicate this thesis to my grandma, who passed away 10 years ago. I hope that I can always make you proud.

Lausanne, 10 Jan 2013 Jingmin Chen

Abstract

In this thesis, we develop methods for modeling route choice behavior using smartphone data. The developing global positioning system (GPS) technology and the popularity of smartphones have revolutionized the revealed preference route choice data collection. Nowadays, smartphones are embedded with various kinds of sensors that are able to provide mobility related information. These sensors include GPS, accelerometer and bluetooth. The recorded raw data is not directly applicable to travel behavior study, information such as the paths and transport modes of travels have to be inferred. The inference procedure is challenging due to the poor quality and the variety of the data. This thesis deals with these challenges by proposing probabilistic methods that account for errors in the data, and fusing various kinds of smartphone data in an integrated framework. Based on the inference methods, a route choice modeling framework exploiting GPS data is developed.

The low cost sensors of smartphones observe measurements with significant errors. Moreover, due to practical constraints, such as the limits on smartphone battery volume and the cost of data transmitted via wireless networks, data are usually recorded in a relatively large time interval (low frequency). These drawbacks preclude path identification (a.k.a. map-matching, MM) algorithms that are designed for dense and accurate data from dedicated GPS devices. Therefore, we first propose a probabilistic unimodal MM method that infers the traveled paths from GPS data recorded during a car trip. Instead of deterministically matching a sequence of GPS points to one path, it generates a probabilistic path observation which is composed of a set of candidate paths, and a measurement likelihood for each path. The candidate paths are generated by a candidate path generation algorithm from GPS data. It is capable of dealing with both accurate and dense data (1 second interval) from dedicated GPS devices, and noisy and sparse data (more than 10 seconds interval) from smartphones. A probabilistic measurement model is constructed to calculate the measurement likelihood, which is the likelihood that the observed GPS data is recorded along a given path. The probabilistic measurement model employs structural equation modeling techniques, and the latent status for each measurement is defined as the true location where the measurement is observed. A GPS sensor measurement model relates the status to each GPS measurement; a structural travel model captures the status over time in the network. In this approach,

besides geographical coordinates, speed and time recorded from GPS also contribute to the identification of the true path. Applications and analyses on real data illustrate the robustness and effectiveness of the proposed approach.

Based on the framework designed for the unimodal MM, a multimodal MM method is developed to deal with a more general problem where the trips can be multimodal and the modes are unknown. We infer both path and mode information simultaneously from various kinds of data. The candidate path generation algorithm is extended to deal with multimodal networks, and to generate multimodal paths, of which a transport mode is associated with each road. The latent status includes both location and mode, and the correlation between them is exploited. For example, if the mode is bus, the path should follow bus routes. Besides the most useful GPS data, acceleration and bluetooth also contribute mobility information, so they are integrated in the probabilistic measurement model by constructing a sensor measurement model for each. ACCEL provides motion status that can be used to infer the transport mode. BT data gives the amount of nearby BT devices, which can be used to recognize, for instance, a public transport environment if there are a lot of BT devices nearby. This approach is flexible in two aspects. First, any kind of sensor data can be integrated as long as a corresponding sensor measurement model is provided. Second, any transport network can be added or removed according to necessity and availability. Data recorded from a trip does not need to be preprocessed into unimodal travel segments, so the risk of wrong segmentation is attenuated. Numerical experiments include map visualizations of some example trips, and an analysis of the performance of the transport mode inference.

In the last part of the thesis, we develop a comprehensive and operational route choice modeling framework for estimating route choice models from GPS data. It integrates three components: the probabilistic unimodal MM method for generating probabilistic path observations from GPS data; the "network-free" data approach proposed by Bierlaire & Frejinger (2008) for estimating route choice models from probabilistic path observations; and a new importance sampling based algorithm for sampling path alternatives for the choice model estimation. The proposed path sampling algorithm produces more relevant alternatives by exploiting the GPS data. Numerical analyses using a real transportation network and synthetic choices empirically show that the proposed path sampling algorithm yields more precise parameter estimates than other importance sampling algorithms. The proposed framework accounts for the imprecision in GPS data. The necessary modifications of each method for GPS data are presented. A route choice model estimated from smartphone GPS data shows the viability of applying the proposed route choice modeling framework with real data.

Keywords: route choice model, smartphone data, GPS data, probabilistic measurement model, map-matching, transport mode inference, sampling of alternatives

Résumé

Cette thèse développe des méthodes de modélisation du comportement pour le choix d'itinéraire à l'aide de données issues de smartphones. La technologie GPS et la popularité et le taux de pénétration élevés des smartphones ont révolutionné la collecte de données de préférences révélées pour le choix d'itinéraire. Aujourd'hui, les smartphones contiennent différents types de capteurs capables de fournir des informations sur la mobilité de leur utilisateur. Ces capteurs incluent le GPS (Global Positioning System), l'accéléromètre et Bluetooth. Les données brutes enregistrées ne sont pas directement utilisables pour étudier le comportement de mobilité ; le chemin parcouru et le mode de transport utilisé pendant le trajet doivent être déduits. Cette procédure est rendue difficile de par la faible qualité des données et leur variété. Cette thèse répond à ce challenge en proposant une méthode probabiliste prenant en compte les erreurs dans les données, et fusionnant les différents types de données issues du smartphone dans un modèle intégré. Fondé sur la méthode probabiliste, un modèle de choix d'itinéraire utilisant les données du GPS est développé.

Les capteurs bon marché des smartphones fournissent des mesures avec des erreurs significatives. De plus, pour des raisons pratiques telles que la durée de vie limitée de la batterie ou le coût de la transmission de données, les données sont habituellement enregistrées à de grands intervalles temporels. Ces inconvénients empêchent l'utilisation d'algorithmes d'identification des chemins (« map-matching » en anglais), de même que d'algorithmes d'identification du mode de transport créés pour des données de GPS professionnels plus précis et dont les mesures sont plus fréquentes. C'est pourquoi nous proposons tout d'abord une méthode probabiliste et unimodale d'identification de chemin déduisant le chemin parcouru à partir de données GPS enregistrées pendant un trajet en voiture. Au lieu de faire correspondre déterministiquement une séquence de points GPS à un chemin, la méthode génère un ensemble de chemins candidats et la vraisemblance pour chacun de ces chemins candidats d'avoir été mesurés. Les chemins candidats sont générés par un algorithme dédié à partir des données GPS. L'algorithme est capable d'utilisé à la fois des données précises et denses (intervalle d'une seconde) provenant de capteurs GPS dédiés, et des données bruyantes et éparses (intervalle de plus de 10 secondes) provenant de smartphones. Un modèle de mesure probabiliste est construit pour calculer la vraisemblance de la mesure, c'est-à-dire la vraisemblance que

la localisation GPS observée est générée en suivant un chemin donné. Le modèle de mesure probabiliste utilise des équations structurelles comme technique de modélisation, et le statut latent de chaque mesure est défini comme étant la vraie localisation où la mesure est observée. Un modèle de mesure pour le capteur GPS associe le statut à chaque mesure GPS ; un modèle structurel de mobilité (« travel model » en anglais) déduit le statut dans le réseau à travers le temps. Avec cette approche, en plus des coordonnées géographiques, la vitesse et le temps enregistrés à partir du GPS contribuent aussi à l'identification du chemin réellement parcouru. Des applications et des analyses sur des données réelles illustrent la robustesse et l'efficacité de l'approche proposée.

En s'appuyant sur le cadre défini pour l'identification de chemin dans le cas unimodal, une méthode « map-matching » multimodale est développée pour résoudre le problème plus général où les trajets peuvent être effectués à l'aide de différents modes et que ce dernier est inconnu. Nous déduisons à la fois le chemin et le mode simultanément à partir de données variées. L'algorithme de génération de chemins candidats est étendu pour gérer les réseaux multimodaux, et pour générer des chemins multimodaux, où un mode de transport est associé à chaque tronçon de route. Le statut latent contient le mode et le lieu, and la corrélation entre les deux est utilisées. Par exemple, si le mode de transport est le bus, the chemin doit suivre les lignes de bus existantes. En plus du signal GPS, l'accélération et le signal Bluetooth participent aussi à la collecte d'information sur la mobilité, et ils sont donc intégrés dans le modèle probabiliste de mesure à l'aide d'un modèle de mesure correspondant à chaque capteur. Le modèle pour l'accélération fournit un indicateur de déplacement qui peut être utilisé pour déterminer le mode de transport. Les données du signal Bluetooth fournissent le nombre d'appareils Bluetooth dans les environs, ce qui peut-être exploité par exemple pour identifier l'utilisation des transports publics dans le cas d'un grand nombre d'appareils identifiés. Cette approche est flexible pour deux raisons. D'abord, d'autres types de données de capteurs peuvent être intégrés, pour autant qu'un modèle de mesure correspondant au type de capteur est fourni. Ensuite, tout réseau de transports publics peut être ajouté ou supprimé en fonction des besoins et de la disponibilité. Les données enregistrées lors d'un trajet ne nécessitent pas de subir un prétraitement sous forme de segment de mobilité unimodal ; ainsi, le risque de fausse segmentation est atténué. Des expériences numériques présentent des visualisations sur carte de certains exemples de trajets, et une analyse de la performance de la détection du mode de transport.

Dans la dernière partie de la thèse, nous développons un cadre pour un modèle de choix d'itinéraire opérationnel et complet pour estimer des modèles de choix d'itinéraires à partir de données GPS. Cela comprend trois éléments : la méthode probabiliste unimodale « map-matching » pour générer des observations de chemins probabilistes à partir de données GPS ; l'approche « network-free » proposée par Bierlaire et Frejinger (2008) pour estimer des modèles de choix d'itinéraire à partir d'observations de chemin probabilistes ;

et un nouvel algorithme d'échantillonnage préférentiel pour échantillonner des chemins alternatifs pour l'estimation du modèle de choix. L'algorithme d'échantillonnage de chemin proposé génère plus d'alternatives pertinentes en exploitant les données GPS. Des analyses numériques utilisant un réseau de transport réel et des choix synthétiques montre empiriquement que l'algorithme d'échantillonnage des chemins fournit des estimateurs plus précis des paramètres que d'autres algorithmes d'échantillonnage préférentiels. Le cadre proposé tient compte de l'imprécision des données GPS. Les modifications nécessaires de chaque méthode pour les données GPS sont présentées. Un modèle de choix d'itinéraire estimé à partir de données GPS issues de smartphones montre la viabilité de l'application du cadre de modélisation du choix d'itinéraire proposé à partir de données réelles.

Mots-clés : modèle de choix d'itinéraire, données de smartphones, données GPS, modèle probabiliste de mesure, map-matching, déduction du mode de transport, échantillonnage d'alternatives.

Contents

List of Figures

List of Tables

1 Introduction

Route choice models study individual behavior in making decisions about which route to travel from one location to another. Individual route choice decisions aggregated in the transportation network result in the demand on the transportation infrastructures. Congestion occurs when the demand is concentrated in a geographical area at the same time and exceeds the capacity. So new infrastructures need to be built in order to meet the increasing demand. Transportation service providers want the demand to be well distributed temporally and spatially. For the sake of the environment, policy makers try to attract people to use public transport in order to improve the transportation efficiency. Understanding individual route choice behavior is a fundamental step towards these objectives.

In a real transportation network, there are a large amount of available paths that connect two locations. People evaluate theses alternatives in terms of many factors, such as length, travel time, monetary cost, traffic lights, etc., and choose the overall best one according to her preferences. Discrete route choice model is the most widely used approach for capturing this decision making process based on observed route choice decisions. However, traditional survey methods using interviews are expensive, and are not able to collect long-term or large scale revealed preference (RP) route choice data.

Nowadays, people also try to understand the traffic situation better by accessing real time traffic information through smartphone applications (APP), e.g. Inrix (Inrix 2012). People are not just satisfied with homogeneous information services, but they also want customized services. Therefore, some APPs even support customized configurations, for example, home and work locations, in order to provide information in a more user-friendly and efficient way. Different people have different preferences. For example, many people do not just deterministically prefer the fastest path, but instead, they make trade-offs among different factors, such as travel time and monetary cost. Apparently, quantitatively specifying such behavior in an APP is an infeasible task for the smartphone

user, and an automatic behavior learning algorithm is needed.

The developing GPS technology and the popularity of smartphones have revolutionized the RP route choice data collection. Smartphone is not just an information terminal, but can be also used as a sensor to learn its user's travel behavior. Nowadays, smartphones are embedded with various kinds of sensors, such as global positioning system (GPS), accelerometer (ACCEL) and bluetooth (BT). These sensors provide "rich" data that can be used to reveal context information, in particular mobility information that interests us. The GPS device records locations during the journeys, hence is able to provide information about the traveled paths. ACCEL records motion status, hence it is able to discover the transport mode sometimes and under certain conditions. BT monitors the nearby discoverable BT devices, which can be used to analyze the immediate environment, and obtain hints about the current transport mode. All this information is relevant to RP route choice modeling, but the raw smartphone data has to be treated beforehand. This thesis develops methods that link smartphone data and route choice models in an operational framework.

In 2009, Nokia Research Center in Lausanne (NRCL) launched the Lausanne Data Collection Campaign (LDCC) in Switzerland. They recruited 200 individuals who reside in the Geneva Lake area, and gave a Nokia N95 smartphone to each one. Each person used the smartphone as her personal communication tool, while an application, *EPFLScope*, automatically recorded various kinds of data constantly in 2 years. *EPFLScope* was jointly developed by NRCL, IDIAP Switzerland, and Transport and Mobility Laboratory at Ecole Polytechnique Fédérale de Lausanne. It records almost all kinds of data that are available on a commercial smartphone, including GPS, ACCEL, nearby BT, nearby WIFI access points, SMS logs, call logs, calendar entries, etc. The full list of recorded data is included in Appendix A.

1.1 Contributions

In order to link the smartphone data and the route choice models, route choice observations, describing the traveled path and the transport modes, are collected from the raw smartphone data. We deal with challenges arising from the characteristics of the smartphone data: first, the low quality due to the low cost sensors; second, the variety of the data which comes from different sensors. We also exploit GPS data in developing a new route choice modeling framework. The contributions are summarized based on the outline of the thesis. For each chapter, the reference to the corresponding publication is given.

Chapter 2: literature review We review the literature in route choice modeling, with a focus on methods driven by data collection techniques. Challenges and opportunities in using smartphone data for route choice modeling are identified.

Chapter 3: probabilistic unimodal map-matching method for GPS data The discrete sequences of GPS data need to be associated with the transportation network, in order to generate meaningful paths for route choice models. The poor quality of GPS data collected from smartphones precludes the use of state of the art map-matching (MM) methods. In this paper, we propose a probabilistic unimodal MM approach to perform the association in a probabilistic manner, such that the errors in the GPS are taken into account. This approach produces probabilistic path observations from a sequence of GPS data recorded when the smartphone user traveled with a car. This approach includes two components: a probabilistic measurement model and a candidate path generation algorithm. The probabilistic measurement model calculates the likelihood that a sequence of GPS data has been recorded from a traveler along a given path. It accounts for the inaccuracy of both the smartphone GPS data and the representation of the underlying transportation network. The candidate path generation algorithm produces a set of candidate true paths from sparse smartphone GPS data. This method can reduce the impact of noise in GPS readings, and provides probabilistic path observations for further applications. Numerical experiments on real smartphone GPS data illustrate the effectiveness and robustness of the proposed method in recognizing path from GPS data. This chapter has been published as:

Bierlaire, M., Chen, J., and Newman, J. P (2013). A probabilistic map matching method for smartphone GPS data, Transportation Research Part C: Emerging Technologies 26: 78 - 98.

Chapter 4: probabilistic multimodal map-matching method for rich smartphone data The probabilistic unimodal MM method is extended to deal with a more general problem: the transport mode is unknown and the trip might be multimodal. The proposed multimodal MM method identifies not only the paths, but also the transport mode of each road. This method synthesizes multiple kinds of data from smartphone sensors which provide relevant location or transport mode information. GPS data is used to identify the path and the transport mode. BT data is used to collect hints about the surrounding people. ACCEL data is used to differentiate walk, bike and motor modes. Since path and mode are inferred simultaneously, the correlation between them is exploited. For example if the mode is bus, the path should follow bus routes. Data recorded from a multimodal trip does not need to be preprocessed into multiple unimodal segments. Real smartphone data case studies illustrate that the generated multimodal

paths resemble multimodal travels of smartphone users in different circumstances. Numerical analysis shows good performance of the proposed method in identifying the transportation mode. This chapter has been published as:

Chen, J., and Bierlaire, M. (forthcoming). Probabilistic multimodal map-matching with rich smartphone data. Journal of Intelligent Transportation Systems.

Chapter 5: route choice models estimated from GPS data A comprehensive and operational route choice modeling framework for GPS data is proposed. It integrates: (i) the proposed unimodal MM algorithm for generating probabilistic path observations from GPS data; (ii) the state of the art "network-free" model for estimating discrete route choice models from the probabilistic path observations; and (iii) a new importance sampling algorithm that exploits GPS data. Numerical experiments show that it yields more precise parameter estimates than other importance sampling methods. The necessary modifications of each method are presented such that the integrated route choice modeling framework is applicable to real GPS data. We illustrate an example of estimating route choice behavior of a driver from real smartphone GPS data.

Chapter 6: conclusions We conclude the thesis, and discuss future topics in research and application.

2 Literature review

This chapter reviews the state of the art in route choice modeling, with a focus on methods driven by data collection techniques. For more general literature review on discrete route choice model structures and choice set generation methods, we refer to Bekhor, Ben-Akiva & Ramming (2006), Prato (2009) and Bovy (2009).

We start with an introduction to discrete route choice modeling, followed by an investigation on existing route choice data collection techniques. With the advanced GPS and smartphone data collection techniques, the raw measurements have to be treated in order to obtain route choice observations. Section 2.3 reviews methods for inferring the traveled path and transport mode. Section 2.4 discusses two different ways of conceptualizing and specifying the route choice set. Section 2.5 discusses challenges and opportunities that are related to the topic of this thesis.

2.1 Route choice models

This section introduces discrete choice models applied to route choice modeling problems. We first present the basis of the random utility based discrete choice model. Different model structures designed for route choice are discussed. In particular, we focus on the route choice modeling methods which motivate this thesis. In this thesis, the proposed methods are applicable to multimodal route choice modeling, hence this topic is also introduced.

2.1.1 Discrete choice model overview

Discrete choice models assume that, in observation n, an individual chooses the best alternative i from a discrete set of available options $\mathcal{C}_n$. Each alternative $j \in \mathcal{C}_n$ is evaluated according to a vector of variables x_{jn}, which contains the attributes of

alternative j and the socio-economic characteristics of the decision maker. The evaluation is based on a linear combination of x_{jn}, and the decision maker perceives a utility function $V_{jn} = \beta_j x_{jn}$, where the parameters β_j describe the importance of each factor for the decision maker. The decision maker may not have a perfect knowledge of each alternative, or some factors are uncertain in reality. Hence, a random error term is introduced, and the utility function is modeled in a random form: $U_{jn} = V_{jn} + \varepsilon_{jn}$. The decision makers are assumed to be utility maximizers, and the probability that i is chosen is defined as the probability that U_{in} is the highest utility among all alternatives in $\mathcal{C}_n$:

$$\Pr(i|\mathcal{C}_n) = \Pr(U_{in} = \max_{j \in \mathcal{C}_n} U_{jn}). \tag{2.1}$$

The parameters β are estimated from choice data, which records the choice decisions that have been made by the decision makers. The maximum likelihood estimator is used for estimating the parameters:

$$\widehat{\beta} = \arg\max_{\beta} \prod_{n \in N} \Pr(i|\mathcal{C}_n), \tag{2.2}$$

where $\widehat{\beta}$ denotes the estimated values, and N denotes the set of choice observations that are used in the estimation. Depending on the assumption of the error terms ε_{jn}, there are different forms of discrete choice models. For example, the logit model has the independence from irrelevant alternatives property (Ben-Akiva & Lerman 1985), which results from the assumption that ε_{jn} are independent and identically distributed (i.i.d.) extreme value (EV). Then the choice probability (2.1) becomes:

$$\Pr(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}}}, \tag{2.3}$$

where μ is the scale parameter of the EV distribution.

Discrete choice models are suitable for the route choice behavior modeling problem. The choice set is defined as a set of available paths that connect the traveler's origin and destination (OD). Travelers are assumed to be utility maximizers, and they consider path attributes, such as length, travel time, monetary cost and traffic lights in making route choices. The socio-economic characteristics of the traveler also affect the route choice decisions.

2.1.2 Route choice modeling methodologies

A unique problem in discrete route choice modeling is that the path alternatives are highly correlated, due to their overlapping. Therefore, the independence from irrelevant

alternatives property of the logit model is not appropriate for route choice. Various discrete choice model structures have been proposed and applied in route choice modeling. Researchers are constantly struggling with the trade-off between validity and tractability of the models. For example, mixed multinomial logit (Frejinger & Bierlaire 2007) introduce a mixture of logit models where the error component corresponds to a sub-network. The cross nested logit (Vovsha 1997) allows to capture the overlapping by nests, but real case studies report that the estimation results often collapse to logit (Prato 2009). The combinatorial nature of the problem excludes the usage of paired combinatorial logit (Chu 1989) in practice. Multinomial probit assumes multivariate normal distributed error terms (Burrell 1968, Daganzo & Sheffi 1977), however, the estimation procedure is computationally expensive since the model does not have a closed form.

Variations of logit are commonly used, where the utility function is modified to include a similarity measure, such as the path size (PS; Ben-Akiva & Bierlaire 1999) or the commonality factor (CF; Cascetta, Nuzzolo, Russo & Vitetta 1996). Both CF and PS measure the similarity of alternatives in the choice set, so as to overcome the i.i.d. assumption. Ramming (2002) examine both methods theoretically and empirically, and concludes that the PS method is more advantageous. The original PS specification has the following form (Ben-Akiva & Bierlaire 1999):

$$PS_{in} = \sum_{a \in i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_{ps}} \delta_{aj}}, \tag{2.4}$$

where $a \in i$ denotes an arc on path i; L_a and L_i denotes the length of arc a and path i respectively; dummy variable δ_{aj} equals to one if path j contains arc a, and zero otherwise. Variations of the PS formulation, such as the generalized path size and the path size correction are proposed (Ben-Akiva & Bierlaire 1999, Ramming 2002, Bovy 2007). Theoretical analyses performed by Frejinger & Bierlaire (2007) suggest that the original PS formulation or the path size correction is preferred.

The path size logit (PSL) has the following form of the choice probability:

$$Pr(i|\mathcal{C}_n; \boldsymbol{\beta}) = \frac{e^{\mu V_{in} + \beta_{ps} PS_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn} + \beta_{ps} \ln PS_{jn}}}, \tag{2.5}$$

where β_{ps} is the parameter associated with the PS. The original PSL model fixes β_{ps} to one (Ben-Akiva & Bierlaire 1999). Hoogendoorn-Lanser, van Nes & Bovy (2005) suggest that β_{ps} can capture the decision maker's perception of the path overlapping, therefore should be estimated. Indeed, they have found that estimating β_{ps} yields better empirical results.

The specification of the route choice model highly depends on the route choice data

and the network data. In different contexts, the route choice behavior is affected by different attributes. For example, in Switzerland, each vehicle is charged for a fixed annual fee for highway. While in China, highway is tolled according to the traveled distance. The attributes considered in the route choice model are also subject to the network data. For example, the travel time may not be available in the network data, so the distance is often considered instead. The type of route choice data also affects the route choice model specification. In particular, Bierlaire & Frejinger (2008) propose a new discrete choice modeling framework for route choice data that is not associated with the transportation network ("network-free" data). They suggest that the error in the "network-free" data should be treated in a probabilistic manner. Although they also suggest that the methodology may be applied to GPS data, they present only results based on interview data. One objective of this thesis is to develop methods which make this framework applicable to route choice data collected from smartphones. Indirect inference is another approach that deals with sparse GPS data (Oshyani, Sundberg & Karlström 2012).

2.1.3 Multimodal route choice modeling

Little attention has been paid to the modeling of multimodal route choice behavior, which studies the combination of route and mode choice. Multimodal route choice models are more complex than unimodal models in many respects. First, the network containing transfers between modes is complicated, and this complexity challenges the choice set generation. Second, the evaluation of choices is subject to more attributes than the unimodal case. Third, the correlation among alternatives is more difficult to capture. Recently, Bovy & Hoogendoorn-Lanser (2005) propose a method of modeling multimodal route choice behavior for inter-urban trips, in which the train is the main transportation mode. Relevant research in the Delft University of Technology, the Netherlands, aims at resolving issues arising from the introduction of multimodal networks, including a super-network approach to the generation of multimodal networks, a compatible algorithm to generate choice sets (Fiorenzo Catalano 2007), and an adapted PS formulation to capture the choices' correlation in the PSL model (Hoogendoorn-Lanser 2005, Hoogendoorn-Lanser & Bovy 2007). Transfers, important steps in multimodal trips, are modeled in route choice behavior by Hoogendoorn-Lanser, van Nes & Hoogendoorn (2006). However, some resolutions are not applicable to intra-urban networks, where the train is not the main mode and there are more alternatives in the motor transportation networks. A more practical way of modeling multimodal route choice behavior is to only study the mode choice aspect. For example, Bekhor & Shiftan (2010) model mode choice behavior among car, bus and train, with different access modes to bus stops and railway stations.

2.2 Data collection for route choice modeling

Discrete route choice modeling requires that the choice observations are available for the model estimation. Routes are represented in link by link way. Recording them from travelers is challenging because the identification of each link is difficult. This section reviews different techniques for collecting route choice data. In particular, we focus on revealed preference and new data collection techniques.

2.2.1 Stated preference

Stated preference (SP) data collection is useful when the route choice situation is hypothetical or dynamic. Conventional survey methods provide a few alternative paths with descriptions of each one (e.g., Bovy & Bradley 1985, Dia 2002, Hess, Rose & Hensher 2008). Recent technologies enable the respondents to make route choice decisions in a virtual environment created by computer graphics. Orlando Transportation Experimental Simulation Program allows the respondent to drive virtually in a transportation network, and records the route choice decisions influenced by real time traffic information (Abdel-aty & Abdalla 2006). Everscape is a 3D multi-user computer game in which players evacuate from one location to another with options of different modes and paths (Doirado, van den Berg, van Lint, Hoogendoorn & Prendinger 2012). The individual choice decisions are recorded, and can be used to study their multimodal route choice behavior under disaster evacuations. Nonetheless, route choice modeling from SP data is relatively standard, as the complexity of the choice set is controlled in the experimental design.

2.2.2 Traditional revealed preference methods

Before GPS technology became available, few studies on route choice behavior were based on RP data. Traditional RP surveys ask travelers to describe the chosen paths via mail, telephone or computer assisted tools. Some surveys use questionnaires to collect characteristics of the traveler and attributes of the trip, rather than the actual route (e.g., Ben-Akiva, Bergman & Daly 1984, Mahmassani, Joseph & Jou 1993, Abdel-Aty, Kitamura & Jovanis 1997). Other surveys ask the travelers to report locations where they have passed by during their journey. For example, Ramming (2002) ask respondents to report streets' names, and collect a route choice observation from 157 commuters. Vrtic, Schüssler, Erath, Axhausen, Frejinger, Bierlaire, Stojanovic, Rudel & Maggi (2006) collect intermediate cities of long distance trips, and get 940 trips from different individuals. Prato, Bekhor & Pronello (2011) design a web-based interactive map, and respondents indicate the sequence of junctions of their commuting trips. 575 trips are recorded from 236 individuals. These route choice observations include incomplete trip descriptions.

Therefore, the gaps must be filled by connecting adjacent reported locations based on assumptions such as shortest or fastest paths.

2.2.3 GPS capable devices

Passive GPS data collection method has revolutionized the RP survey for travel behavior study. The automatic and real time data recording is more reliable, compared to traditional methods that rely on respondents' memory. Respondents' willingness to participate in the survey is a smaller issue, because carrying a GPS device is not a heavy burden. Especially with smartphones, people do not even have to remember to bring an additional device. They also manage the tasks of charging it, at least as well as for any special survey device. Murakami & Wagner (1999), Jan, Horowitz & Peng (2000), and Wolf, Oliveira & Thompson (2003) provide more numerical evidences for favoring GPS data in quantitative travel behavior studies. In the recent decade, more and more route choice models are produced using GPS data. To mention a few, Broach, Dill & Gliebe (2012) model cyclist route choice behavior; Murakami & Wagner (1999), Jan et al. (2000) and Li, Guensler & Ogle (2005) model route choice behavior of drivers.

Embedded with various sensors, such as GPS and accelerometer, smartphones can be utilized to understand the users' context. They become popular as data collection tools in studying mobility patterns and transportation network performances. González, Hidalgo & Barabási (2008) learn from 100,000 mobile phone users' positions that human mobility is not random, but has a high degree of spatial and temporal patterns; Jenelius, Rahmani & Koutsopoulos (2012) use low frequency GPS data to estimate road travel time; Due to low-cost sensors' poor performances and various practical constraints, the smartphone data are usually sparse and inaccurate. For example, data recording interval for GPS is usually set to be quit large (e.g. 10 seconds), and the smartphone GPS data are not accurate. Moreover, retrieving and synthesizing information from various sensors is also challenging.

There are two potential sources of biases that can be introduced in existing route choice modeling methods using GPS data. First, a particular challenge related to GPS data is its inaccuracy, due to the constraints of the technology. GPS points are often recorded off-road, and they are usually deterministically map-matched to the transportation networks. Thus potential bias can be introduced due to wrong matchings. Another issue of RP route choice data, collected from both traditional methods and GPS devices, is incompleteness. Indeed, usually a fixed time interval is configured for sampling from the GPS sensor, and it ranges from 1 second to several minutes. Current transport studies deal with incompleteness by connecting reported or map-matched locations with some simplistic assumptions. For example, shortest path is usually used in most of the

map-matching procedures (e.g., Schuessler & Axhausen 2009b, Oshyani et al. 2012). Negative exponential (Hunter, Abbeel & Bayen 2012) or normal (Liao, Patterson, Fox & Kautz 2007) is sometimes assumed so as to fit into the mathematical form of some machine learning methods. However, these assumptions may not correspond to the real behavior. Consequently, bias could be introduced.

2.3 Obtaining route choice observations from smartphone and GPS

In order to provide useful information for travel behavior studies, the mobility history has to be recovered from the raw smartphone data. For route choice modeling, the transport modes and the paths of trips need to be learned. Traditionally, transport mode inference and path detection (a.k.a map-matching, MM) are applied to GPS data only, and consist of two steps (e.g., Schuessler & Axhausen 2009a): first, split the data into multiple unimodal segments, and infer the transport mode of each segment; second, perform MM for each segment independently.

2.3.1 Transport mode inference

Dedicated GPS devices provide good quality GPS data in terms of high accuracy and high density. Speed, acceleration and deceleration can be calculated from dense GPS coordinates. Deterministic rule based methods distinguish modes by some predefined deterministic criteria. Bohte & Maat (2009) calculate the average speed from a sequence of GPS data to determine the transport mode of the corresponding journey. Stopher, Clifford, Zhang & Fitzgerald (2008) use more criteria, including 85th percentile of speed, acceleration and deceleration. They also use rail and ferry networks to recognize these two modes. Chung & Shalaby (2005) determine the sequence of modes from a predefined set of reasonable mode change chains. Deterministic rule based methods suffer from data outliers, therefore, possibilistic and probabilistic approaches are often proposed. Schuessler & Axhausen (2009a) use fuzzy logic to identify modes from speed and acceleration. Machine learning methods are convenient because transport mode inference can be treated as a standard classification problem. Zheng, Liu, Wang & Xie (2008) apply decision trees, Baysian networks, support vector machine and conditional random fields to speed data, and conclude that decision trees performs the best.

Smartphone provides more sparse and less accurate GPS data, but records more kinds of data such as ACCEL. Machine learning methods are particularly convenient when there are multiple kinds of data, hence they are widely used. They can exploit many features of the raw dense and noisy ACCEL data: mean, variance, fast Fourier transform,

time between peaks (Ravi, Dandekar, Mysore & Littman 2005, Nham, Siangliulue & Yeung 2008, Kwapisz, Weiss & Moore 2010, Ding, Zhang & Wang 2010) They can also fuse multiple kinds of data to infer the status of the phone carrier. For example, Reddy, Burke, Estrin, Hansen & Srivastava (2009) use speed data from GPS, acceleration data from accelerometer, and apply naive Bayes, decision trees, k-nearest neighbor, support vector machine, and hidden Markov model to classify transport modes. Stenneth, Wolfson, Yu & Xu (2011) also use the proximity to bus stops and train stations, and they have tested naive Bayes, Bayesian network, decision trees, random forest, and multilayer perceptron.

Rule based methods are simple, but they are not robust with respect to data outliers. Machine learning methods aim at fast inference. But the problems often have to be modeled or simplified in a way that fits into the standard frameworks. This constraint results in that the proposed methods might not be able to systematically or correctly capture how the data is produced. Thus errors introduced in this procedure are difficult to identify, and may result in biases in travel behavior studies. Both kinds of methods produce deterministic inference results. To the best of our knowledge, BT data has not been used in transport mode inference. But as discussed in last section, BT data can also help to recognize the public transport environment.

2.3.2 Unimodal map-matching of a segment using GPS

The raw GPS data are usually matched to the transportation network in order to be useful for many applications. In particular, navigation systems motivate the study of such MM techniques. A comprehensive review of 35 MM algorithms for navigation applications since 1989 is presented by Quddus, Ochieng & Noland (2007). And a validation strategy for MM algorithms is proposed by Quddus, Noland & Ochieng (2005). Since they are designed for navigation applications, current MM algorithms aim at providing on-line deterministic identification of the real road from a single GPS point. However, they don't guarantee that detected roads are connected to form a meaningful path, even if some MM algorithms (e.g., Greenfeld 2002, Ochieng, Quddus & Noland 2003) do consider connectivity and contiguity of the arcs. In some transport studies where on-line identification is not required and intensive computation is allowed, researchers are also interested in the actual path for the whole trip. For example, some novel navigation techniques learn "routing" strategies from GPS data recorded from experienced road network users (e.g., Yuan, Zheng, Zhang, Xie, Xie, Sun & Huang 2010); "route" travel time can be estimated from GPS data recorded from floating cars in the transportation network (e.g., Ebendt, Sohr, Tcheumadjeu & Wagner 2010). In route choice modeling, "path" observations are the input for route choice models (Bierlaire & Frejinger 2008).

The adaptation of multiple hypotheses technique (MHT) (Pyo, Shin & Tae-Kyung 2001)

in MM enables modelers to generate a connected path from a GPS trace representing geographical locations during a trip. Several algorithms (e.g., Marchal, Hackney & Axhausen 2005, Schuessler & Axhausen 2009b) maintain at each GPS point a set of path candidates. For each candidate, a score is calculated based on the dissimilarity between GPS points and arcs in terms of distance, speed and/or heading difference, though heading was found to be unreliable for this application (Schuessler & Axhausen 2009b). The work by Schuessler & Axhausen (2009b) focuses on the computational efficiency of the MM method, and shows excellent results along that line, with dense and accurate GPS data. However, from the experiments that we have conducted (see Section 3.2.4), it appears that the method is not suitable for smartphone data, where the focus should be in managing the inaccuracy and low density of the data. Moreover, the scores calculated by MM algorithms with MHT techniques, while often heuristically effective, in general lack the theoretical foundation necessary to serve as the probabilities that the corresponding paths are the true path. The simplicity of the score calculation can not ensure its correctness if there are data outliers. Moreover, in such a post-processing algorithm (as opposed to real time algorithm for navigation tools), "inaccurate" data is eliminated in the process of data filtering (Schuessler & Axhausen 2009a), with the risk that some useful information is also excluded. Conditional Random Field (CRF) based method, proposed by Hunter et al. (2012), aims at a fast inference procedure via CRF. A "driver model" is constructed to capture the driver's movement in the network, but its specification tends to favor shorter paths. Probabilistic MM algorithms in the literature rely on Dead Reckoning equipping cars or other sensors that smartphones don't embed (e.g. Ochieng et al. 2003).

2.3.3 Multimodal map-matching of entire trips

The two step technique poses a high risk of yielding wrong results, because potentially wrong segmentations in the first step are not recoverable. Many algorithms assume that walking is necessary for a transition between different modes, and they rely on dense GPS data (1 second interval) to detect the mode (e.g., Zheng, Li, Chen, Xie & Ma 2008, Zhang, Dalyot, Eggert & Sester 2008, Schuessler & Axhausen 2009a). The validity of this assumption is questionable, especially for smartphone data, because GPS data could be missing due to the unavailability of the GPS signal while walking indoor. Moreover, due to the sparsity of the GPS data on a smartphone, they may not provide sufficient information for a proper segmentation and mode inference.

An integrated particle filter modeling framework for simultaneously detecting transportation modes and traveling roads is proposed by Liao et al. (2007). In their approach, a *state* combines various mobility patterns, including the transportation mode and the

current road. A Rao-Blackwellized particle filter is used as the framework, while the probability of the traveler switching from one mode to another depends on his proximity to available transportation facilities. A Kalman filter is used to model the dynamic process of traveling on the network and retrieving the GPS fix. In order to fit in the Kalman filter framework, a great deal of simplification is required.

2.4 Route choice set specification

In real transportation networks, the number of paths that connect a given pair of origin and destination (OD) is huge and cannot be enumerated. Therefore, the identification of the universal choice set is not feasible (Bovy 2009, Prato 2009). Hence, a subset of path alternatives are usually generated for the model estimation. These alternatives should contain relevant paths for the traveler in order to enable the discrete choice model estimator to correctly identify her compensatory decision making process. Similar to choice model structures, a good compromise between tractability and behavioral relevance is also needed in the path generation procedure.

2.4.1 Consideration set

Some methods are based on behavioral assumptions and assume that a traveler only considers a number of attractive alternatives in their "consideration set" when she makes route choice decisions. The formation of the consideration set is a different mental process, before the actual choice from considered alternatives is made (Bovy 2009). Hence, a route choice set generation algorithm is designed to build the consideration set. For example, shortest-path based algorithms (e.g. Ben-Akiva, Bergman, Daly & Ramaswamy 1984, Azevedo, Santos Costa, Silvestre Madeira & Vieira Martins 1993, de la Barra, Perez & Anez 1993) assume that people consider the shortest paths in terms of some generalized cost functions; two stochastic variants (Ramming 2002, Bovy & Fiorenzo Catalano 2007) randomly perturb the parameters and attributes of the cost function according to predefined distributions, and select the shortest paths repeatedly; constrained enumeration approach uses branch-and-bound to generate all paths satisfying some constraints (Friedrich, Hofsaess & Wekeck 2001, Prato & Bekhor 2006). Although these algorithms are motivated by behavioral assumptions, the modelers hardly get information to validate that the generated choice set is the actual consideration set. Actually, it has been often reported that the chosen alternative does not even belong to the choice set (Ramming 2002, Bekhor et al. 2006).

2.4.2 Importance sampling

Recent importance sampling approaches do not model the consideration set. Instead, they assume the choice set to include all paths. Although it is clearly not consistent with behavior, it guarantees that no important path for the decision maker is omitted. In order to make the model tractable, a subset of paths are sampled for the model estimation using importance sampling. Importance sampling of alternatives is just an intermediate statistical procedure for the model estimator, unlike the "consideration set" that captures a separate mental process. To obtain consistent estimates, the sampling bias must be corrected in the model specification. The random walk (RW) algorithm, proposed by Frejinger, Bierlaire & Ben-Akiva (2009), is the first importance sampling algorithm that provides consistent model estimators in route choice context. They report unbiased parameter estimates using a reasonable number of path samples in a synthetic route choice experiment. However, due to the structure of the sampling probability definition, which is link multiplicative, the RW algorithm can not avoid cyclic paths. Paths sampled from real networks are often found to have many loops (Schüssler 2010). These cyclic paths are in fact irrelevant to the decision maker, thus contribute very little to the identification of the parameters. As a result, in practice, the asymptotically consistent estimator needs prohibitively large samples of alternatives in order to achieve unbiasedness. Recently, Flötteröd & Bierlaire (2013) propose a Metropolis-Hasting path sampling (MHPS) technique that allows to sample acyclic paths from any given distribution. This technique offers more flexibility than the random walk in terms of sampling distributions.

2.5 Discussions

Although smartphone is a convenient data collection tool, inferring route choice data is critical. Fusing various kinds of data is difficult, so usually standard machine learning methods are used. Most of the existing methods deal with transport mode inference and map-matching in two different stages. The risk of this approach has been identified. They provide deterministic mode inference and map-matching results, and bias can be introduced to route choice models if the results are wrong. Bierlaire & Frejinger (2008) propose a new discrete choice modeling framework for route choice data that is not associated with the transportation network ("network-free" data). They suggest that the error in the "network-free" data should be treated in a probabilistic manner. Although they also suggest that the methodology may be applied to GPS data, they present only results based on interview data.

Importance sampling of alternatives provides an alternative way of specifying the route

choice set. Especially the recent MHPS technique provides a flexible way of designing a sampling algorithm.

3 Probabilistic map-matching for smartphone GPS data

An important feature of most GPS capable cell phones is Assisted-GPS, which reduces warm-up time for getting the first GPS reading to seconds. This advantage provides more opportunities to observe full tracks of the user's trips without losing the beginning parts of trips. However, the GPS device consumes a great deal of energy. Due to practical constraints, such as limited phone storage space and expensive data transmission cost, data cannot be recorded at a high rate. *EPFLScope* specifies a time interval of 10 seconds. Also, the data is not as accurate as those collected from dedicated GPS devices. For instance, in Nokia N95, the GPS antenna is embedded under the keyboard, which is generally covered by the screen when the phone is not being actively used. Furthermore, most people carry the cell phone in their pocket or handbag. This weakens the GPS signal.

We conducted an experiment where a N95 smartphone and a dedicated GPS device (a MobilityMeter, of the type used by Flamm, Jemelin & Kaufmann (2007)), were both carried by the same person during a day. Both devices are configured to record GPS fixes with 1 second interval. The two tracks are reported in Figure 3.1, where the blue circles (appearing darker on a black-and-white copy) represent the tracks provided by the MobilityMeter, and the red x's (appearing lighter on a b&w copy) represent the tracks provided by the N95 smartphone. Figure 3.1a shows 6083 points from N95, and 12165 points from MobilityMeter. The availability rate of N95 is 88.7%, while that of MobilityMeter is 99.0% [1]. Throughout this thesis, the transportation network data used is provided by *OpenStreetMap* (www.openstreetmap.org), which is an open source map data service. Although statistical investigation of GPS data accuracy (e.g. Blewitt, Heflin, Webb, Lindqwister & Malla 1992, Wing, Eklund & Kellogg 2005) is out of the scope in this thesis, we can still observe from the visualization that, intuitively, the MobilityMeter

[1] Warming time is not accounted in calculating the availability rate. If a device doesn't record data in more than 10 minutes, it is considered as 'off' and this time period is not accounted in calculating the availability rate.

GPS data are more consistent with each other in terms of continuity, while N95 GPS data are more scattered. Also, the MobilityMeter GPS data seem closer to the roads. The poor quality of smarpthone data precludes the use of state of the art deterministic map-matching (MM) methods, as an example shown in Section 3.2.4.

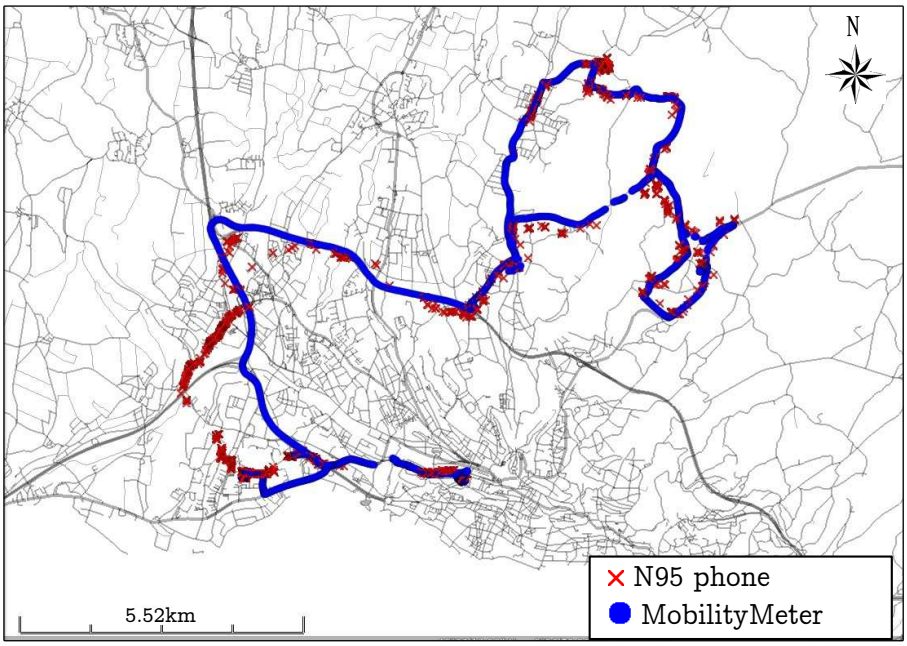
The best MM techniques generate a unique best fitting path, but in some applications a unique path is not required. One such application is route choice modeling with network-free data, as presented by Bierlaire & Frejinger (2008). They have introduced an estimation procedure for route choice models that accepts a probabilistic representation of the observed paths, accounting for errors in measurement. An observation does not need to be a unique path, but can be represented by a set of potential paths, along with a probability that the location measurements are indeed recorded from each path.

This chapter proposes and implements an advanced and practical probabilistic MM algorithm. It takes advantage of both geographical and temporal information in GPS data to measure the likelihood that the data has been generated along a given path. The likelihood measurement accounts for the inaccuracy of both the smartphone GPS data and the representation of the underlying transportation network. The proposed path generation procedure is capable of dealing with the sparsity of the smartphone GPS data. This method can reduce the impact of noise in GPS readings, and provides probabilistic path observations for further applications.

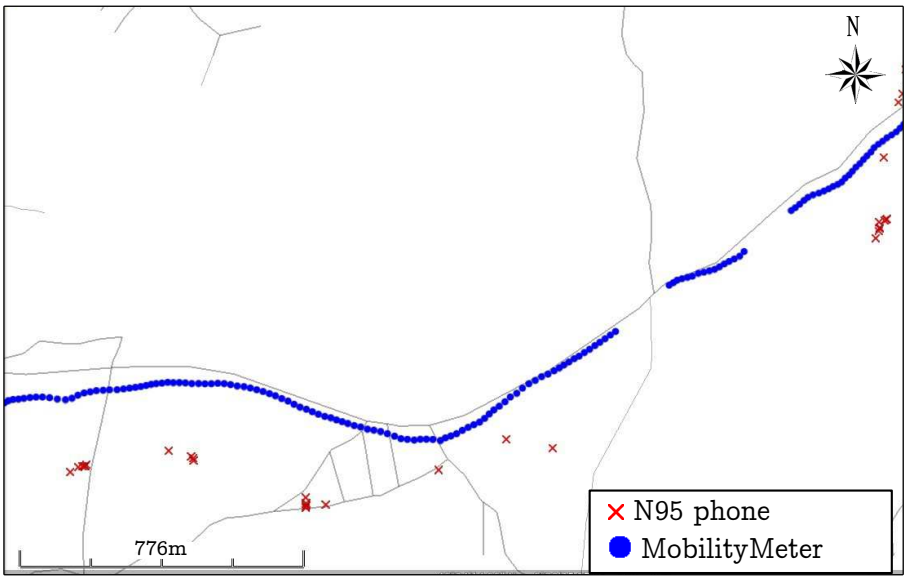
The next section introduces the GPS data recorded from the smartphones, and the context where the data was recorded. Section 3.2 derives the probabilistic measurement model for measuring the likelihood that a GPS trace is recorded while traveling on a path. This model relies on a network performance model. Although stand-alone traffic simulators can be used, a simple travel model using only information available from the GPS records is presented. The probabilistic measurement model is illustrated on some example paths with a real smartphone GPS trace. Potential paths need to be generated before their likelihoods can be calculated. As MM algorithms are not suitable for the smartphone GPS data, a new path generation algorithm, accounting for the sparsity of the smartphone GPS data, is proposed in Section 3.3. The proposed approach is applied on 25 real smartphone GPS traces, and some examples are illustrated. In Section 3.4 we perform sensitivity analyses on model parameters and GPS sampling interval. Some conclusions are included in Section 3.5.

3.1 Context and data

Let $G = (N, A)$ denote a transportation network, where N is the set of all nodes and A is the set of all arcs. The horizontal position of each node $n \in N$ is represented by



(a) in a region



(b) zoom in

Figure 3.1: GPS traces from N95 and a GPS device

$x_n = (\text{lat}, \text{lon})$, which is a pair of coordinates consisting of latitude and longitude. The shape of the physical route of arc a is described by an application

$$\mathcal{L}_a : [0, 1] \to \mathbb{R}^2. \tag{3.1}$$

For a point on the arc, its position x is generated from a unique number ℓ between 0 and 1 such that $x = \mathcal{L}_a(\ell)$. In particular, $\mathcal{L}_a(0)$ is the coordinates of the up-node, and $\mathcal{L}_a(1)$ is the coordinates of the down-node of arc a. For example, if the arc is a straight line from node u to node d, then

$$\mathcal{L}_a(\ell) = (1 - \ell)\, x_u + \ell x_d. \tag{3.2}$$

Indeed, straight lines are used in transport network data to represent arcs in practice. The performance of the network is characterized by a model

$$x = S(x^-, t^-, t, p) \tag{3.3}$$

predicting the position x at time t of an individual at position x^- at time t^-, and following path p. It is a random variable with probability distribution function

$$f_x(x|x^-, t^-, t, p). \tag{3.4}$$

Typically, this model is obtained from a calibrated traffic simulator. However, for practical purposes, analytical models can also be used (see Section 3.2.3).

EPFLScope triggers a GPS reading event every 10 seconds. A GPS measurement

$$\widehat{g} = \left(\widehat{t}, \widehat{x}, \widehat{\sigma}^x, \widehat{v}, \widehat{h}\right),$$

is extracted for each GPS reading, and it contains:

- $\widehat{t}$, a time stamp ; $\widehat{x} = (\widehat{x}_{\text{lat}}, \widehat{x}_{\text{lon}})$, a pair of coordinates;
- $\widehat{\sigma}^x$, the standard deviation of the horizontal error in the location measurement;
- $\widehat{v}$, a speed measurement (km/h) and,
- $\widehat{h}$, a heading measurement, that is the angle to the north direction, from 0 to 359, clockwise.

Sometimes, the GPS sensor fails to measure the speed and heading values for a measurement, and it reports the exact same values as in the previous measurement. In this case, the speed and heading values of the measurement have to be calculated. If we denote

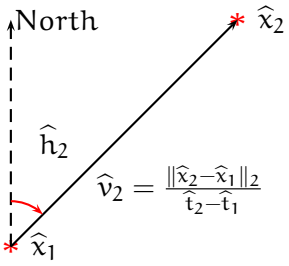


Figure 3.2: Calculate speed and heading

two consecutive GPS measurements as $\widehat{g}_1$ and $\widehat{g}_2$ respectively, Figure 3.2 shows how to calculate $\widehat{v}_2$ and $\widehat{h}_2$.

We assume that the data has been preprocessed so that we have access to a sequence of measurements $(\widehat{g}_1, \ldots, \widehat{g}_T)$, which is abbreviated as $\widehat{g}_{1:T}$[2], corresponding to a given trip.

The experiments described in this chapter use smartphone data extracted from the LDCC database. Dataset A, presented in Section 3.2.4, contains only one GPS trace with 10 points. It has been collected by one of the authors, with known true path; Dataset B, presented in Section 3.4, contains 25 GPS traces and has been collected by 3 anonymous individuals, without known true paths. These GPS traces are recorded while the users are traveling in urban and outskirt areas.

3.2 Probabilistic measurement model

In this section, we focus on the derivation of the probabilistic measurement model for a set of GPS data. More precisely, we compute the likelihood of observing GPS points $\widehat{g}_{1:T}$ on a hypothetical path p at time $t_{1:T}$ respectively:

$$\Pr(\widehat{g}_{1:T} | t_{1:T}, p).$$

We assume that the time is recorded without error. Therefore, the model will return a non zero probability only when the sequence $t_{1:T}$ exactly matches the sequence of time stamp $\widehat{t}_{1:T}$ in the data. This probability is an essential input for the network-free data modeling approach (Bierlaire & Frejinger 2008). In this section, we introduce a new modeling framework to derive this model and its components.

[2]This thesis deals with data sequences, and the notation follows this abbreviation convention throughout the thesis.

3.2.1 Measurement equations

We now derive the probability that a given path p generates the data $\widehat{g}_{1:T}$. For the sake of simplification, we focus on the measurement equation for the locations $\widehat{x}_{1:T}$, that is

$$\Pr(\widehat{x}_{1:T}|t_{1:T}, p), \tag{3.5}$$

which is decomposed recursively:

$$\Pr(\widehat{x}_{1:T}|t_{1:T}, p) = \Pr(\widehat{x}_T|\widehat{x}_{1:T-1}, t_{1:T}, p)\Pr(\widehat{x}_{1:T-1}|t_{1:T-1}, p). \tag{3.6}$$

The recursion starts with the model $\Pr(\widehat{x}_1|t_1, p)$:

$$\Pr(\widehat{x}_1|t_1, p) = \int_{x_1 \in p} \Pr(\widehat{x}_1|x_1, t_1, p)\Pr(x_1|t_1, p)dx_1, \tag{3.7}$$

where the integral spans all locations x_1 on path p. For the first point, if we do not have any prior on the location, $\Pr(x_1|t_1, p)$ is a constant equal to the inverse of the length L_p of p. The model $\Pr(\widehat{x}_1|x_1, t_1, p) = \Pr(\widehat{x}_1|x_1)$ describes the measurement error of the smartphone device.

It is generally assumed that the errors in longitudinal and latitudinal directions (e_{lon} and e_{lat} respectively) are independently normally distributed (van Diggelen 1998). Therefore, the distance between the true location and the measured coordinates $e = \sqrt{e_{lon}^2 + e_{lat}^2}$ follows a Rayleigh distribution. The probability that coordinates $\widehat{x}_1$ is recorded from a location x_1 is defined as the probability that the distance between x_1 and $\widehat{x}_1$ is less than the true error. Then, we have

$$\Pr(\widehat{x}_1|x_1) = \Pr(e > \|\widehat{x}_1 - x_1\|_2) = \exp\left(-\frac{\|\widehat{x}_1 - x_1\|_2^2}{2\widehat{\sigma}^2}\right). \tag{3.8}$$

As the variance σ^2 is unknown, we use $\widehat{\sigma}^2 = \sigma_{\text{network}}^2 + (\widehat{\sigma}_1^x)^2$ as an estimate, where $\sigma_{\text{network}}^2$ captures the difference between the coded network and the actual roads and paths, and $(\widehat{\sigma}_1^x)^2$ captures the measurement error of the GPS device. Quddus et al. (2005) explain that errors in network data effect the quality of the MM results, therefore the network error parameter σ_{network} is introduced here.

Combining (3.7) and (3.8), we obtain

$$\Pr(\widehat{x}_1|t_1, p) = \frac{1}{L_p}\int_{x_1} \exp\left(-\frac{\|\widehat{x}_1 - x_1\|_2^2}{2\widehat{\sigma}^2}\right) dx_1. \tag{3.9}$$

The next step of the recursion derives

$$\Pr(\widehat{x}_1, \widehat{x}_2 | t_1, t_2, p) = \Pr(\widehat{x}_2 | \widehat{x}_1, t_1, t_2, p) \Pr(\widehat{x}_1 | t_1, p). \tag{3.10}$$

$\Pr(\widehat{x}_1 | t_1, p)$ is defined by (3.9). For the first term, we have

$$\Pr(\widehat{x}_2 | \widehat{x}_1, t_1, t_2, p) = \int_{x_2 \in p} \Pr(\widehat{x}_2 | x_2, \widehat{x}_1, t_1, t_2, p) \Pr(x_2 | \widehat{x}_1, t_1, t_2, p) dx_2. \tag{3.11}$$

The first term $\Pr(\widehat{x}_2 | x_2, \widehat{x}_1, t_1, t_2, p) = \Pr(\widehat{x}_2 | x_2)$, is again modeling the measurement error of the device, and can also be defined by (3.8), combined with the same simplifications as described above. The second term predicts the position of the traveler at time t_2. It is written as

$$\Pr(x_2 | \widehat{x}_1, t_1, t_2, p) = \int_{x_1 \in p} \Pr(x_2 | x_1, \widehat{x}_1, t_1, t_2, p) \Pr(x_1 | \widehat{x}_1, p) dx_1. \tag{3.12}$$

The first term in (3.12) changes to $\Pr(x_2 | x_1, t_1, t_2, p)$, where we only need the true location x_1 and the time t_1 when the measurement $\widehat{x}_1$ is recorded. It models the movement of the traveler, which is captured by (3.3), so

$$\Pr(x_2 | x_1, t_1, t_2, p) = f_x(x_2 | x_1, t_1, t_2, p),$$

where f_x is the density function (3.4) of the travel model. The second term in (3.12) can be derived from Bayes rule:

$$\Pr(x_1 | \widehat{x}_1, p) = \frac{\Pr(\widehat{x}_1 | x_1, p) \Pr(x_1 | p)}{\int_{x_1} \Pr(\widehat{x}_1 | x_1, p) \Pr(x_1 | p) dx_1}.$$

As $\Pr(x_1 | p) = 1/L_p$ is constant for a given p, we have

$$\Pr(x_1 | \widehat{x}_1, p) = \frac{\Pr(\widehat{x}_1 | x_1, p)}{\int_{x'_1 \in p} \Pr(\widehat{x}_1 | x'_1, p) dx'_1} \tag{3.13}$$

which is a normalized version of (3.8). This completes the definition of (3.10).

The recursion in (3.6) requires that, at iteration k, the probability

$$\Pr(\widehat{x}_k | \widehat{x}_{1:k-1}, t_{1:k}, p) = \Pr(\widehat{x}_k | \widehat{x}_{1:k-1}, t_k, p)$$

is calculated. It can be generalized from the above derivation that

$$\Pr(\widehat{x}_k|\widehat{x}_{1:k-1}, t_k, p) = \int_{x_k} \Pr(\widehat{x}_k|x_k) \int_{x_{k-1}} \Pr(x_k|x_{k-1}, t_{k-1}, t_k, p) \Pr(x_{k-1}|\widehat{x}_{1:k-1}, p) d_{x_{k-1}} dx_k, \quad (3.14)$$

where $\Pr(\widehat{x}_k|x_k)$ is given by (3.8), and $\Pr(x_k|x_{k-1}, t_{k-1}, t_k, p)$ is the travel model $f_x(x_k|x_{k-1}, t_{k-1}, t_k, p)$. The last part of (3.14), $\Pr(x_{k-1}|\widehat{x}_{1:k-1}, p)$, is the posterior pdf of the true location x_{k-1} given observed GPS trace $\widehat{x}_{1:k-1}$ and path p. This distribution is not tractable, and we must simplify it, and replace it by

$$\Pr(x_{k-1}|\widehat{x}_{1:k-1}, p) \approx \Pr(x_{k-1}|\widehat{x}_{k-1}, p). \quad (3.15)$$

Therefore, we can use the same derivation that leads to (3.13) to obtain

$$\Pr(x_{k-1}|\widehat{x}_{k-1}, p) = \frac{\Pr(\widehat{x}_{k-1}|x_{k-1}, p)}{\int_{x'_{k-1} \in p} \Pr(\widehat{x}_{k-1}|x'_{k-1}, p) dx'_{k-1}}. \quad (3.16)$$

The derivation above involves many integrals over the full path. Although these integrals have low dimension, they can be cumbersome to compute, especially when the path p is long. In Section 3.2.2, we describe how to decompose the integrals, and to use the concept of Domain of Data Relevance (DDR) introduced by Bierlaire & Frejinger (2008) to simplify the computation.

3.2.2 Computing integrals

The measurement equations involve various integrals along a path p of the form

$$I = \int_{x \in p} f(x) dx, \quad (3.17)$$

that are complicated to compute in real applications. We describe here how to exploit the topology of the network to compute these integrals.

First, we decompose the path into arcs to obtain

$$I = \sum_{a \in p} \int_{x \in a} f(x) dx. \quad (3.18)$$

For each arc, we use the shape model (3.1) to obtain a unidimensional integral

$$\int_{x\in a} f(x)dx = \int_{\ell=0}^{1} f(\mathcal{L}_a(\ell))|\partial\mathcal{L}|d\ell, \tag{3.19}$$

where

$$|\partial\mathcal{L}| = \sqrt{\left(\frac{d(\mathcal{L}_a(\ell))_{lat}}{d\ell}\right)^2 + \left(\frac{d(\mathcal{L}_a(\ell))_{lon}}{d\ell}\right)^2}. \tag{3.20}$$

For example, if the linear model (3.2) is used, we have

$$|\partial\mathcal{L}| = \|x_u - x_d\|_2. \tag{3.21}$$

Second, we truncate the domain of the integrals to save computation time where negligible quantities are involved. For a given GPS observation $\widehat{x}$, Bierlaire & Frejinger (2008) define the DDR as the physical area where the piece of data is relevant. In our context, a point x is considered to be in the DDR of $\widehat{x}$ if following conditions are satisfied:

- the probability $Pr(\widehat{x}|x)$ is above a given threshold θ;
- if $\widehat{v} > 10$km/h, the difference between the GPS heading and the arc direction is less than 60 degrees. The arc direction is approximated as the direction from its up node to down node.

In our implementation, we have used a value $\theta = 0.65$. It corresponds roughly to points in a diameter of 100m when the σ parameter of the GPS device is 100m, and the σ for the network coding is assumed to be 30m. Indeed,

$$\exp\left(-\frac{\|\widehat{x} - x\|_2^2}{2\widehat{\sigma}^2}\right) \geq \theta$$

is equivalent to

$$\|\widehat{x} - x\|_2 \leq \sqrt{-2\left(\widehat{\sigma}\right)^2 \ln\theta},$$

and the upper bound 96.9 is obtained with $\theta = 0.65$ and $\widehat{\sigma} = 104.4 = \sqrt{100^2 + 30^2}$. This is illustrated in Figure 3.3, where the parts of arcs AB and AC represented by a solid red line are inside the DDR of the data point $\widehat{x}$.

Clearly, the value of the parameters should be adjusted to account for the features of the relevant application, and the quality of the associated data. Also, the complexity of the computation of the integrals increases with the size of the DDR. A large DDR

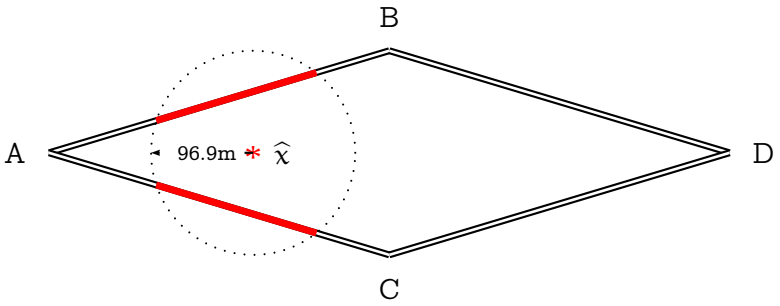


Figure 3.3: Domain of Data Relevance

means more computation. On the other hand, too small a DDR may artificially produce a zero probability for the measurement equation, which is undesirable. As discussed by Bierlaire & Frejinger (2008), the specification of the DDR should correspond to a good trade-off between accuracy and computational burden. Some sensitivity analyses regarding these parameters are performed in Section 3.4.

3.2.3 Travel model

In our framework, the travel model is designed to predict the position of the GPS device over time. More precisely, it predicts the position x_k of the device at time t_k if the position at time t_{k-1} is x_{k-1}, and the device is traveling along path p.

This is the typical role of dynamic traffic simulators (such as AIMSUN (Barceló & Casas 2005), MITSIM (Yang & Koutsopoulos 1996), DynaMIT (Ben-Akiva, Bierlaire, Burton, Koutsopoulos & Mishalani 2001), Dynasmart (Mahmassani 2001), among many). However, it is not always practical to use a calibrated traffic simulator in a MM context. Therefore, we suggest to use a simple analytical model such as the one described below.

First, we define the operator that computes the distance between two points x_{k-1} and x_k lying on path p, and denote it by

$$d_p(x_{k-1}, x_k). \tag{3.22}$$

This operator is easily implemented using the same decomposition of paths into arcs described in Section 3.2.2. We write the travel model in terms of speed instead of position, considering the random variable

$$v = \frac{d_p(x_{k-1}, x_k)}{t_k - t_{k-1}} \tag{3.23}$$

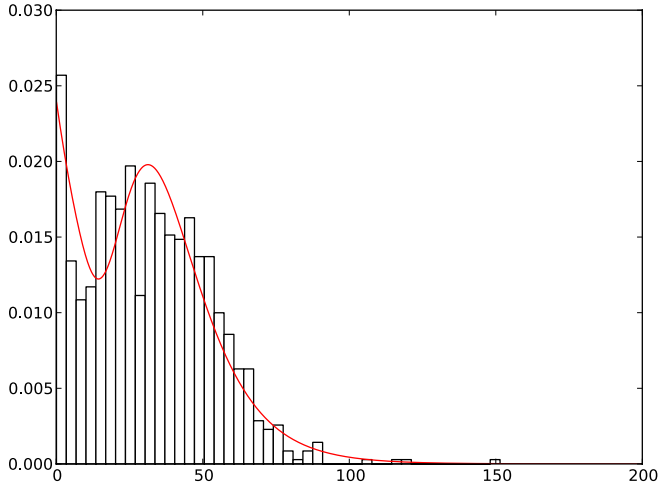


Figure 3.4: The speed distribution

with pdf

$$f_v\left(\frac{d_p(x_{k-1}, x_k)}{t_k - t_{k-1}}\right). \tag{3.24}$$

In our experiments, the traveling speed of the device is recorded every 10 seconds, therefore its distribution can be derived from the observed speed data. For the distribution of speed, we assume a mixture of a negative exponential distribution and a log normal distribution. The first is designed to capture the instances where the vehicles are stopped at intersections, or traveling at low speed before or after that stop. The second is designed to capture vehicles moving at regular speed. The distribution is

$$f_v(v) = w\lambda \exp^{-\lambda v} + (1-w)\frac{1}{v\sqrt{2\pi\tau^2}} \exp^{-\frac{(\ln v - \mu)^2}{2\tau^2}}, \tag{3.25}$$

where w (the weighting), λ (the scale parameter of the negative exponential distribution), μ (the location parameter of the log normal distribution), and τ (the scale parameter of the log normal distribution) are parameters to be estimated. Dataset B, including 1041 GPS records, is used for the estimation. Following are some statistics of dataset B: total number of GPS points (1041); number of GPS points per trace (minimum: 16, mean: 35.9, maximum: 53); duration of the trip (minimum: 180 seconds, mean: 387 seconds, maximum: 795 seconds). Figure 3.4 shows the normalized histogram of the recorded speed data and the estimated speed distribution. Table 3.1 reports the parameters estimated by maximum likelihood.

parameter	estimate	standard error
w	0.423	0.0636
λ	0.057	0.0097
μ	3.672	0.0314
τ	0.396	0.0282
Parameters estimated by R.		

Table 3.1: Parameters estimates for the speed distribution

3.2.4 Illustration

We illustrate likelihood results for 4 example paths associated with a real GPS trace (dataset A) recorded from a N95 smartphone. The 4 paths are shown in Figure 3.5 as red solid lines. The GPS points are also shown in each figure as blue points. The direction of the trajectory is illustrated by the arrow besides each path and by the the GPS points's annotation with their orders being recorded. This particular trip is chosen to be analyzed because it is recorded while traveling by car in a dense transportation network. And the path is known with certainty.

Figure 3.5a shows the true path. If we look at the GPS points, the ambiguity of the coordinates readings and the density of the transportation network makes the actual path difficult to be recognized from the N95 data alone. It can be observed that some of the GPS points (e.g. 7 and 8) deviate more than 30 meters from the actual path. Consequently, another path shown in Figure 3.5b also seems intuitively reasonable enough to be the actual path if we only compare the geographical dissimilarities. Another path candidate shown in Figure 3.5c is intuitively less possible to be the actual path. The last one (Figure 3.5d) seems very problematic, but is actually generated by the deterministic MM algorithm developed by Schuessler & Axhausen (2009b) (without using speed penalty term in the score function).

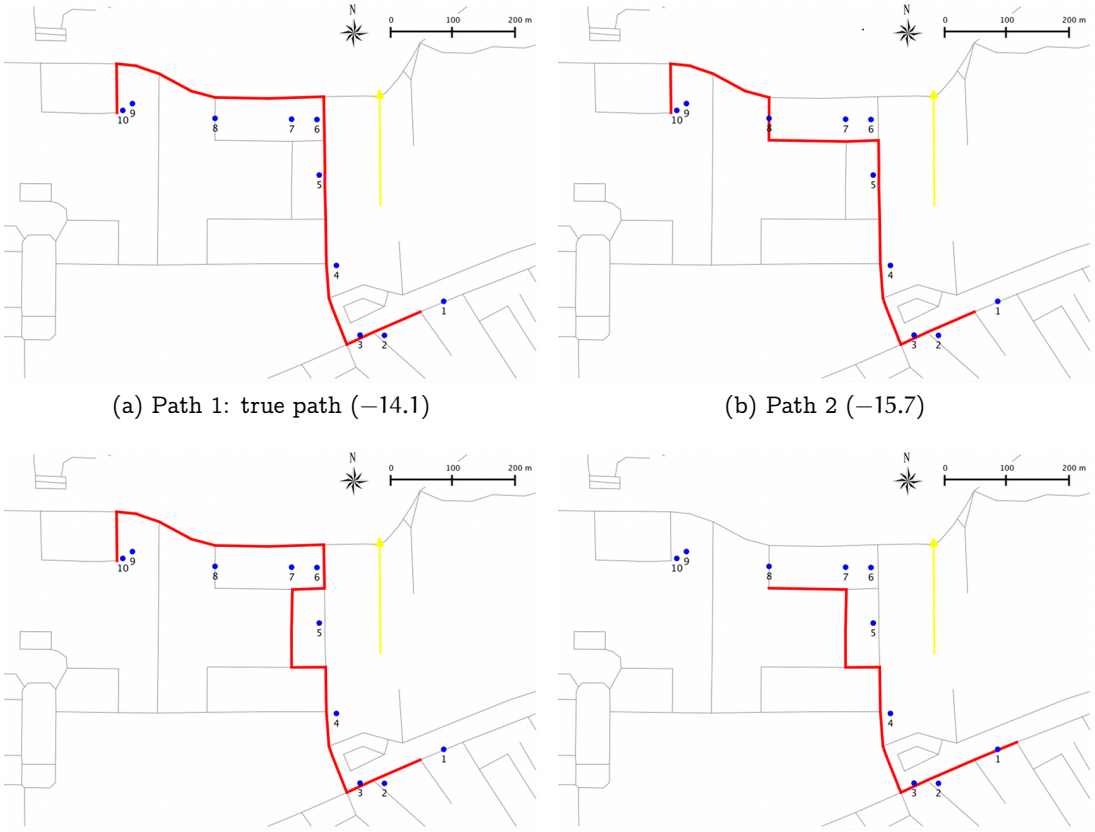
We calculate the natural logarithm of the measurement likelihood (3.5), termed the measurement loglikelihood, for all paths [3],

$$\ln \Pr(\widehat{x}_{1:T}|\widehat{t}_{1:T}, p), \tag{3.26}$$

where the time $\widehat{t}_{1:T}$ is directly taken from the GPS data.

We notice that the real path gains the highest loglikelihood, -14.1. The loglikelihood is

[3] If we further expand (3.6), the measurement likelihood (3.5) becomes $\Pr(\widehat{x}_{1:T}|t_{1:T}, p) = \Pr(\widehat{x}_1|t_1, p)\prod_{k=2}^{T}\Pr(\widehat{x}_k|\widehat{x}_{1:k-1}, t_{1:T}, p)$, which is the multiplication of many probability values that are smaller than 1. Consequently, the measurement likelihood (3.5) is close to zero. Throughout this thesis we present the logarithm of it, as it is common for likelihood.



(a) Path 1: true path (−14.1)

(b) Path 2 (−15.7)

(c) Path 3 (−16.3)

(d) Path 4 ($-\infty$)

Figure 3.5: Results from N95 GPS data

lower for paths that are intuitively less possible to be the true one (−15.7 and −16.3 for path 2 and path 3 respectively). The value for path 4 is $-\infty$, because the path does not pass through DDR of some GPS points (e.g. the last one).

Path 4 generated by the deterministic MM algorithm seems strange due to the incapability of the algorithm to deal with sparse data. In fact, the beginning of the path is correctly identified. The path terminates earlier than the real destination because the number of arcs in the path is constrained by the algorithm not to be higher than the number of GPS points. Indeed, the matched path has exactly 10 arcs, and it gains the lowest MM score among all paths with not more than 10 arcs. The drawback of this algorithm is described in Section 3.3 in details. In fact, if the GPS data has higher density, the deterministic MM algorithm may be able to identify the true path. For example, Figure 3.6 shows the MobilityMeter data recorded at the same time. The deterministic MM result produced by Schuessler & Axhausen (2009b)'s algorithm is the true path with this data.

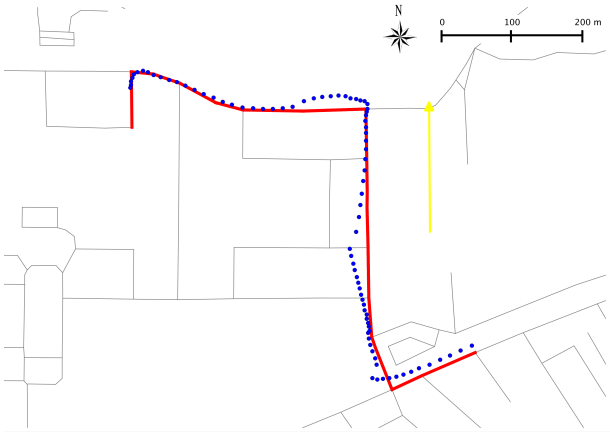


Figure 3.6: MobilityMeter data and deterministic MM result

3.3 Candidate path generation

The probabilistic measurement model $\Pr(\widehat{g}_{1:T}|t_{1:T}, p)$ calculates the likelihood of observing measurements $\widehat{g}_{1:T}$ on a given path p at time $t_{1:T}$. Given a set of candidate paths $\mathcal{P}$, it can be further used to infer how likely p is to be the true path:

$$q(p) = \frac{\Pr(\widehat{g}_{1:T}|t_{1:T}, p)\Pr(p)}{\sum_{p' \in \mathcal{P}} \Pr(\widehat{g}_{1:T}|t_{1:T}, p')\Pr(p')}, \tag{3.27}$$

This path probability can be used as a score function in traditional deterministic map-matching algorithms (e.g., Schuessler & Axhausen 2009b) for determining which path is the true one. Because we assume that the time is measured without error, the values of $t_{1:T}$ are taken directly from the data $\widehat{t}_{1:T}$. Considered as the prior probability, $\Pr(p)$ represents how likely p would be traveled without having the smartphone data. It actually models the smartphone user's route choice behavior. A multivariate extreme value (MEV) model estimated from historical data or external data sources can be used[4]. It might not provide precise route choice behavior, or can be very simple in model specification, but at least can give some basic information about the route choice preferences. If no route choice model is available, as in this thesis, the distribution is assumed to be uniform.

State of the art deterministic MM algorithms are designed for dense data, where it can be safely assumed that nearly every arc on a path generates at least one GPS point. For instance, Marchal et al. (2005) and Schuessler & Axhausen (2009b) generate path candidates by considering each GPS point one by one in the chronological order. At

[4] In fact, the choice probability for a MEV model contains a normalizing constant which requires path enumeration. If a MEV model is specified, the same constant appears in both the numerator and the denominator of (3.27). Thus, it is trivial to prove that this normalizing constant cancels out, and the path enumeration is avoided.

each iteration k, they generate a set $\mathcal{P}_k$ of path candidates assumed to match the GPS points up to k. These paths are generated by topologically extending the paths in $\mathcal{P}_{k-1}$ by not more than one arc. Hence, in order to allow for correctly identifying the true path, the GPS device has to record at least one GPS point on each arc. It can clearly be observed from Figure 3.1, Figure 3.5 and Figure 3.6 that the dedicated GPS device data is consistent with the "high density" hypothesis, while the smartphone data is not. Also, the example in Figure 3.5d shows that the deterministic MM algorithm is not appropriate for smartphone GPS data.

In order to address this problem, we propose a path generation algorithm designed for sparse data. It uses a similar iterative process as the one described above. But the path extension procedure at each iteration is not limited to only one arc. The algorithm ignores GPS points that have a speed lower than 8km/h, labeled as "stationary". When the device is more or less stationary, while it may generate data that is relevant for comparing path likelihood, it is not generating information that is useful in path extension. Two exceptions are the first and the last GPS points; even if their speed values are low, they reveal information about the origin and the destination. Therefore, they are not labeled as "stationary". This algorithm is described in Algorithm 1. Detailed explanation of some procedures (numbered lines) are given as follows:

12. The bound for the shortest path tree is derived from an assumption about the maximum possible speed and the time interval between t_{k-1} and t_k. The leaf nodes of the bounded shortest path tree are the first nodes detected by the Dijkstra algorithm that violate the bound. In our experiments, the bound is defined by $1.5(t_k - t_{k-1})\widehat{v}_{max}$, where $\widehat{v}_{max}$ is the maximum speed value among the observed speeds $\widehat{v}_{k-1}$, $\widehat{v}_k$, and the speed calculated by $\|\widehat{x}_k - \widehat{x}_{k-1}\|_2/(t_k - t_{k-1})$. The factor 1.5 is a safety margin to minimize the risk of missing a relevant observation.

21. In the update of likelihood, all the GPS points up to $\widehat{g}_k$, including "stationary" points, are included. As explained in Section 3.2.1, the likelihood (3.6) is updated recursively and at each iteration, only equation (3.14) needs to be calculated.

22. The path elimination procedure limits the size of $\mathcal{P}_k$ at each iteration if it is too large. After many tests, we use 20 as the threshold, which balances the tradeoff between algorithm speed and result effectiveness. It is designed to speed up the algorithm by eliminating less relevant branches produced from the path extension procedure. The path elimination procedure is performed by selecting and keeping following paths:

    1. The 2 shortest paths in $\mathcal{P}_k$ are selected.

Algorithm 1: Path generation algorithm

Input: A GPS trace $\{\widehat{g}_{1:T}\}$ with non-"stationary" GPS points.
Input: The underlying transportation network $G = (N, A)$.
Result: A set of candidate paths $\mathcal{P}_T$.

```
// Deal with the first GPS point.
P_1 ← empty set of paths;
DDR_1 ⇐ the DDR of the first GPS point;
for each arc a ∈ A do
    if a intersects DDR_1 then
        include a as a path in P_1;

// Iterative process.
for k ← 2 to T do
    P_k ← empty set of paths;
    foreach p ∈ P_{k−1} do
        // Path extension procedure.
        n ← the end node of p;
        spt ← a bounded shortest path tree rooted at n;
        foreach arc a ∈ spt do
            if a intersects DDR_k then
                sp ← shortest path connecting p and a;
                a_0 ← first arc of sp;
                a_1 ← last arc of p;
                if a_0 ∈ DDR_k or a_0 is not reverse of a_1 then
                    p_new ← join p, sp and a;
                    include p_new in P_k;
                    update likelihood Pr(x̂_{1:k} | t̂_{1:k}, p_new);
    if ‖P_k‖ > 20 then
        eliminate some paths from P_k;
```

2. Paths are randomly selected from $\mathcal{P}_k$ according to the probability (3.27). Path candidates are drawn using simulation until the cumulative normalized likelihood exceeds a predefined number (e.g. 0.8).
3. For each arc $a \in spt_k$ and a intersects DDR_k, $\mathcal{P}_{ak}$ is defined as $\mathcal{P}_k$'s subset that only contains paths going via a. We then apply a similar simulation procedure as in Step 2 on $\mathcal{P}_{ak}$, but only to draw one path. This is meant to guarantee that each arc associated with the latest GPS point has at least a path in $\mathcal{P}_k$ after the elimination procedure.

The algorithm is implemented as a software package in C++. We illustrate some results generated from 25 real GPS traces (dataset B). Figure 3.7 shows 4 examples, and 6 more examples are included in Appendix B. Again, each GPS trace is associated with many path candidates, and they overlap to a large extent, as shown on the maps. The results in general look reasonable, as each path is close to its corresponding GPS trace.

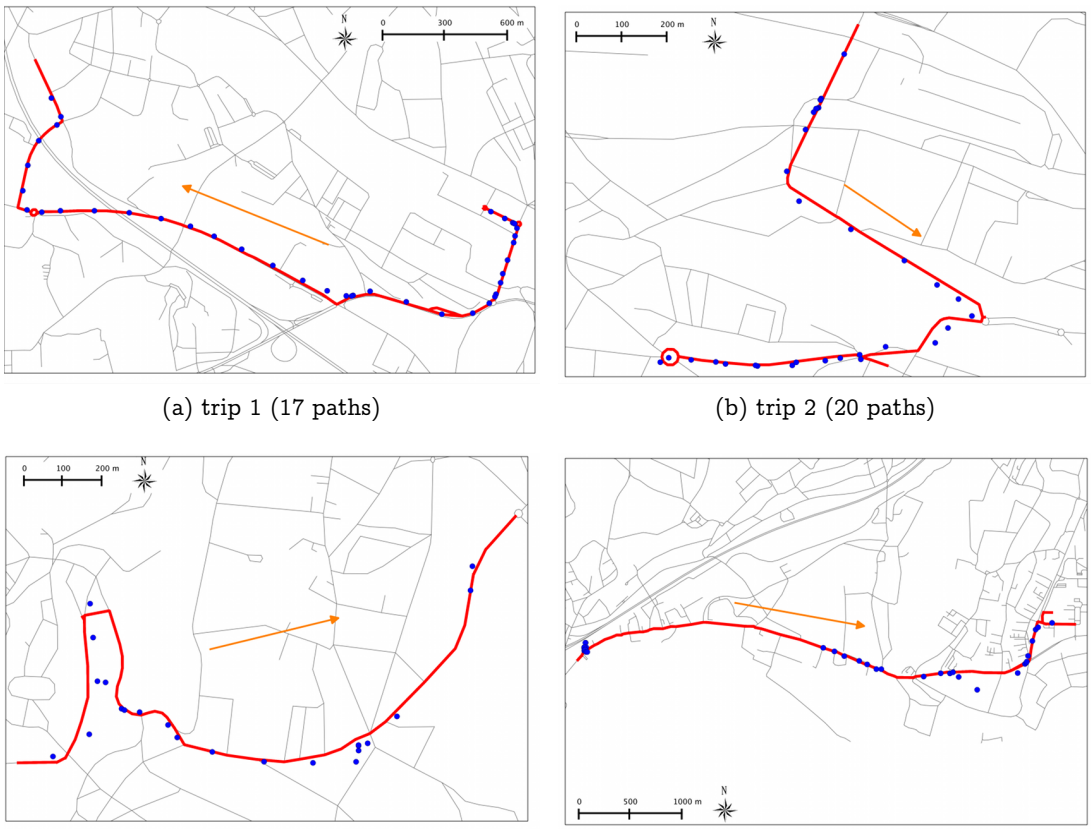
For the same trip, the differences of the generated paths show the uncertainty of the probabilistic map matching result. On one hand, the uncertainty is due to the imprecision of the GPS data. On the other hand, most of the uncertainty belongs to the end of the trips. This can be explained by the mechanism of the likelihood model. The likelihood model utilizes the dependency between adjacent GPS points. Each GPS point in fact provides information to help in identifying its upstream trajectory. The end of a trip always gains less information since it has less (or none) downstream GPS points.

3.4 Sensitivity analysis

In the probabilistic measurement model, some of the parameters' values are based on engineering intuition. These parameters are $\sigma_{network}$, the standard deviation of network error; θ, which defines the diameter of the DDR; and the heading constraint (60 degrees) for excluding arcs from the DDR. Sensitivity analyses, presented in on these parameters to test the robustness of the proposed probabilistic MM approach to these somehow arbitrary values. A sensitivity analysis is also performed on the sampling interval of the GPS data.

3.4.1 Experimental design

For any of the above mentioned parameters, although its precise value is not easy to decide, a reasonable bound can be derived based on available information. Therefore, the sensitivity analysis is performed as applying the proposed approach on the same dataset with the parameter's value varying within these bounds, and analyzing how the



(a) trip 1 (17 paths)

(b) trip 2 (20 paths)

(c) trip 3 (18 paths)

(d) trip 10 (20 paths)

Figure 3.7: Examples for some GPS traces

variation affects the results.

A probabilistic MM result for a GPS trace is a set of paths with associated measurement likelihood values. It contains a lot of information which takes effort to read. Hence, we have to define aggregate indicators for the sensitivity analyses. Intuitively, all paths in a result should be almost the same, as have been illustrated with some examples in the last section. The differences among the paths can be understood as the uncertainty of the result, which is caused by the ambiguity in the GPS data. Hence, we first define an aggregate similarity indicator $O_1(P)$ to measure the overall overlapping of all paths P in a result. Second, we also want to compare results produced by different parameter values. Therefore, another indicator $O_2(\mathcal{P}_1, \mathcal{P}_2)$ is defined to measure the similarity (overlapping) between two sets of paths, $\mathcal{P}_1$ and $\mathcal{P}_2$.

We start by defining how one path p overlaps with all paths in a set P ($p \in P$):

$$O(p, P) = \begin{cases} 1 & \text{if } \|P\| = 1 \\ \sum_{a \in p} \frac{L_a}{L_p} \frac{\sum_{p' \in P} \delta_{ap'} - 1}{\|P\| - 1} & \text{otherwise} \end{cases}, \tag{3.28}$$

where $\|P\|$ denotes the size of the path set; $\delta_{ap'}$ is a dummy variable, valued 1 if path p' goes via arc a, and 0 otherwise. This definition is inspired by the concept of Path Size, which is widely used in route choice modeling (see Ben-Akiva & Bierlaire 2003) to measure how an alternative overlaps with other paths in the choice set. The O value, between 0 and 1, can be roughly understood as the proportion of the path p that overlaps with all paths in P. The more the overlapping, the higher the value. If p is the only path in P ($\|P\| = 1$), i.e. perfect overlapping in P, $O = 1$; if p doesn't overlap with any other path at all, $O = 0$.

Based on this definition, we simply take the average to measure P's overall overlapping:

$$O_1(P) = \frac{1}{\|P\|} \sum_{p \in P} O(p, P).$$

So O_1 also values between 0 and 1. And a higher value indicates a higher degree of overlapping in P. We expect this indicator close to 1 because all paths in a result should be similar. Indeed, The O_1 values for example results shown in Figure 3.7 are: 0.966, 0.952, 0,962, 0.963, which are all close to 1.We also expect that the uncertainty of the result is insensitive to parameter variations. Hence O_1 value should be stable with respect to parameter variations.

Similarly, O_2 indicator for comparing two path sets is defined as:

$$O_2(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{\|\mathcal{P}_1\|} \sum_{p \in \mathcal{P}_1} O(p, \mathcal{P}_2 \cup \{p\}).$$

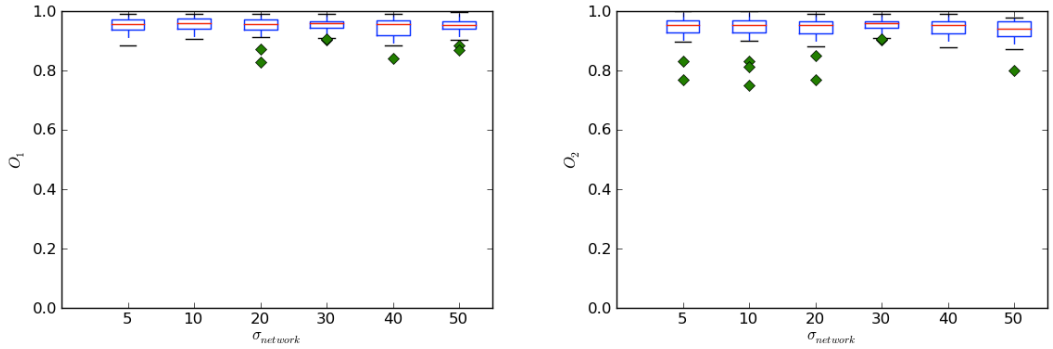
In particular, if $\mathcal{P}_1 = \mathcal{P}_2$, then $O_2(\mathcal{P}_1, \mathcal{P}_2) = O_1(\mathcal{P}_1) = O_2(\mathcal{P}_2)$. Similar to O_1, O_2 indicator also values between 0 and 1. A higher value indicates a higher degree of overlapping between the two path sets. If any path in one set does not overlap with any path in the other set, $O_2 = 0$. If all paths in $\mathcal{P}_1$ and $\mathcal{P}_2$ are identical, $O_2 = 1$. In the sensitivity analysis for a parameter, $\mathcal{P}_2$ is always set to be the path set produced by the default value. For example, in analyzing σ_{network}, $\mathcal{P}_2$ is fixed to be $\mathcal{P}_{\sigma_{\text{network}}=30}$, which is the path set produced by using $\sigma_{\text{network}} = 30$ (we follow the same notation convention throughout this section). Then, O_2 always indicates the overlapping of the paths generated from an alternative setting (e.g., $\sigma_{\text{network}=5}$) against $\mathcal{P}_{\sigma_{\text{network}}=30}$. We expect the proposed method is robust in the sense that results produced by different parameter values are similar to each other. Therefore, this indicator should have high value (close to 1) for any parameter value being used in the analyses.

3.4.2 Network error

There are two sources of the network error. First, the *OpenStreetMap* network data are collected from GPS devices, so the error in the GPS records is introduced. The amplitude of this error is difficult to be estimated because many people contribute to the network data and they use different kinds of GPS devices. A survey on commercial GPS devices suggests that the error from commercial GPS receivers is less than 15m in 95% of all cases (Ehsani, Buchanon & Salyani 2009). Second, in the network data, a road is represented as an abstract line without width. Therefore, we need to account for the width of the real road in the network error. This part of the error is also difficult to estimate because the details about the infrastructure are not easily accessible. In Switzerland, a third class road with one lane has minimum 2.8m width and a first class road with 2 lanes has minimum 6m width (Swisstopo 2011). A motorway has not more than 4 lanes per direction (according to Swiss motorway website, http://www.autobahnen.ch), and according to Switzerland standard ($3.20\text{m} - 3.75\text{m}$ width per lane, OFROU (2011)), the maximum width per direction is 15m.

Based on the above analysis, we believe that in most of the cases, σ_{network} is very unlikely to go beyond 50m or below 5m. So we perform a sensitivity analysis on σ_{network} with values 5m, 10m, 20m, 30m, 40m and 50m.

Figure 3.8 reports the distribution of O_1 and O_2 across the 25 trips in dataset B, using



(a) O_1 (b) O_2

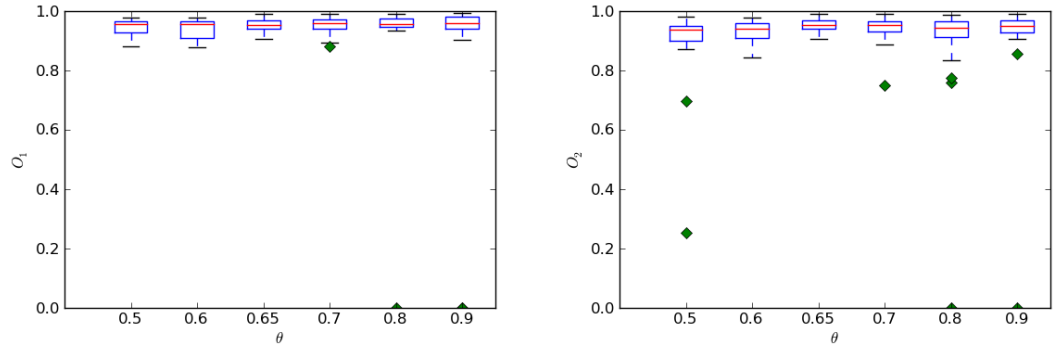
Figure 3.8: Sensitivity analysis for network error parameter

a boxplot representation. We can notice from this figure that O_1 is in general close to 1, which indicates a high degree of overlapping and a low degree of uncertainty in the path results. And this indicator is insensitive to the σ_{network} value variation. Moreover, no mater which parameter value is used, all the paths generated from the same trip should be, by the DDR definition, close to the GPS points, and hence, similar to each other. This is verified by Figure 3.8b, in which high value of O_2 shows that all the path results are similar to $\mathcal{P}_{\sigma_{\text{network}}=30}$. So, we can conclude that the probabilistic MM result is robust to σ_{network} variations.

3.4.3 The DDR diameter

The parameter θ defines the diameter of a GPS point's DDR. Blunck, Kjæ rgaard & Toftegaard (2011) conduct an experiment that studies smartphone (Google Nexus One and Nokia N95 used) GPS data accuracy using a high performance dedicated GPS device as the benchmark device. They report that in open-sky urban conditions, in the worst case, at least 90% of the smartphone GPS points have the error distance less than 60 meters and 100% of them have error distance less than 100 meters. Therefore, in our experiment, we assume the DDR diameter to be 100m.

We perform a similar sensitivity analysis as in Section 3.4.2 with θ to be 0.50 (123m), 0.60 (106m), 0.65 (97m), 0.70 (88m), 0.80 (70m) and 0.90 (48m) respectively (numbers in parentheses denote the corresponding diameters of the DDR), with 0.65 being the default setting. The O_1 and O_2 indicators are reported in Figure 3.9. Generally high values of O_1 and O_2 suggest that the results are robust with respect to the variations of θ. However, from both graphs, we notice that some indicators are close to or equal 0 with θ being 0.50, 0.80 and 0.90. Actually, they correspond to the trip shown in Figure 3.7d.



(a) O_1 (b) O_2

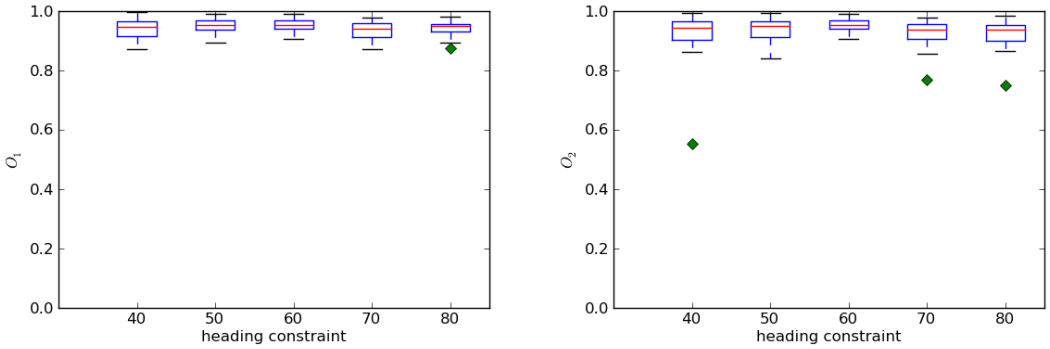
Figure 3.9: Senstivity analysis for θ parameter

This GPS trace is especially of low quality. It contains 15 stationary GPS points in the beginning of the trip (shown as a cluster in Figure 3.7d), and there is a huge gap in the GPS trace. It is intentionally selected to analyze the robustness of the algorithm. With θ being 0.50, 0.80 and 0.90, the algorithm fails to produce any reasonable path. But the results with θ around 0.65 (0.6, 0.65 and 0.7) are reliable.

3.4.4 Heading constraint

In the total 1041 GPS points used in our experiment, 896 of them have speed greater than 10km/h. For these 896 GPS points, the mean of the recorded standard deviation of the heading error is 2.85 while the maximum is 36. So we safely assume that the heading error does not exceed 60 degrees when the speed is greater than 10km/h. It forms a rule for excluding unreasonable arcs from the DDR. At low speed status, heading measurements from the GPS are generally not reliable. Hence, for GPS points with speed less than 10km/h, this heading constraint is not applied. Here, we analyze how much the variation of the heading constraint affects the result.

The analysis is performed on the values 40, 50, 60, 70, 80, where 60 is the default setting. The O_1 and O_2 indicators are reported in Figure 3.10. As expected, the results are robust with respect to the parameter value variation, in the sense that both indicators for most of the cases are close to 1. In Figure 3.10b, the outlier point in the boxplot for the constraint being 40 can be understood as an exceptional case when 40 degrees is too tight for few GPS points. Overall, we can conclude that 60 degrees is a suitable value for the heading constraint.



(a) O_1 (b) O_2

Figure 3.10: Senstivity analysis for heading constraint parameter

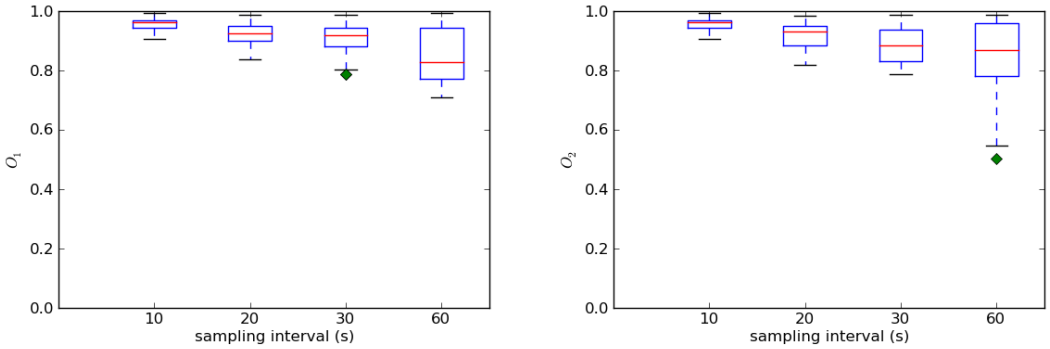
3.4.5 GPS sampling interval

The experiments are implemented for GPS data collected with sampling interval to be 10 seconds. However, we are also interested in applying the same method in more general situations. Therefore, we want to test the robustness of the method with respect to the data density. We artificially decrease the density of the data by manually increasing the GPS data interval to be $\kappa \in \{20, 30, 60\}$ seconds. The process is performed by selecting GPS points from the original data with following procedures:

1. select the first GPS point;
2. if the next GPS point is less than κ seconds later than the last selected GPS point, it is neglected; otherwise, it is selected;
3. repeat step 2 until the last GPS point.

We also perform a sensitivity analysis using O_1 and O_2 indicators. For the calculation of $O_2(\mathcal{P}_{\text{interval}=\kappa}, \mathcal{P}_{\text{interval}=10})$, the original data with interval 10s is truncated such that it has the same first and last GPS point as the processed data with interval κ. This is to guarantee that the GPS traces being compared have the same beginning and end, hence correspond to the same trip.

First, we notice that the algorithm fails to proceeds at some GPS points in some cases (2 cases for $\kappa = 20$, 4 for $\kappa = 30$ and 4 for $\kappa = 60$). The temporary path set $\mathcal{P}_k$ produced at a certain iteration k is empty in these cases. Figure 3.11 reports the O_1 and O_2 indicators for the successfully generated results. Figure 3.11a shows a trend that the larger sampling interval, the higher uncertainty of the path result. This is consistent with the intuition that more GPS data brings more information, thus less uncertainty.



(a) O_1 (b) O_2

Figure 3.11: Senstivity analysis for GPS sampling interval

Relatively high value of O_2 tells that results from different sampling interval settings are similar to the default setting. Since the paths generated from higher sampling interval are more heterogeneous, overall, they are less similar to the paths produced by the default setting, as reported by O_2 indicators. So, we conclude that the performance of the proposed algorithm decreases with the density.

3.5 Conclusions

We propose a probabilistic MM method for matching a set of paths with GPS data. A probabilistic measurement model is derived, which calculates the probability that a GPS recording device would have generated a sequence of measurements while following a given path. It is based on a structural model and a measurement model, which captures the movements and the recordings of the GPS device respectively.

The uncertainty derived from the inaccuracy of both the GPS data and the transportation network is explicitly taken into account. The application to real data shows that the probability values of the actual path and some other paths are realistic and meaningful.

A path generation algorithm is also proposed that accounts for the sparsity of the data. The methodology has been applied on real smartphone data collected in Switzerland. In the probabilistic measurement model, some of the parameters' values are based on engineering intuition. These parameters are σ_{network}, the standard deviation of network error; θ, which defines the diameter of the DDR; and the heading constraint (60 degrees) for excluding arcs from the DDR. Sensitivity analyses presented in Section 3.4 prove the robustness of the proposed probabilistic MM approach with respect to these somehow arbitrary values. A sensitivity analysis is also performed on the sampling interval of the

GPS data. The sampling interval, originally 10 seconds, is increased to be 20, 30 and 60 seconds, thus the data become sparser. The uncertainty in the path results increases with the sparsity of the GPS data. This is consistent with the intuition that more GPS data brings more information, thus less uncertainty.

4 Probabilistic multimodal map-matching with rich smartphone data

Last chapter deals with the map-matching problem when a trip is unimodal and the mode is known. And we have developed an implementation for the GPS data recorded from car trips. This chapter deals with a more general problem when a trip is multimodal, and the modes are unknown. We aim at identifying the traveled path and modes from various kinds of smartphone data.

The drawbacks of the two stage mode identification and map-matching approach have been discussed in Section 2.3. In this chapter, a novel algorithm is proposed to overcome these drawbacks. The data collected from a trip do not need to be segmented. The algorithm infers the physical path and the transport mode of each road simultaneously. This algorithm is called as *probabilistic multimodal map-matching*:

- Multimodal, because the output is a set of multimodal paths. Each arc on a path is associated with a specific transport mode. The transport modes on different arcs may differ (see Section 4.1 for the definition and an example of multimodal path).
- Probabilistic, because the algorithm generates a set of candidate paths, each of which is associated with a probability to be the true one.

Smartphone data is "rich" in the sense that more than one kind of data are available from various built-in sensors, including GPS, BT and ACCEL. For example, iPhone and Nokia 95, are usually embedded with a 3-axis accelerometer with $\pm$2G sensitivity. It has been found that ACCEL data from accelerometers are useful in recognizing the motion status of the phone carrier (e.g., Reddy et al. 2009, Kwapisz et al. 2010). Moreover, the BT sensor also provides valuable information about the smartphone's context. For example, the BT sensor detects more nearby BT devices in a public transport environment than those in a private mode. We propose a framework that can exploit rich smartphone data. Although only GPS, BT and ACCEL data are studied in this thesis, the method can be

extended to any type of sensor, such as gyroscope, if it provides information about the location or the transport mode.

The proposed algorithm is an extension of the unimodal map-matching algorithm proposed in Chapter 3, which is capable of dealing with both dense and sparse GPS data (time interval ranges from 1 second to 1 minute). In this chapter, a probabilistic measurement model is derived for each sensor to capture the data generation process. An integrated smartphone measurement model is constructed to integrate all sensor models in a unified framework. The smartphone measurement model calculates the likelihood of observing the smartphone measurements on a multimodal path. The travel model accounts for different transport modes and mode changes in the multimodal network. The candidate path generation algorithm deals with multimodal networks. Numerical experiments are illustrated. Finally, discussions and conclusions are given.

4.1 Smartphone data and transport networks

Two types of inputs are used for the proposed method: a model of the transport networks, and the smartphone data collected during travels.

4.1.1 Multimodal transport network and multimodal path

Different from the unimodal transport network defined in Section 3.1, another dimension, transport mode, is brought into the multimodal transport network. In a multimodal network, each arc $a \in A$ represents a road segment or a rail track segment, and accommodates one particular transport mode m. A road that can be traveled with bus and car, is represented by two arcs. A unimodal transport network G_m contains only arcs with the same transport mode m. In this chapter, we assume that the smartphone data are recorded while the carrier is traveling on a multimodal transport network. A multimodal transport network is represented by a union of several unimodal transport networks, and virtual arcs that connect them. This multimodal network representation is inspired by the *supernetwork* approach (Carlier, Fiorenzo-Catalano, Lindveld & Bovy 2003). A virtual arc is associated with a change of transport mode, and connects two nodes belonging to two different unimodal networks but having the same geographical location. This chapter models urban transport modes, private walk, bike, car, and public bus, metro.

A position $\mathbf{x} = (x, m)$ in a multimodal network is characterized by horizontal coordinates $x = (x_{lat}, x_{lon})$ consisting of latitude and longitude, and transport mode $m \in \{\text{walk}, \text{bike}, \text{car}, \text{bus}, \text{metro}\}$. A path is an ordered list of connected arcs. A multi-

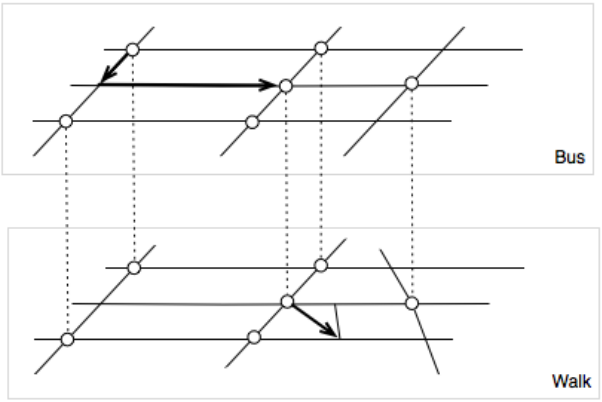


Figure 4.1: A multimodal network and a multimodal path

modal path p is a path in a multimodal network [1]. A multimodal path may contain only one single mode or several different modes. Figure 4.1 gives an example of a multimodal network with two unimodal networks (bus and walk), and a multimodal path. For the sake of clarity of the drawing, the arcs are represented bidirectional, and the directions are not drawn. Dashed lines represent virtual arcs that connect two unimodal networks. A multimodal path with direction starts from the bus network, and changes to walk via a virtual arc connecting $\mathbf{x}_c^- = (x_c, \mathrm{bus})$ and $\mathbf{x}_c^+ = (x_c, \mathrm{walk})$, where x_c denotes the coordinates of the mode change location.

OpenStreetMap is used as the source of the transport networks data. In this database, the transport mode accessibility on each road is specified and the public transport lines are also available. The *OpenStreetMap* data structure is only designed for visualization, and the PT network data have to be pre-processed for routing and map-matching usages. The metro stops are sometimes disconnected from other networks. We assume that people can access/egress them by walking from/to the nearest nodes, and create walking arcs to connect metro stops to the 5 nearest nodes. For the sake of simplicity, each arc is created as a straight line.

4.1.2 Smartphone data

When a sensor is activated, *EPFLScope* triggers sensor reading events periodically and logs the data. The availability of data is also subject to practical constraints. For example, GPS data are observed only if the GPS signal is available. Sometimes, the user may turn off the BT sensor. The raw sensor readings, e.g. a list of MAC addresses of nearby BT devices, are usually not ready to be used directly. So useful measurements need to

[1] In this chapter, the definition of path p is "multimodal", which is different from last chapter. By default , "path" refers to "multimodal path" in this chapter; "physical path" refers to a path without mode information.

be extracted from the raw data. This process is termed feature extraction in pattern recognition literature. The mechanism of reading sensors, and the feature extraction methods are explained below. The GPS data has been introduced in Chapter 3, and we focus on BT and ACCEL data here.

Bluetooth sensor *EPFLScope* configures the BT sensor to scan for nearby BT devices every 180 seconds. Each scan returns a list of nearby BT devices with their unique identifiers (MAC addresses). Nowadays, many people carry BT-enabled personal electronic devices, such as smartphones and tablets. These devices are visible to each other if, they are in a range of approximately 10 meters; and they do not move out of this range for a short time, which is about 1.92 seconds (Naya, Noma & Kogure 2005). The number of nearby visible BT devices varies with the context. In public transport, people are more compact in the vehicle, and they are stationary relative to each other. Hence a smartphone has a higher chance to observe more BT devices than in private transport. Therefore, we utilize the information about nearby BT devices in differentiating public/private transport context. A measurement $(\widehat{b}, \widehat{t})$ is extracted from each BT scan, and $\widehat{b}$ is equal to 1 if there is at least one BT device nearby, and 0 otherwise. It is also associated with a time stamp $\widehat{t}$.

Accelerometer sensor Accelerometer readings provide motion status of the phone user. It has been proposed in the literature to use them to detect the transport mode of the traveler (e.g., Reddy et al. 2009). A N95 smartphone is embedded with a 3-axis accelerometer with the sensitivity of ± 2G. An accelerometer reading is a triplet that contains the accelerations measured from 3 axes. The unit of the acceleration is $\frac{1}{280} m/s^2$, in which 280 is a normalization factor. *EPFLScope* triggers an accelerometer reading event every 120 seconds. Every reading event lasts for 10 seconds in a frequency of 40Hz. Therefore, it returns 400 accelerometer readings. Table 4.1 gives an example of data returned from a reading event.

We assume random orientation of the smartphones, and calculate the acceleration for each reading by taking the 2-norm of the triplet. Due to the high frequency of recording noisy acceleration data, an aggregation method is needed here. The aggregation takes a time resolution of 2 seconds, and split the 10 seconds data into 5 equal time windows. This aggregation technique is generally used in practice in order to reduce the noise in the acceleration data (e.g., 1 second time resolution is used by Reddy et al. 2009). In each time window, a measurement $(\widehat{a}, \widehat{t})$ is generated with $\widehat{a}$ as the mean of the accelerations in this time window. The measurement time $\widehat{t}$ is set to be the middle of the time window. Consequently, 5 accelerometer measurements are generated by each reading event.

Table 4.1: Acceleration data returned from a reading event

index	time stamp[a]	x-axis[b]	y-axis[b]	z-axis[b]
1	1272678293.03	15.0	-15.0	-324.0
2	1272678293.05	21.0	-15.0	-324.0
3	1272678293.09	21.0	-15.0	-329.0
...	...	...	...	...
398	1272678302.91	21.0	-15.0	-324.0
399	1272678302.94	15.0	-10.0	-319.0
400	1272678302.95	15.0	-15.0	-324.0

[a] Unix time stamp in seconds.
[b] Acceleration readings from 3 axes.

4.1.3 Measurements sequence from a trip

EPFLScope records data from independent sensors with a pre-defined schedule. We assume that the data have been preprocessed so that we have access to all the measurements recorded during a trip, and store them in a chronologically ordered sequence $\widehat{y}_{1:T}$, where T is the total number of measurements. The entire sequence is composed of 3 subsequences: the GPS, the BT, and the ACCEL. For example, we denote all the GPS measurements as $\widehat{g}_{1:I}$, where I is the total number of GPS.

Since only GPS provides valuable geographical location information, the measurements sequence $\widehat{y}_{1:T}$ is processed such that the first and the last measurements are GPS. All BT and ACCEL measurements recorded before the first GPS or after the last GPS are excluded. In $\widehat{y}_{1:T}$, if different types of measurements have the same time stamp, the order is set to be BT, ACCEL, and then GPS. If two GPS measurements have a large time gap, they do not provide reliable location information to BT and ACCEL data observed between them. Therefore, we decide to discard BT and ACCEL data if the time gap is large, and 20 seconds is chosen as the threshold.

Some sensor data (dataset C) with annotated transport modes are used to calibrate sensor measurement models and a speed distribution for each transport mode. These data are collected from 3 smartphone users while they are traveling with various transport modes. The true transport modes of the travels are known. The numerical experiments in this chapter use measurements sequences (dataset D) that are collected from 2 smartphone users while they are traveling in urban and outskirt areas. More details about the data will be provided in the corresponding sections.

4.2 Sensor measurement models

In this section, measurement models are defined to represent the sensors' operations in a multimodal transport network context. A measurement model has the form $\Pr(\widehat{y}|\mathbf{x}, t)$, where the state variable $\mathbf{x} = (x, m)$ is the position of the phone carrier in the network, and $\widehat{y}$ denotes a sensor measurement collected at time $\widehat{t}$. Conditional on the state $\mathbf{x}$, the measurement $\widehat{y}$ is derived for a value of t that equals to the time stamp $\widehat{t}$ of the measurement. Therefore, the model can be denoted as $\Pr(\widehat{y}|\mathbf{x})$. The rest of this section defines a sensor measurement model for each type of measurement, i.e. GPS, BT and ACCEL.

4.2.1 GPS measurement model

The GPS measurement model proposed here focuses on the location only, and is therefore denoted by $\Pr(\widehat{x}|\mathbf{x})$. It is an extension of the measurement model proposed in Section 3.2 with one more dimension, transport mode m, in the latent status variable $\mathbf{x}$. In this multimodal context, it is assumed that the error in the GPS coordinates is indepedent of the transportation mode,

$$\Pr(\widehat{x}|\mathbf{x}) = \Pr(\widehat{x}|x) = \exp\left(-\frac{\|x - \widehat{x}\|_2^2}{2\widehat{\sigma}^2}\right), \tag{4.1}$$

where $\|x - \widehat{x}\|_2$ calculates the distance (in meters) between the recorded coordinates $\widehat{x}$ and the coordinates x in the transport network; the variance $\widehat{\sigma}^2$ is approximated by $\widehat{\sigma}^2 = \sigma^2_{\text{network}} + (\widehat{\sigma}^x)^2$, where $\sigma_{\text{network}} = 30\text{m}$ is the standard deviation of the horizontal error in network data (see Section 3.2 for more details).

4.2.2 BT measurement model

We assume that the BT measurement $\widehat{b}$ only depends on whether the transport mode is public or private, then we have:

$$\Pr(\widehat{b}|\mathbf{x}) = \Pr(\widehat{b}|m) = \begin{cases} \Pr(\widehat{b}|m \in \text{PT}) & \text{if } m \text{ is PT} \\ \Pr(\widehat{b}|m \notin \text{PT}) & \text{if } m \text{ is non-PT} \end{cases}$$

where $\text{PT} = \{\text{bus}, \text{metro}\}$ denotes the set of public transport modes. The PT and non-PT models are based on empirical distributions. They are calibrated from the annotated BT data of dataset C, and reported in Table 4.2. The number of measurements used for calibration is 869 for PT and 1826 for non-PT respectively. We observe that the chance of observing a BT device is higher in public transport.

Table 4.2: Probability density mass of $\widehat{b}$

$\Pr(\widehat{b}\|m)$	$\widehat{b}=0$	$\widehat{b}=1$
$m \in PT$	0.19	0.81
$m \notin PT$	0.60	0.40

4.2.3 ACCEL measurement model

Acceleration merely provides information about the transport mode, so we assume it to be independent of the location. As for the BT data, we derive a model based on an empirical distribution. Then we have $\Pr(\widehat{a}|x) = f_a(\widehat{a}|m)$, where $f_a(\widehat{a}|m)$ denotes the probability density function of the ACCEL measurement for mode m. Furthermore, we assume that motor-based transport modes (including car, bus and metro) have a similar pattern of acceleration. Then we calibrate a probability density function for walk, bike and motor-based transport modes respectively, and denote them as $f_a(\widehat{a}|\text{walk})$, $f_a(\widehat{a}|\text{bike})$ and $f_a(\widehat{a}|\text{motor})$.

For each density function, a finite mixture of normal is used to model the distribution of the acceleration measurement:

$$f_a(\widehat{a}) = \sum_{j=1}^{J} w_j \phi(\mu_j, \sigma_j^2). \tag{4.2}$$

The following parameters need to be estimated: J, the number of normal components; w_j, the proportion of component j ($w_j \geq 0$, $\sum_{j=1}^{J} w_j = 1$); μ_j and σ_j^2, the mean and the variance of the normal distribution $\phi(\mu_j, \sigma_j^2)$. These parameters are estimated from the annotated ACCEL data of dataset C. The estimation technique is described by Park, Zhang & Lord (2010) where the same method is applied to model the heterogeneous speed data. A *R* package *mixAK* using Markov chain Monte Carlo methodology is employed for the estimation (Komárek 2009). The optimal number of components J is selected according to deviance information criterion. The histograms of the ACCEL measurements and the predictive densities are drawn in Figure 4.2. Table 4.3 reports the parameter estimates. The gravity 1G corresponds to 280 in the ACCEL measurement, so deviation from 280 means acceleration caused by the smartphone's movement. Acceleration less than gravity is usually caused by vertical movements. We can observe distinct patterns from the distributions. *Walk* is the least stable movement status since it has a higher chance to observe a high acceleration value. *Bike* has a peak near 1G, which means that the movement is quite stable with little acceleration. *Motor* has a peak centered at less than 1G, which depicts vertical movements caused by the road condition (e.g., bumps and uphills) and the usage of the phone by the user.

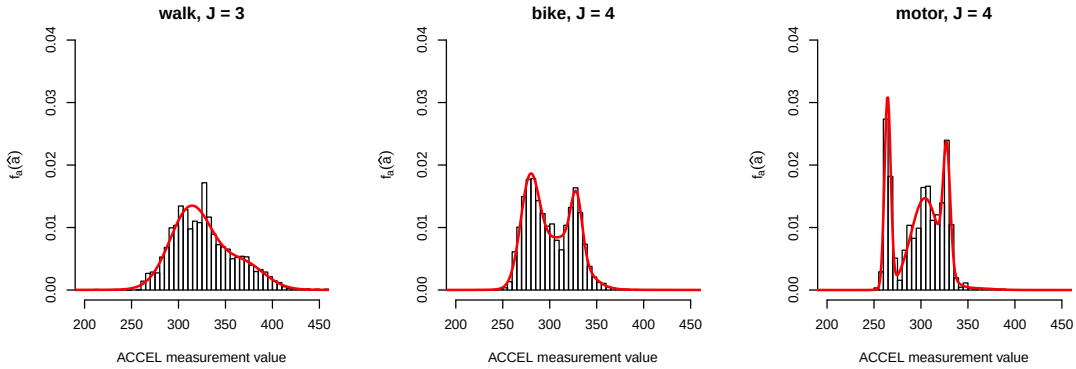


Figure 4.2: Acceleration distributions for walk, bike, and motor

Table 4.3: Parameter estimates for acceleration distributions

	walk	bike	motor
	4501[a]	11924[a]	11801[a]
J	3	4	4
w_1	$7.773e-01$ $(2.782e-02)$[b]	$2.106e-01$ $(5.727e-02)$	$1.922e-01$ $(4.496e-03)$
μ_1	$3.152e+02$ $(1.202e+00)$	$2.929e+02$ $(3.038e+00)$	$2.647e+02$ $(9.726e-02)$
σ_1	$2.206e+01$ $(7.589e-01)$	$9.907e+00$ $(1.327e+00)$	$3.496e+00$ $(7.860e-02)$
w_2	$2.151e-01$ $(2.814e-02)$	$3.217e-01$ $(5.676e-02)$	$7.537e-01$ $(8.748e-03)$
μ_2	$3.712e+02$ $(2.938e+00)$	$2.766e+02$ $(1.063e+00)$	$3.102e+02$ $(2.324e-01)$
σ_2	$1.879e+01$ $(1.711e+00)$	$8.039e+00$ $(4.633e-01)$	$1.333e+01$ $(2.061e-01)$
w_3	$7.668e-03$ $(3.924e-03)$	$2.047e-01$ $(2.102e-02)$	$4.450e-02$ $(4.700e-03)$
μ_3	$3.983e+02$ $(4.307e+01)$	$3.284e+02$ $(3.289e-01)$	$2.864e+02$ $(2.463e-01)$
σ_3	$1.059e+02$ $(2.843e+01)$	$6.411e+00$ $(4.536e-01)$	$2.633e+00$ $(1.872e-01)$
w_4	-	$2.631e-01$ $(4.501e-02)$	$9.626e-03$ $(5.186e-03)$
μ_4	-	$3.163e+02$ $(2.055e+00)$	$3.545e+02$ $(1.213e+01)$
σ_4	-	$2.075e+01$ $(9.360e-01)$	$2.158e+01$ $(5.339e+00)$

[a] The number of measurements used for the calibration.

[b] The figure in parentheses reports the standard deviation of the estimate.

4.3 Smartphone measurement model

In this section, an integrated smartphone measurement model is proposed to combine the sensor measurement models in a unified framework. This smartphone measurement model $\Pr(\widehat{y}_{1:T}|t_{1:T}, p)$ is intended to calculate the likelihood of observing all the smartphone measurements $\widehat{y}_{1:T}$ on a multimodal path p at time $t_{1:T}$ respectively. As the same as in Chapter 3, we assume that the time is recorded without error. Therefore, the model will return a non zero probability only when the sequence $t_{1:T}$ exactly matches the sequence of time stamp $\widehat{t}_{1:T}$ in the data.

4.3.1 Derivation of the smartphone measurement model

The derivation of the smartphone measurement model builds on the procedure described in Section 3.2. In this chapter, we focus on the main differences introduced by the multimodal context and the integration of various sensors.

The measurement equation is decomposed as:

$$\Pr(\widehat{y}_{1:T}|t_{1:T}, p) \quad = \Pr(\widehat{y}_1|t_1, p) \prod_{k=2}^{T} \Pr(\widehat{y}_k|\widehat{y}_{1:k-1}, t_{1:k}, p), \tag{4.3}$$

where $\Pr(\widehat{y}_k|\widehat{y}_{1:k-1}, t_{1:k}, p)$ is the conditional probability for observing $\widehat{y}_k$, and is calculated iteratively. The complex dependency in the sequentially observed measurements is modeled by this conditional probability. In order to simplify its derivation, we assume that the observation of measurement $\widehat{y}_k$ on path p at time t_k only depends on the previous observation. Then the conditional probability for observing $\widehat{y}_k$ simplifies to

$$\Pr(\widehat{y}_k|\widehat{y}_{1:k-1}, t_{1:k}, p) \approx \Pr(\widehat{y}_k|\widehat{y}_{k-1}, t_{k-1}, t_k, p), \tag{4.4}$$

For the first measurement, which is always a GPS measurement by construction, we derive

$$\Pr(\widehat{y}_1|t_1, p) \quad = \int_{x_1 \in p} \Pr(\widehat{y}_1|x_1) \Pr(x_1|t_1, p) dx_1 \tag{4.5}$$

$$= \int_{x_1 \in p} \Pr(\widehat{x}_1|x_1) \Pr(x_1|t_1, p) dx_1, \tag{4.6}$$

where the probability $\Pr(x_1|t_1, p)$ captures a prior knowledge of the initial position of the device. If nothing is known, it can for instance be defined as $\frac{1}{L_p}$ where L_p is the

length of path p. For each subsequent observation $k \geq 2$, we have

$$\Pr(\widehat{y}_k|\widehat{y}_{k-1}, t_{k-1}, t_k, p) = \int_{\mathbf{x}_k \in p} \Pr(\widehat{y}_k|\mathbf{x}_k) \Pr(\mathbf{x}_k|\widehat{y}_{k-1}, t_{k-1}, t_k, p) d\mathbf{x}_k, \tag{4.7}$$

where $\Pr(\mathbf{x}_k|\widehat{y}_{k-1}, t_{k-1}, t_k, p)$ represents the prior probability that the device is at (multimodal) position $\mathbf{x}_k$ at time t_k given the last observed measurement $\widehat{y}_{k-1}$ at t_{k-1}, and can be derived by:

$$\Pr(\mathbf{x}_k|\widehat{y}_{k-1}, t_{k-1}, t_k, p) = \int_{\mathbf{x}_{k-1} \in p} \Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p) \Pr(\mathbf{x}_{k-1}|\widehat{y}_{k-1}, p) d\mathbf{x}_{k-1}, \tag{4.8}$$

where $\Pr(\mathbf{x}_{k-1}|\widehat{y}_{k-1}, p)$ is the posterior distribution of $\mathbf{x}_{k-1}$ from last iteration,

$$\Pr(\mathbf{x}_{k-1}|\widehat{y}_{k-1}, p) = \frac{\Pr(\widehat{y}_{k-1}|\mathbf{x}_{k-1})}{\int_{\mathbf{x}_{k-1} \in p} \Pr(\widehat{y}_{k-1}|\mathbf{x}_{k-1}) d_{\mathbf{x}_{k-1}}}. \tag{4.9}$$

Putting everything together, we have

$$\begin{aligned} &\Pr(\widehat{y}_k|\widehat{y}_{k-1}, t_{k-1}, t_k, p) \\ &= \frac{\int_{\mathbf{x}_k \in p} \int_{\mathbf{x}_{k-1} \in p} \Pr(\widehat{y}_{k-1}|\mathbf{x}_{k-1}) \Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p) \Pr(\widehat{y}_k|\mathbf{x}_k) d_{\mathbf{x}_{k-1}} d_{\mathbf{x}_k}}{\int_{\mathbf{x}_{k-1} \in p} \Pr(\widehat{y}_{k-1}|\mathbf{x}_{k-1}) d_{\mathbf{x}_{k-1}}}. \end{aligned} \tag{4.10}$$

There are two kinds of essential components in this equation. One is the sensor measurement models, $\Pr(\widehat{y}_1|\mathbf{x}_1)$, $\Pr(\widehat{y}_{k-1}|\mathbf{x}_{k-1})$ and $\Pr(\widehat{y}_k|\mathbf{x}_k)$, which are already described; the other is the travel model $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p)$, which we define next.

4.3.2 Travel model

The travel model with the form

$$\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p) \tag{4.11}$$

essentially predicts the position $\mathbf{x}_k = (x_k, m_k)$ at time t_k, given that the state at time t_{k-1} is $\mathbf{x}_{k-1} = (x_{k-1}, m_{k-1})$, and the smartphone user is traveling along path p. There are several ways of implementing the travel model, for instance, via a traffic simulator or real-time traffic information. In this chapter, we extend the empirical model proposed in Section 3.2.3 to multimodal context. It is based on the speed distribution for each transport mode.

Table 4.4: Parameter estimates for speed distributions

mode	measurements[a]	w_m	λ_m	μ_m	τ_m
walk	9350	0.46 (0.01)[b]	0.20 (0.00)	4.41 (0.03)	1.51 (0.03)
bike	11899	0.39 (0.01)	0.09 (0.00)	2.88 (0.00)	0.30 (0.00)
metro	1142	0.52 (0.02)	0.17 (0.01)	3.51 (0.03)	0.43 (0.02)
bus	1669	0.48 (0.07)	0.13 (0.03)	3.16 (0.05)	0.46 (0.02)
car	2069	0.20 (0.03)	0.12 (0.03)	3.76 (0.03)	0.62 (0.02)

[a] The number of measurements used for the calibration.
[b] The figure in parentheses reports the standard deviation of the estimate.

Speed distributions

Researchers have been using speed profiles to infer transport modes (e.g., Liao et al. 2007, Zheng, Li, Chen, Xie & Ma 2008, Reddy et al. 2009, Bohte & Maat 2009). Studies have also been performed on estimating the speed profiles of transport modes (e.g., Knoblauch, Pietrucha & Nitzburg 1996, Thompson, Rebolledo, Thompson, Kaufman & Rivara 1997).

A speed distribution for car has been estimated in Section 3.2.3. The distribution is assumed to be a mixture of a negative exponential and a log-normal. The first is designed to capture the period when the traveler is stopped, or traveling at low speed before or after that stop. The second is designed to capture the traveler moving at regular speed. In this chapter, this method is adapted to estimate a speed distribution $f_v(v|m)$ for each transport mode. Speed measurements from dataset C are used for the estimation. And the probability density function for mode m is written as:

$$f_v(v|m) = w_m \lambda_m e^{-\lambda_m v} + (1 - w_m) \frac{1}{v\sqrt{2\pi\tau_m^2}} e^{-\frac{(\ln v - \mu_m)^2}{2\tau_m^2}}. \tag{4.12}$$

Our data analysis shows that a mixture of negative exponential and normal fits better for walk. The distribution for walk is therefore

$$f_v(v|\text{walk}) = w_{\text{walk}} \lambda_{\text{walk}} e^{-\lambda_{\text{walk}} v} + (1 - w_{\text{walk}}) \frac{1}{\sqrt{2\pi\tau_{\text{walk}}^2}} e^{-\frac{(v - \mu_{\text{walk}})^2}{2\tau_{\text{walk}}^2}}. \tag{4.13}$$

The parameters to be estimated are: w_m, the weight for the mixture; λ_m, the scale parameter of the negative exponential distribution; μ_m, the location parameter of the normal and log-normal distributions respectively; τ_m the scale parameter of the normal and log-normal distributions respectively. Figure 4.3 shows the normalized histograms of the recorded speed data and the estimated speed distributions for all modes. Table 4.4 reports the parameters estimated by maximum likelihood.

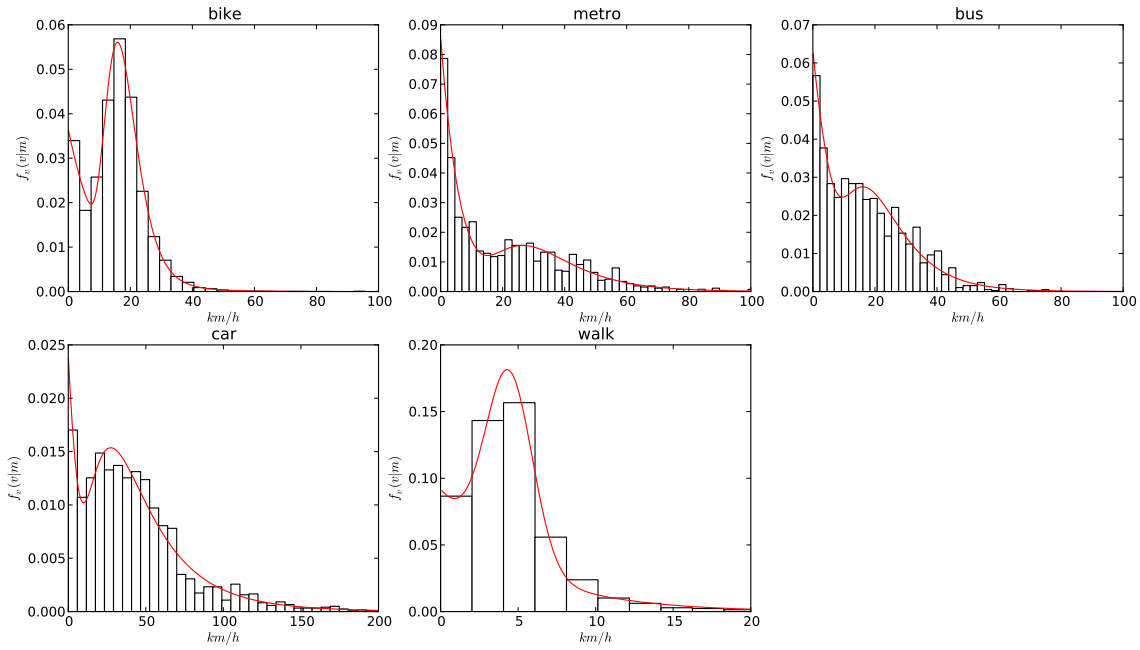


Figure 4.3: Speed distributions of 6 transport modes.

Derivation of the travel model

A multimodal path differs from a unimodal path in that there are possible mode changes along a path. Therefore, there are two situations that need to be considered in deriving the travel model: the presence or absence of a virtual mode transfer arc between x_{k-1} and x_k along p.

If there is no mode change between x_{k-1} and x_k. The smartphone carrier travels from x_{k-1} to x_k along path p with the same mode $m_k = m_{k-1}$. Then the probability density function of the travel model (4.11) can also be written as

$$f_x(x_k|x_{k-1}, t_{k-1}, t_k, m_k, p), \tag{4.14}$$

which predicts the next location x_k of the unimodal (m_k) travel along path p since the previous location x_{k-1}. We assume that the travel speed follows the speed distribution of the transport mode m_k being used in this uni-modal travel segment, then we have the following model:

$$f_x(x_k|x_{k-1}, t_{k-1}, t_k, m_k, p) = f_v(\frac{d_p(x_{k-1}, x_k)}{t_k - t_{k-1}}|m_k), \tag{4.15}$$

where $d_p(x_{k-1}, x_k)$ calculates the distance from x_{k-1} to x_k on path p; so that $\frac{d_p(x_{k-1},x_k)}{t_k - t_{k-1}}$ calculates the travel speed; and the probability density function f_v is given by Equation (4.12) or (4.13).

If there are mode changes between $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$. Considering the fact that the GPS data are recorded every 10 seconds, we assume that it is not possible to have more than one mode change in such a short time. The mode change between $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$ is represented by a virtual arc on p, associated with coordinates x_c. We denote upstream node of the virtual arc by $\mathbf{x}_c^- = (x_c, m_{k-1})$ and the downstream node by $\mathbf{x}_c^+ = (x_c, m_k)$. The time at which the mode change happens is unknown and denoted by $t_c \in [t_{k-1}, t_k]$. Then, we have the following model:

$$\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_{k-1}, t_k, p) = \int_{t_c=t_{k-1}}^{t_k} \Pr(t_c|\mathbf{x}_{k-1}, t_{k-1}, p) \Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_c, t_k, p) dt_c, \quad (4.16)$$

The two probabilities in the right hand side (RHS) of (4.16) describe the two unimodal travel segments before and after the mode change. The first predicts the mode change time t_c; the second predicts the position $\mathbf{x}_k$ at time t_k given the mode change time t_c. Following the derivation of (4.15), we also assume that the travel speed of each segment follows the speed distribution of the corresponding transport mode. Then, they can be re-written as:

- $\Pr(t_c|\mathbf{x}_{k-1}, t_{k-1}, p) = f_v(\frac{d_p(x_{k-1},x_c)}{t_c - t_{k-1}}|m_{k-1})$,
- $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, t_c, t_k, p) = f_v(\frac{d_p(x_c,x_k)}{t_k - t_c}|m_k)$.

4.3.3 Computing integrals

The above formulations involve a lot of integrals along path p with the form of $\int_{\mathbf{x}\in p} f(\mathbf{x})d\mathbf{x}$. In order to save computation time, Section 3.2.2 defines a Domain of Data Relevance (DDR) of each GPS point as a physical area nearby. Then, the domain of the integral is truncated to the part of the path that is inside the DDR.

Obviously, this simplification method only works for GPS data, because other data do not contain location information. In this chapter, the domain of integral for BT and ACCEL measurements is illustrated in Figure 4.4. For each BT or ACCEL measurement $\widehat{y}_k$, its previous and next GPS measurements are denoted as $\widehat{y}_{k^{g-}}$ and $\widehat{y}_{k^{g+}}$ ($t_{k^{g-}} \le t_k \le t_{k^{g+}}$). Since measurements are observed sequentially, $\widehat{y}_k$'s state $\mathbf{x}_k \in p$ has to be downstream of $\mathbf{x}_{k^{g-}}$ and upstream of $\mathbf{x}_{k^{g+}}$. Then, the domain of integral for $\mathbf{x}_k$ only includes the

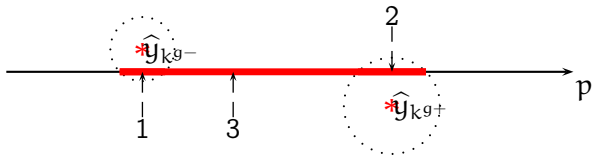


Figure 4.4: Integral domain for Bluetooth or Acceleration measurement $\widehat{y}_k$

domains of integral for $\mathbf{x}_{k^{g-}}$ and $\mathbf{x}_{k^{g+}}$ (part 1&2 in Figure 4.4), plus the part of p that connects $\mathbf{x}_{k^{g-}}$ and $\mathbf{x}_{k^{g+}}$ (part 3). This definition highly relies on the two adjacent GPS measurements. If the GPS measurements happen to have a large time gap, they do not provide reliable location information for BT and ACCEL data observed between them. Therefore, as mentioned earlier, the BT and ACCEL measurements observed between them are discarded. For the implementation of the DDR of GPS measurements and the integral, we refer to Section 3.2.2 for more details.

4.4 Candidate path generation

We propose a multimodal candidate path generation algorithm as an extension of the unimodal algorithm described in Section 3.3. This algorithm has special features compared to the conventional map-matching and transport mode inference algorithms:

- The algorithm builds the physical path and the transport modes simultaneously.
- Smartphone data recorded from a trip are not required to be preprocessed into several unimodal segments.
- Transport networks also contribute to the inference of the transport mode, especially in differentiating PT and non-PT modes.

The algorithm produces a set of multimodal paths, denoted as $\mathcal{P}$, along with a likelihood for each one. Based on $\mathcal{P}$, the probability for each path being the one can be calculated:

$$q(p) = \frac{\Pr(\widehat{y}_{1:T}|t_{1:T}, p)\Pr(p)}{\sum_{p' \in \mathcal{P}} \Pr(\widehat{y}_{1:T}|t_{1:T}, p')\Pr(p')}, \tag{4.17}$$

Again, because we assume that the time tags are measured without error, the values of $t_{1:T}$ are taken directly from the data $\widehat{t}_{1:T}$. The prior probability $\Pr(p)$, representing a route choice model, is specified as uniform.

Usually, a map-matching algorithm takes two sources of information: location data, and transport networks data. In this algorithm, although BT and ACCEL measurements do

not contain significant location information for generating path candidates, they are implicitly used in the process that eliminates paths according to the path probability (4.17). The output is a set of candidate multimodal paths. For each path, the likelihood (4.3) of observing the smartphone measurements on it, and the probability (4.17) that it is the true path are also calculated.

The algorithm iterates over the sequence of GPS measurements $\widehat{g}_{1:I}$. At each iteration i, it generates a set of candidate paths $\mathcal{P}_i$ that are matched to the sequence of all measurements (including BT and ACCEL) up to $\widehat{g}_i$. In the next iteration $i+1$, each path is extended from its end node, and downstream segments are appended in order to map the new measurements up to the next GPS $\widehat{g}_{i+1}$. In fact, this iterative method connects the DDR of the GPS points to build candidate paths. Practically, the number of candidate paths grows exponentially because each DDR is relatively large with about $100m$ radius. Therefore, heuristics are proposed to reduce the computational burden. The result from the last iteration I is the final output of the algorithm. The pseudo-code is briefly described as Algorithm 2. Detailed explanations of some steps are given as follows:

9. At iteration i, the path extension process is carried out only when the GPS point $\widehat{g}_i$ is far enough away from $\widehat{g}_{i'}$, where i' corresponds to the last iteration when a path extension happened, and $100m$ is chosen as the distance threshold. Otherwise, the traveler is considered to be immobile, and a path extension is not necessary.

15. The path extension from end node n takes place in each unimodal network G_m, if n appears in G_m or connects to G_m with a virtual arc (mode change). A new path candidate p_{new} is created by joining the current path candidate $p \in \mathcal{P}_{i'}$ with the newly discovered downstream segment (see line 22). The transport mode m of the downstream segment is the mode of G_m. It can be different from the mode of the last arc on p. In other words, a mode change is allowed to happen at the connecting node n, which is the end node of p.

16. Each transport mode has a speed limit. For example, walk is not expected to have a speed over $18km/h$. If the observed speed $\widehat{v}_i$ of the GPS exceeds the maximum speed of a transport mode m, the corresponding unimodal transport network G_m is neglected. This is mainly designed to neglect walk and bike networks when the GPS is in fact observed from higher speed motor modes, hence to reduce the amount of irrelevant paths as candidates. The maximum speed for walk and bike is set to be $18km/h$ and $40km/h$ respectively, while no constraint is imposed to motor modes. These values correspond to 99% percentile in the speed data of dataset C.

Algorithm 2: Candidate path generation algorithm

Input: The smartphone measurements sequence $\widehat{y}_{1:T}$ with GPS subsequence $\widehat{g}_{1:I}$
Input: The underlying multimodal transportation network G with multiple unimodal networks G_m, $m \in \{\text{walk}, \text{bike}, \text{car}, \text{bus}, \text{metro}\}$.
Result: A set of candidate paths $\mathcal{P}_I$.

```
// Deal with the first GPS point.
P_1 ← empty set of paths;
DDR_1 ← the DDR of the first GPS measurement;
for each arc a ∈ A do
    if a intersects DDR_1 then
        include a as a partial path in P_1;
i' ← 1: the temporary index for processed GPS;
for i ← 2 to I do
    // Iterative path extension process.
    if ||x̂_i − x̂_i'|| > 100m then
        i' ← i;
        foreach p ∈ P_i' do
            if p intersects DDR_i then
                include p in P_i;
            n ← the end node of p;
            foreach unimodal network G_m do
                if v̂_i ≤ the maximum speed of mode m then
                    spt ← a bounded shortest path tree rooted at n in G_m ;
                    foreach link a ∈ spt do
                        if a intersects DDR_k then
                            sp ← shortest path connecting p and a;
                            p_new ← join p, sp and a;
                            include p_new in P_i;
        limit the size of P_i;
```

18. Shortest path trees are used to link the DDRs of adjacent GPS points. For the sake of computational efficiency, the shortest path trees are bounded. The leaf nodes of a bounded shortest path tree are the first nodes detected by the Dijkstra algorithm that violates the bound. The bound is the same as in Section 3.3. It is based on an assumption about the maximum possible speed of the traveler within the time interval $[t_{i'}, t_i]$. In our experiments, the bound is defined by $1.5(t_i - t_{i'})\widehat{v}_{max}$, where $\widehat{v}_{max}$ is the maximum speed value among the observed GPS speeds $\widehat{v}_{i'}$ and $\widehat{v}_i$, and the speed value calculated from the coordinates: $\|\widehat{x}_i - \widehat{x}_{i'}\|_2/(t_i - t_{i'})$. The factor 1.5 is a safety margin to minimize the risk of missing a relevant observation.

24. The path elimination procedure is designed to speed up the algorithm by eliminating less relevant branches. It eliminates unreasonable paths deterministically according to various criteria. The deterministic elimination procedure includes:

    1. We assume that walk is necessary for mode changes. Therefore, if a path has a mode change without walk involved, it is eliminated.
    2. A path with loops is considered to be unreasonable, hence is excluded, unless the loops involve walk.
    3. A path might be too long to be consistent with the observed travel time approximated by $t_I - t_1$. The mean travel speed $\bar{v}_m$ for each mode m is taken from the speed data of dataset C. Then the mean travel time for a path can be calculated by summing up the mean travel time $\frac{L_a}{\bar{v}_{m_a}}$ for each arc a, where m_a is the mode of arc a and L_a is the length of the arc. We assume the lower bound of a path's travel time as half of the mean travel time. If the observed travel time is lower than the lower bound, the path is considered too long to be realistic, and removed.

    Clearly, more behavioral rules could be considered here, possibly involving a calibrated behavior model, or a Markov sequential conditional probability for mode changes (e.g., Zheng, Li, Chen, Xie & Ma 2008).

    In order to control the complexity of the algorithm, the number of paths generated at each iteration should be reasonably small. For example, Marchal et al. (2005) and Schuessler & Axhausen (2009b) suggest to maintain 30 paths at each iteration for unimodal map-matching. We use a random sampling procedure to select paths according to the path probability (4.17). Since the algorithm is multimodal, we decide to maintain more paths, but not more than 60 paths in our experiments. The random selection procedure includes three steps.

    1. Randomly draw some paths from $\mathcal{P}_i$ according to the path probability (4.17). In our experiments, 20 paths are selected.

2. Let $\mathcal{P}_i^1 \subseteq \mathcal{P}_i$ denote the set of paths with the least mode changes. Then, randomly sample some paths from $\mathcal{P}_i^1$ according to the path probability (4.17). Before the random sampling, the probabilities are normalized such that they sum up to one for the paths in $\mathcal{P}_i^1$. In this step, paths with the least mode changes are favored because they are more behaviorally reasonable. In our experiments, 10 paths are selected in this step.

3. The likelihood that the GPS measurement is observed on a is defined:

$$\Pr(\hat{x}|a) = \int_{x \in a} \Pr(\hat{x}|x) \Pr(x|a) dx = \frac{1}{L_a} \int_{x \in a} \Pr(\hat{x}|x) dx. \tag{4.18}$$

We create the set of arcs that intersect with DDR_i and have the transport mode m, and denote the set as A_{im}. For each mode m, we sample some arcs according to the likelihood (4.18) (as in Step 2, the normalization of the likelihood is required before the random sampling). In our experiments, 5 arcs for each mode are selected. For each sampled arc a, we denote $\mathcal{P}_{ia} \subseteq \mathcal{P}_i$ as the set of paths that go via a and have the least mode changes. We then apply a similar random sampling procedure as in Step 2 on $\mathcal{P}_{ia}$, but only to draw one path. In this step, the sampled paths go through different arcs with different modes that intersect with DDR_i. Therefore, this step ensures sufficient variability in the generated paths.

If a trip is unimodal and the mode is known, the algorithm can be used to only identify the physical path. It is simply accomplished by supplying the unimodal transport network of the known mode. Since this technique is essentially unimodal map-matching, it is denoted as Algorithm-U, while the original multimodal algorithm is denoted as Algorithm-M. In the next section, the results generated by Algorithm-U with the correct transport mode will be used as the benchmark when we analyze the mode inference performance of Algorithm-M.

4.5 Numerical experiments

The proposed method is implemented as a software package in C++. It reads smartphone data and OSM network data as inputs, and produces probabilistic map-matching results. In this section, numerical experiments are performed with smartphone data collected in different circumstances. Some examples are first illustrated with map visualization to gain an intuitive impression of the results. Then, numerical analyses focus on the performance in inferring the modes. The contributions of BT and ACCEL data are also analyzed.

4.5.1 Result illustration

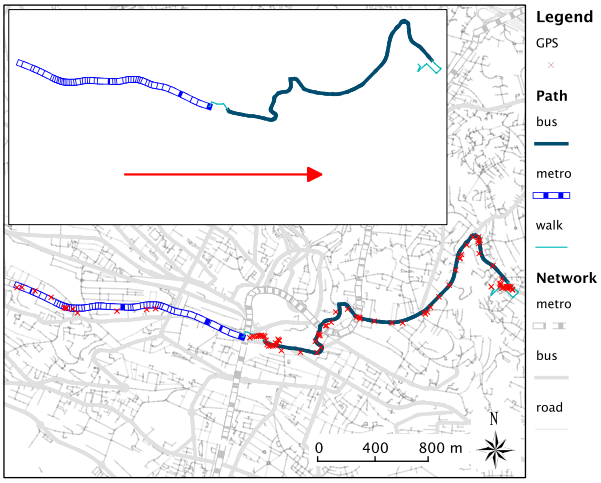
Some common trip patterns are chosen for the illustration, including bike, car, and public transport with changes. Using information from dataset C, the main transport modes for each trip are known while the exact path is unknown.

The first example in Figure 4.5(a) shows a complex multimodal trip with metro $\rightarrow$ walk $\rightarrow$ bus $\rightarrow$ walk. The total travel time is 20 minutes, and there are 91 GPS measurements generated, with 8 BT, and 395 ACCEL. The background network is shown in gray lines, while a generated path is drawn with color. According to the smartphone user who provided the data, this path resembles the trip that he made. The graph without background network shows the same path. The red arrow shows the direction of the trip. There are 43 paths generated by the multimodal map-matching algorithm (Algorithm-M). The measurement log likelihood $\ln(\Pr(\widehat{y}_{1:T}|t_{1:T}, p)$ for each path is plotted in Figure 4.6, and the x-axis shows the id of each path. We notice that some paths have much higher log likelihood than the others. The path 22 drawn in Figure 4.5(a) gains the highest log likelihood (-347.9).

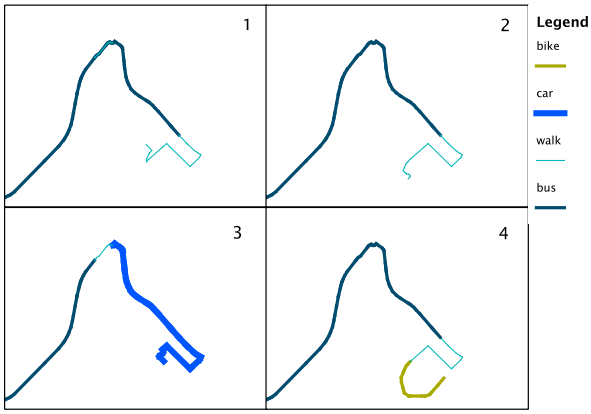
The differences of the generated paths show the uncertainty of the result. On the one hand, the uncertainty is due to the imprecision of the smartphone data and the network. On the other hand, the uncertainty mainly belongs to the end of the trip, since we notice that the generated paths mainly differ at the end of them. This can be explained by the mechanism of the smartphone measurement model. The model utilizes the dependency between adjacent measurements (see Equation 4.4). Each measurement in fact provides information to identify its upstream trajectory. The end of a trip always gains less information since it has less (or none) downstream measurements. We focus on the differences of the paths by showing the end of them in Figure 4.5(b). Graph 1 shows path 22's end; Graph 2 shows path 23's end, which has a different destination (log likelihood -348.7); Graph 3 shows path 14, of which a part is identified as car (log likelihood -384.0); Graph 4 shows path 40 with its end identified as bike (log likelihood -381.9).

The second example in Figure 4.7 shows a car trip. All the generated 30 paths are drawn in the figure, and they greatly overlap with each other. Except for the uncertainty of the trip end, there is also uncertainty (marked by a circle) due to the data noise and the density of the network.

The third example in Figure 4.8 shows a trip with bike as the main mode. There are 33 paths generated, the left graph draws a path with the highest log likelihood -117.7. The same path without background network is drawn in the top right graph. The end of the path is identified as walk because the smartphone user is entering a parking place. The bottom right graph shows another representative path, which gains a little lower log



(a) Data and a generated path.



(b) Trip end uncertainty.

Figure 4.5: A multimodal trip.

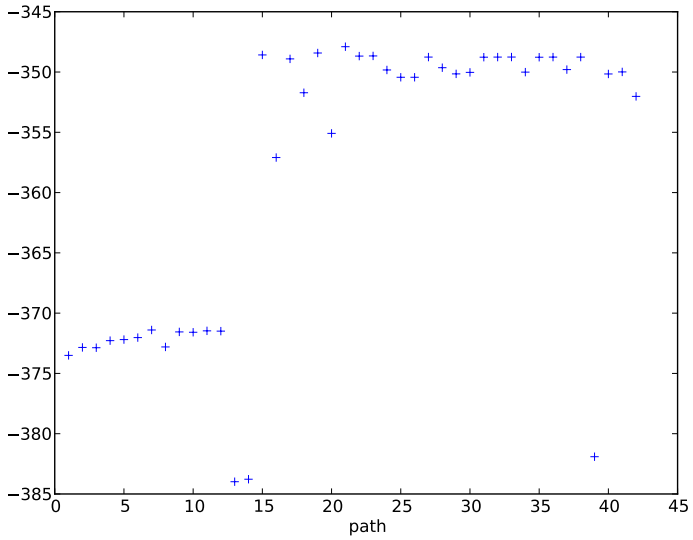


Figure 4.6: Measurement log likelihood for paths

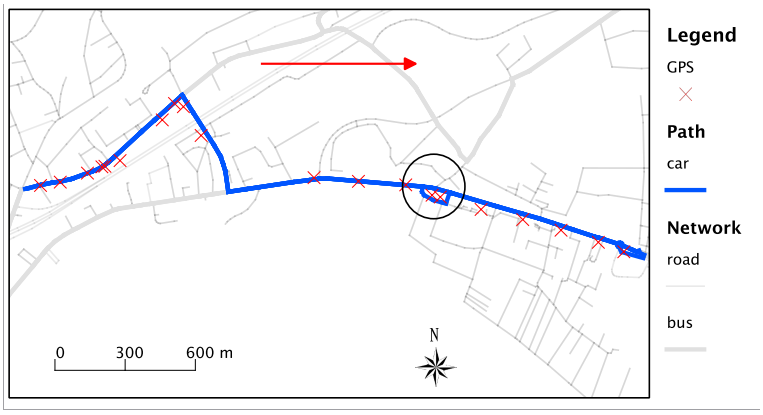


Figure 4.7: A car trip

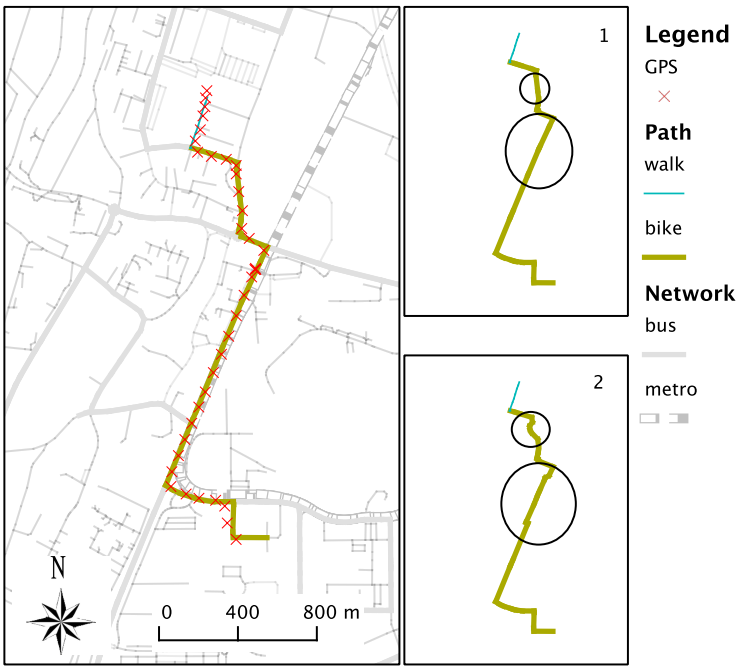


Figure 4.8: A bike trip

likelihood −118.0. The differences between two paths are highlighted in circles.

In these examples, the paths with the highest likelihoods are chosen for illustration purpose only. In practice, we suggest to carry the uncertainty in the application, together with the associated probability. For example, in route choice modeling, the probabilistic path observations from the map-matching results can be used with network-free method (Bierlaire & Frejinger 2008).

4.5.2 Performance analysis

In order to gain more systematic understanding of the performance of the algorithm, analysis with more data is provided. The analysis focuses on the most important aspects of the method proposed in this chapter: the identification of the modes, and the usage of various kinds of data. For the sake of convenience, we extract from dataset C data sequences that are known to have one single mode. 36 data sequences are used for the analysis. The transport mode, the travel time, the number of GPS, BT and ACCEL measurements for each data sequence is given in the left part of Table 4.5.

Since each data sequence has a known transport mode, Algorithm-U with the known mode can be applied to generate unimodal paths that have the correct transport mode. Algorithm-M are then applied, and the map-matching results (path sets) are denoted as

follows:

P^0: Algorithm-U is applied to only GPS data, with known mode.

P^1: Algorithm-M is applied to only GPS data.

P^2: Algorithm-M is applied to GPS and BT data, if BT data are available.

P^3: Algorithm-M is applied to all data, if ACCEL data are available.

Since we don't have access to the ground truth of the traveled path, we use uni-modal map-matching result P^0 with correct transport mode as the benchmark. We refer to Chapter 3 for more discussion on the route identification accuracy of the probabilistic unimodal map-matching algorithm.

The performance of Algorithm-M with different data is evaluated by comparing P^1, P^2, P^3 against P^0. We expect that P^1, P^2 and P^3 are similar to P^0, if Algorithm-M correctly identifies the paths and the modes. In order to compare the path sets, we first define quantitative similarity indicators.

Similarity indicators

First, the overlapping indicator $O(p, P)$ is defined to measure how much a path p overlaps with all the paths in a path set P:

$$O(p, P) = \sum_{a \in p} \frac{L_a}{L_p} \sum_{p' \in P} q(p') \delta_{ap'}, \tag{4.19}$$

where $\delta_{ap'}$ is a dummy variable, valued 1 if path p' contains arc a, and 0 otherwise; L_a and L_p are the lengths of arc a and path p respectively; $q(p')$ is the path probability (4.17). This definition is similar to (3.28), but we take the path probability into account such that the outliers in the results are not overrepresented in the similarity indicator. For example, the result may contain a path with totally different modes than other paths, and this path may gain very low probability. Then, this path contributes less to the similarity indicator due to the probability term. This overlapping indicator is valued between 0 and 1, and can be roughly understood as the average proportion of the path p overlapping with all paths in P. When p is the same as any path in P, then the overlap is total and $O(p, P) = 1$; when p does not overlap with any path in P at all, $O(p, P) = 0$.

Then $S(P', P)$ is defined to compare another path set P' against P,

$$S(P', P) = \sum_{p \in P'} q(p) O(p, P). \tag{4.20}$$

S is also valued between 0 and 1. When all paths in both P and P' are the same, $S(P', P) = 1$; when all paths are distinct without any overlap, $S(P', P) = 0$. If $P' = P$, $S(P, P)$ in fact calculates the similarity of the paths in the same set. When P is a map-matching result, $S(P, P)$ indicates the level of the uncertainty in the result. The higher the similarity, the lower the uncertainty. For example, $S(P, P)$ for the public transport trip shown in Figure 4.5 is 0.967, for the car trip (Figure 4.7) is 0.956, while for the bike trip (Figure 4.8) is lower 0.854 because more uncertainty is observed.

For the comparison of the different results, we select P^0 as benchmark, and analyze the values $S^0 = S(P^0, P^0)$, $S^1 = S(P^1, P^0)$, $S^2 = S(P^2, P^0)$, $S^3 = S(P^3, P^0)$. The uncertainty of the unimodal matching result S^0 tells the degree of the data noise and the density of the network. S^1, S^2, S^3 are expected to have a high value, but not higher than S^0.

Analysis

Table 4.5: Numerical comparisons of results

	mode	time (s)	GPS	BT	ACCEL	S^0	S^1	S^2	S^3
1	bus	479	12	0	19	0.99	0.07	-	0.15
2	bus	399	40	0	25	0.98	0.93	-	0.93
3	bus	234	24	1	11	0.96	0.64	0.65	0.93
4	bus	499	47	0	23	0.98	0.81	-	0.98
5	bus	255	24	0	23	0.96	0.96	-	0.97
6	bus	412	42	0	41	0.97	0.94	-	0.86
7	bus	417	39	2	34	0.98	0.98	0.98	0.98
8	bus	479	35	0	7	0.98	0.27	-	0.50
9	car	229	20	0	23	0.97	0.95	-	0.96
10	car	180	16	0	0	0.95	0.87	-	-
11	car	241	23	0	23	0.92	0.91	-	0.90
12	car	229	24	0	0	0.93	0.93	-	-
13	bike	290	29	0	0	0.91	0.55	-	-
14	bike	289	27	0	0	0.80	0.68	-	-
15	bike	313	32	0	0	0.93	0.80	-	-
16	bike	369	38	1	23	0.83	0.76	0.77	0.76
17	bike	1153	115	0	98	0.96	0.89	-	0.81
18	bike	1021	100	4	73	0.97	0.95	0.95	0.77
19	metro	892	62	1	34	0.99	0.99	0.99	0.97
20	metro	560	34	1	23	0.99	0.77	0.82	0.85
21	metro	259	16	0	0	0.98	0.94	-	-

Continued on next page

	mode	time (s)	GPS	BT	ACCEL	S^0	S^1	S^2	S^3
22	metro	409	33	1	23	0.98	0.99	0.98	0.99
23	metro	594	49	2	40	0.99	0.95	0.96	0.96
24	metro	716	38	2	0	0.96	0.80	0.91	-
25	metro	601	16	0	20	0.97	0.13	-	0.89
26	metro	449	39	0	20	0.99	0.98	-	0.99
27	metro	230	22	0	23	0.95	0.76	-	0.94
28	metro	579	7	0	10	0.98	0.08	-	0.48
29	walk	1269	114	0	0	0.88	0.71	-	-
30	walk	719	62	0	0	0.65	0.58	-	-
31	walk	659	47	0	0	0.76	0.80	-	-
32	walk	998	97	5	0	0.93	0.90	0.88	-
33	walk	240	21	0	0	0.72	0.59	-	-
34	walk	359	27	1	0	0.80	0.75	0.73	-
35	walk	488	38	0	0	0.85	0.84	-	-
36	walk	490	35	2	26	0.83	0.81	0.77	0.61

The similarity indicators for all trips are reported in Table 4.5. A empty cell means that the corresponding data is not available, hence no result is generated. S^2 is empty when BT data are unavailable, S^3 is empty when ACCEL data are unavailable.

We first notice that all S^0 have high value with low uncertainty in the results. The value is 0.921 in average. Since the uncertainty is mainly due to the error of the GPS data and the density of the transport network, S^0 for walk data is lower because the walk network is usually denser.

When Algorithm-M is applied with the multimodal network, we have $S^1, S^2, S^3 < S^0$, because the algorithm is not aware of the true mode and there is a chance of mis-identifying it. However, in the majority of the cases, S^1 is close to S^0, and the average of S^1 is 0.757, which is 82.2% of the average of S^0. Considering the complexity of the multimodal network and the sparsity of the GPS data, Algorithm-M achieves quite high accuracy in the transport mode inference. We observe some exceptional cases (case 1, 8, 25, and 28), where S^1 has very low value. They are mainly due to two reasons. First, in case 1, 25 and 28, the GPS data are too sparse, therefore the measurements do not provide enough information to find out the correct mode. Indeed, in case 1, there are only 12 GPS measurements observed in 479 seconds. Second, in case 8, the data are observed when the bus was running slowly in peak hour in the city center. Therefore, the chance

of identifying the mode as bike increases. In these two cases, some generated paths have very strange mode change behavior, such as, using both bike and car in 10 minutes. We believe that if an appropriate route choice behavior model is incorporated in the candidate path generation algorithm, such paths will be less favored by the algorithm.

By comparing S^1 among different modes, we observe that non-PT cases have high values, because people do not follow the PT lines during the entire trip when they use private mode. Hence the chance of mis-identifying the mode as public transport (bus and metro) is low. In this situation, the transport network helps in identifying the mode. From another perspective, if a route choice behavior model can consider the fact that private mode travel does not follow PT lines, the results could further be improved.

In 12 cases where BT data are available, the average of S^2 (0.888) is greater than the average of S^1 (0.858). In 22 cases where ACCEL data are available, the average of S^3 (0.826) is greater than the average of S^1 (0.751). Therefore, in general, the additional BT and ACCEL data contribute to the accuracy of the results. Additional ACCEL data are particularly helpful when S^1 has very low value (case 8, 25 and 28). In some cases, although there are drops in S^3 with additional data, the values are still acceptable. Still, ACCEL data need to be used more carefully. We notice that ACCEL changes a lot when the vehicle is traveling at low speed, because the vehicle is accelerating or decelerating frequently then. This information can be used to improve the ACCEL measurement model in the future. The candidate path generation procedure involves random sampling of a subset of candidates, thus stochasticity is introduced. We run P^3 for the second time and compare the resulted S^3 indicators to the ones reported in Table 4.5. It is found that the absolute value of the relative difference is only 8.1% in average, which indicates that the results are not sensitive to the stochasticity.

This chapter aims at a mathematically sound probabilistic measurement model, and it involves numerical integrations that are computational intensive. The computation time is roughly linear to the number of GPS points involved in the candidate path generation algorithm. In each path extension iteration for a GPS point, the computation time depends on the complexity of the network and the amount of ACCEL and BT data recorded since the previous GPS point. It varies from less than a second for car trip in suburban area, to a couple of minutes for bus trip in city center, on a MacbookPro using single thread (CPU 2.66 GHz).

The numerical experiments with real smartphone data show that the proposed multimodal map-matching algorithm performs well in identifying the multimodal paths from the smartphone data. The algorithm works when at least GPS data is available. The inclusion of BT and ACCEL data improves the inference accuracy. We have tested the proposed method on relatively sparse GPS data with 10 seconds time interval. In fact,

the proposed method should also work with denser or sparser GPS data. Section 3.4.5 provide some evidence that the same framework performs well in unimodal map-matching for both dense and sparse data (time interval ranges from 1 second to 1 minute).

4.6 Conclusions and discussions

This chapter proposes a probabilistic method to infer the path and the modes of a trip from smartphone data. A smartphone measurement model is derived to calculate the likelihood that a smartphone would have generated a sequence of measurements while traveling on a multimodal path. It is based on a structural travel model that captures the dynamic of the smartphone user's state in the transport network, and sensor measurement models that capture the sensors' operation. This smartphone measurement model synthesizes information available from various sensors, such as GPS, BT and ACCEL.

An algorithm is developed to generate candidate paths from the smartphone data. This algorithm identifies the physical path and the modes of a trip simultaneously. Hence, the transport network information is also utilized to identify the transport modes of a trip. Data recorded from a multimodal trip do not need to be divided into unimodal travel segments. The result of the algorithm is a set of candidate multimodal paths, along with a probability for each being the true one.

The proposed method is flexible in two aspects. First, in the smartphone data aspect, this method works when at least GPS data is available, but richer data, e.g. with BT and ACCEL, results in better accuracy. For example, under congested traffic conditions, a car might have the similar speed pattern as a bus, then the travel model might not be sufficient to distinguish between car and bus, although it could recognize with high confidence that the mode is not walk. Nonetheless, we expect that a car driving may not always follow a bus line. Even if it does, a smartphone observes less BT devices in a car than in a bus. In extreme cases that the car behaves exactly as a bus, the result is still probabilistic with several candidates, possibly containing both car and bus. The probabilistic result avoids deterministic wrong identification, thus further information will be able to be supplemented to improve the inference precision. For example, the bus schedule will help in this case. Second, in the transport network aspect, we can remove or add networks depending on need and availability. For instance, if we know that a travel is absolutely not bike, the bike network can be removed. For another example, if we want to distinguish between local bus and express bus, the lines can be defined in two different networks. Since different services usually do not have large physical overlaps, the type of the service can be recognized from the GPS location data. Moreover, we can construct different travel models for different types of bus services, if such information is

available, e.g., from the bus companies.

The visualized examples show that the results are intuitively reasonable, and the measurement likelihood values are realistic and meaningful. A complex multimodal trip example shows the capability of the algorithm in dealing with mode changes. Numerical analysis shows the performance of the algorithm in identifying the transport modes. Apart from the most useful GPS data, BT and ACCEL also contribute in identifying the transport mode.

Future works involve more investigations on the usage of BT and ACCEL data. They need to be used more carefully, because they do not contain any location information, and highly rely on the adjacent GPS measurements to have prior information about where they are observed. More sophisticated BT measurement model can be considered. In particular, we may want to capture the fact that walk might happen in a crowded place where BT devices are observed. In this case the location and time of the day should play roles in the BT measurement model. The measurement variable can be defined as the number of nearby devices in order to utilize more information from the BT data. Besides the mean, more features of the ACCEL data can also be used. The travel model can adopt more external information, for instances, the timetable of public transport, and the real time traffic information observed from sensors such as loop detectors. If a route choice model, e.g. estimated from historical observations, can be supplied for the prior probability in candidate path generation, the results will be further improved. The computation speed will be improved with more efficient approximation to the integration, and better behavioral roles in the elimination procedure of the candidate path generation algorithm. Finally, the probabilistic map-matching results will be used to estimate multimodal route choice behavior.

5 Route choice models estimated from GPS data

Bierlaire & Frejinger (2008) propose a new discrete choice modeling framework for route choice data that is not associated with the transportation network ("network-free" data). They suggest that the error in the "network-free" data should be treated in a probabilistic manner. Although they also suggest that the methodology may be applied to GPS data, they present only results based on interview data. The Metropolis-Hastings path sampling (MHPS) technique proposed by Flötteröd & Bierlaire (2013) is an importance sampling algorithm where the sampling probability can be specified explicitly. They mention the generation of choice sets for route choice models as a motivation for the work, but do not investigate it further. We have proposed probabilistic MM methods for smartphone data. In particular, the unimodal MM method proposed in Chapter 3 generates probabilistic route choice observations from GPS data recorded during car trips. In this chapter, these three methods are adapted to the use of GPS data and integrated in a comprehensive and operational route choice modeling framework. We only deal with unimodal route choice, so by default, MM method refers to unimodal MM method presented in Chapter 3, and path refers to unimodal path.

The choice set for route choice model is assumed to include all paths. We propose a new importance path sampling algorithm which is built upon the MHPS technique. Importance sampling of path alternatives is just an intermediate statistical procedure for the route model estimator, unlike the "consideration set" that captures a separate mental process. It yields a consistent model estimator with respect to the universal choice set. Discrete choice models imply compensatory decision making process, in which the decision maker balances the trade-off among attributes of alternatives. Of course, she makes such balances mostly among relevant alternatives. Consequently, relevant alternatives are more useful in identifying the parameters during the model estimation procedure. The new path sampling algorithm is designed to generate relevant alternatives by exploiting GPS data. It yields more precise parameter estimates than

other importance sampling based algorithms.

This chapter is organized as follows. A comprehensive and operational route choice modeling framework for GPS data is proposed in section 5.1, with a focus on the path sampling algorithm. The consistency of the estimated route choice model is empirically tested in section 5.2, and compared with route choice models based on other importance sampling algorithms. Section 5.3 presents a route choice behavior model that is estimated from real smartphone GPS data. Finally, conclusions and future works are discussed in section 5.4.

5.1 Methodologies

We start this section by presenting a route choice modeling framework for GPS data. A logit model with importance sampling of alternatives and path size correction is then introduced. A new sampling algorithm based on MHPS technique is motivated and derived.

5.1.1 Route choice modeling framework for GPS data

Bierlaire & Frejinger (2008) propose a route choice modeling framework for "network-free" route choice data that is not directly associated with the transportation network. The main idea is that the association between the data and the network should be probabilistic to avoid major errors that path imputation using MM algorithms may produce.

A set of route choice data includes multiple route choice observations. Each observation n records a choice decision, measured in "network-free" data i, plus the context information, such as the set of relevant OD pairs S_n. Then the choice probability for the observation n is written as $\Pr(i|S_n)$, and it can be decomposed as:

$$\Pr(i|S_n) = \sum_{s \in S_n} \Pr(s|S_n) \sum_{p \in \mathcal{U}(s)} \Pr(i|p) \Pr(p|\mathcal{U}(s); \boldsymbol{\beta}), \tag{5.1}$$

where

- $\Pr(s|S_n)$ is the probability that the actual OD pair is s.
- $\mathcal{U}(s)$ is the choice set corresponding to OD s.
- $\Pr(i|t_i, p,)$ is a measurement likelihood that calculates the probability that measurement i has been generated by the traveler along path p at time t_i.
- $\Pr(p|\mathcal{U}(s); \boldsymbol{\beta})$ describes a route choice model with unknown parameters $\boldsymbol{\beta}$ to be

estimated.

$\Pr(i|t_i, p)$ is the most important component that is introduced in this framework. We have derived this probability in Chapter 3 when i is GPS data, i.e. $i = \widehat{g}_{1:T}$ and $t_i = \widehat{t}_i$. It deals with the incompleteness and the inaccuracy of the smartphone GPS data without introducing potential biases. The incompleteness is solved by a more realistic probabilistic travel model that captures the traveler's movements in the network, instead of an arbitrary shortest path assumption. The inaccurate data is dealt with in a probabilistic manner. The probabilistic unimodal MM algorithm produces, from a sequence of GPS measurements, a set of candidate paths $\mathcal{P}_n$. $\Pr(i|t_i, p) = 0$ if $p \notin \mathcal{P}_n$. Each candidate path has a pair of OD, S_n is then defined as the collection of them.

The maximum likelihood estimation is applied to this formulation in order to estimate $\boldsymbol{\beta}$ from multiple route choice observations. In this thesis, we assume that $\mathcal{U}(s)$ is the universal choice set that contains all acyclic paths that connect OD s. The rest of this section deals with the discrete choice model $\Pr(p|\mathcal{U}(s); \boldsymbol{\beta})$.

5.1.2 Logit model with sampling of alternatives

McFadden (1978) proves that the logit model can be consistently estimated using a subset of alternatives $\mathcal{C}_n(s) \in \mathcal{U}(s)$ by introducing a sampling correction term. In route choice context, this approach had not been applicable until Frejinger et al. (2009) attempt to construct a path sampling protocol using the random walk (RW) algorithm. Following Ben-Akiva (1993) and Frejinger et al. (2009), for a candidate chosen path p associated with observation n, a general sampling protocol for building a subset $\mathcal{C}_n(s)$ is: first, drawing Ψ_n alternatives with replacement using a sampling algorithm such that the sampling probability $q(j)$ of each path $j \in \mathcal{U}(s)$ is known; and second, deterministically adding the chosen alternative p. Then, a sampling correction term is added to the deterministic part of the utility function:

$$\ln q(\mathcal{C}_n(s)|j) = \ln \frac{k_{jn}}{q(j)}, \tag{5.2}$$

where k_{jn} is the number of times that path j appears in $\mathcal{C}_n(s)$. Here, we want to emphasize that $\mathcal{C}_n(s)$ is merely a set of sampled alternatives from $\mathcal{U}_n(s)$, and the actual choice set is still the universal choice set $\mathcal{U}_n(s)$. In the sampling correction term, the sampling probability $q(j)$ can be replaced by its unnormalized form $b(j)$. If we write $q(j) = \frac{b(j)}{K}$, where $K = \sum_{j' \in \mathcal{U}(s)} b(j')$ denotes the normalizing constant. It is trivial to prove that in the logit model, the normalizing constant K cancels out. Thus path enumeration is avoided.

The consistent estimator requires an adequate number of sampled alternatives in order to achieve sufficient precision for the parameter estimates. The estimator identifies the parameters using $\mathcal{C}_n(s)$, therefore, the relevance of alternatives in $\mathcal{C}_n(s)$ is important for correctly identifying the parameters. Hence, the sampling algorithm should be designed to have a higher probability for sampling relevant alternatives, such that a good composition of $\mathcal{C}_n(s)$ can be generated with a reasonably small number of samples.

The above sampling correction derivation works if only one path p is deterministically added to $\mathcal{C}_n(s)$ in the second step of the sampling protocol. However, it is possible that several observations with different chosen paths are collected; moreover, if the raw choice decision measurement i is GPS, the probabilistic unimodal MM method generates multiple candidate chosen paths. These paths are of course relevant to the traveler, hence would be better to be included in $\mathcal{C}_n(s)$. However, deterministically adding multiple paths will invalidate the sampling correction derived above. Therefore, in this thesis, we rely on simulation, and design a sampling algorithm that assigns higher probabilities to these paths that are available from route choice data. It is based on the MHPS technique that offers a flexible way of constructing a path sampling algorithm.

5.1.3 Sampling of alternatives using RP data

Flötteröd & Bierlaire (2013) propose a Metropolis-Hastings (MH) algorithm to sample paths between a given OD from a predefined sampling distribution. The most important feature of this method is that the path sampling distribution can be defined in an unnormalized form, and path enumeration is avoided. Flötteröd & Bierlaire (2013) propose an exponential function for the unnormalized sampling weight:

$$b(j) = \exp(\omega_1 \delta(j)), \tag{5.3}$$

where, ω_1 is a parameter to be specified by the modeler, and $\delta(j)$ denotes the generalized cost of path j. In route choice context, the path length is often the most convenient and reasonable attribute, hence, Flötteröd & Bierlaire (2013) suggest to use:

$$b(j) = \exp(\omega_1 L_j), \tag{5.4}$$

where L_j denotes the length of path j. The ratio of the probabilities for two different paths being sampled only depends on their length difference. It results in that, with the same ω_1, the sampled alternatives for a longer trip is more concentrated around the shortest path, compared to those for a shorter trip. This is an undesired property thus ω_1 should be adjusted according to the scale of the route choice problem. Flötteröd &

Bierlaire (2013) suggest to use a "scale-invariance" parameter ζ to parameterize ω_1,

$$\omega_1 = \frac{\ln 2}{(\zeta - 1)L_{j_{sp}}} \tag{5.5}$$

The underlying assumption is that a path with the length of $\zeta L_{j_{sp}}$ has half of the probability as the shortest path j_{sp}. This concept is more intuitive, and modelers only need to tune ζ once for all observations instead of tuning ω_1 for each one.

We propose an algorithm that requires a smaller size of $\mathcal{C}_n$, in order to achieve precise parameter estimates. The idea is to include more relevant alternatives in the choice set, because they contribute more to the parameter identification. Intuitively, the (candidate) chosen alternatives available in or generated from RP route choice data are relevant to the traveler. Therefore, we suggest to exploit GPS data to assist the generation of $\mathcal{C}_n(s)$. We calculate the score factor P_j, called *observation score*, from GPS:

$$P_j = \frac{1}{|N^s|} \sum_{n \in N^s} q_n(j), \tag{5.6}$$

where N^s is the set of all route choice observations associated with the same OD s. $q_n(j)$ denotes the probability that j is the chosen path given the observation n, and can be calculated by Equation (3.27) derived in Section 3.3. The *observation score* also works for route choice observations with real chosen paths. Then the probability $q_n(j)$ becomes deterministic, and $q_n(j) = 1$ if j is the chosen path in observation n, and 0 otherwise.

We then define the weight function as

$$b(j) = \exp(\omega_1 L_j + \omega_2 \lambda P_j), \tag{5.7}$$

where L_j is the length of path j; P_j is the *observation score* defined above; ω_1 and ω_2 are parameters to be specified. ω_1 can also be parameterized by using Equation (5.5). λ is calibrated such that when $\omega_2 = 1$, the sampling weight of the shortest path j_{sp} and the path with highest *observation score*, $j_o = \arg\max_{j \in \mathcal{U}(s)} P_j$, are the same: $b(j_o) = b(j_{sp})$, then we have:

$$\lambda = \frac{\omega_1 (L_{j_{sp}} - L_{j_o})}{P_{j_o} - P_{j_{sp}}}. \tag{5.8}$$

Note that (5.4) is a special case of (5.7), since they are equivalent if $\omega_2 = 0$.

Note that it is inappropriate to use observed choices as an input to the procedure for the generation of the choice set. This introduces endogeneity. It is not the case here, as the choice set is composed of *all* paths linking the OD and, therefore, its definition does *not* depend on the observed path. It is only the sampling procedure which is exploiting

the observations. In the next section, numerical experiments show that the proposed algorithm achieves unbiased parameter estimates, even requiring less number of samples than other importance sampling approaches.

This method is capable of dealing with GPS data, and it is integrated with the "network-free" data approach (Bierlaire & Frejinger 2008), and the probabilistic unimodal MM approach to form a comprehensive route choice modeling framework for GPS data. This framework is applied to a set of real smartphone GPS data in section 5.3.

5.2 Numerical analysis

In this section, we aim at evaluating the performance of the importance sampling algorithm that exploits GPS data. The experiment method is described, followed by a case study. The parameters of the proposed sampling algorithm, ζ and ω_2, are discussed. ζ is used to parameterize ω_1, and the discussion on ζ is transferable to ω_1 according to Equation (5.5).

5.2.1 Design of the experiment

We focus on the precision of the parameter estimates. The precision of a parameter estimate can be empirically assessed by a t-test against its true value. If a parameter estimate is significantly different from its true value at 5% significance level (critical value 1.96), we report it as imprecise.

The procedure to perform such an analysis for an algorithm is inspired by Frejinger et al. (2009):

1. Define a transportation network and an OD, where a traveler makes route choice decisions.
2. Postulate a route choice model for the traveler with the specification of each parameter's true value.
3. Generate a number of synthetic choices according to the postulated model.
4. Sample a set of alternatives $\mathcal{C}_n(s)$ for each synthetic choice.
5. Estimate the route choice model, and empirically analyze the precision of each parameter estimate by computing the t-test statistic against its true value.

For the sake of simplicity of this experiment, we don't introduce errors into the choice

data, so the route choice observations are deterministic and consist of the real chosen paths. GPS data are introduced when we process real data in section 5.3.

In this thesis, we want to analyze the performance of different importance sampling based algorithms. The first candidate is the random walk (RW) algorithm, which is the first one that can provide a consistent estimator, and the only one before the MHPS technique is proposed. Based on the MHPS technique, we test different specifications:

- MH^g: the algorithm proposed in this chapter.

- MH^ℓ: MHPS algorithm with only the length factor, i.e. $\omega_2 = 0$, corresponding to (5.4), as proposed by Flötteröd & Bierlaire (2013).

- MH^e: MHPS algorithm with equal probability for every path, $b(j) = 1$. It is pure random sampling instead of importance sampling. This is used as a benchmark, as it is not expected to perform well in practice.

- MH^t: sampling with the true choice probability. Lemp, Ridge & Kockelman (2011) propose a strategic sampling method that suggests to use a choice model, which is estimated with a simple choice set generation algorithm, to approximate the true choice probability distribution. Then the approximated choice probability is used to sample the choice set so as to refine the choice model estimation. We also want to test this method in route choice context when the true choice model is provided. In this case, the sampling weight is defined as $b(j) = e^{V_{jn}}$, if a logit model is specified, and a specification of the deterministic part of the utility function V_{jn} is provided.

5.2.2 Conducting an experiment

We conduct a numerical experiment according to the procedure described above.

Step 1: Brief introduction of the experiment scenario

The experiment scenario is defined as: a traveler makes route choice decisions between an OD in a real transportation network. The network and the OD is shown in Figure 5.1. The network is in Lausanne city center, and there are some traffic lights. The traveler tries to avoid them in making route choice decisions.

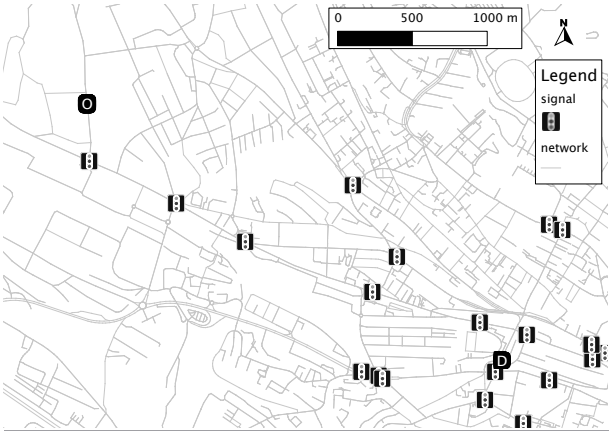


Figure 5.1: The experiment network

Step 2: Postulating a route choice model

We postulate a path size logit (PSL) model (*Model A*) with the deterministic part of the utility function for path p and observation n as

$$V_{pn} = \beta_\ell L_p + \beta_s S_p + \beta_{ps} \ln PS_p, \tag{5.9}$$

where S_i denotes the number of traffic lights along path p; L_p is the length in meter; $\boldsymbol{\beta} = \{\beta_\ell, \beta_s, \beta_{ps}\}$ are the parameters with the values $\{-0.03, -3, 3\}$. The ratio between β_s and β_ℓ, i.e. 100, implies that the traveler would compensate 100 meters of additional drive in order to avoid a traffic light. Throughout this chapter, the scale parameter for the MNL model is set to be 1. PS_i calculates the path size which measures the overlapping between alternatives (Ben-Akiva & Bierlaire 1999):

$$PS_p = \sum_{a \in p} \frac{L_a}{L_p} \frac{1}{\sum_{j \in \mathcal{C}_{ps}} \delta_{aj}}, \tag{5.10}$$

where $\mathcal{C}_{ps}$ denotes a set of paths for the path size calculation; $a \in p$ denotes an arc on path p; dummy variable δ_{aj} equals to one if path j contains arc a, and zero otherwise. Frejinger et al. (2009) argues that the path size should be computed from the universal choice set, and suggests to use a large set of paths for $\mathcal{C}_{ps}$ in practice. Therefore, we decide to sample 1000 paths using MHPS algorithm. The sampling weight for path j is chosen to be $b(j) = e^{-0.01*L_j - 1.0*S_j}$, where the magnitude of the parameters for L_j and S_j is smaller than the ones in the route choice model specification. The objective is to have a higher variance in the sampling probability distribution, thus to generate more distinct paths which have a good coverage over the relevant part of the network (see Figure 5.2 for the visualization). After 1′000′000 *burn-in* samples, every 10′000th sample is drawn and kept.

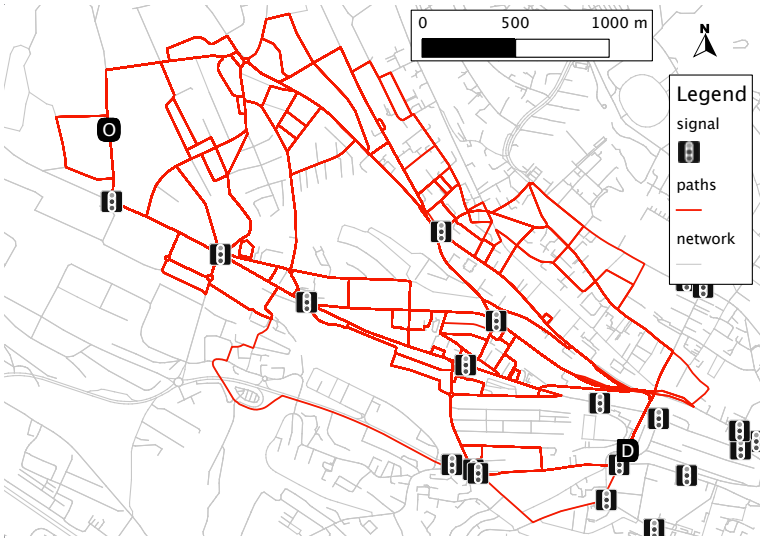


Figure 5.2: Paths $\mathcal{C}_{\text{ps}}$ sampled for path size calculation. They overlap in the network and cover the relevant part of the network.

Step 3: Generating synthetic choices

Given the postulated route choice model, we simulate 50 route choice decisions using MHPS algorithm. Every 100′000th sample is drawn as an route choice decision after a *burn-in* period of 1′000′000. These numbers are both much larger than what are suggested (10′000 and 10′000) by Flötteröd & Bierlaire (2013). Thus independent samples (choices) are guaranteed. The generated choices, along with the shortest path which is not chosen, are plotted in Figure 5.3. We notice that the traveler takes longer paths in order to avoid traffic lights.

Step 4: Sampling alternatives

Section 5.2.1 introduces the importance sampling algorithms to be analyzed. For each algorithm, different specifications of parameters are tested. For RW algorithm, we have tested different Kumaraswamy parameters (Frejinger et al. 2009), $(30, 0)$ and $(50, 0)$ respectively. For MH based algorithms, including MH^{ℓ}, MH^{g}, MH^{t} and MH^{e}, the *burn-in* is 1′000′000 and every 100′000th sample is drawn as an alternative. The specifications of MH^{t} and MH^{e} are straightforward, and do not involve any parameter that needs to be calibrated. We refer to (5.7) for the formulation of the sampling weight function of MH^{ℓ} and MH^{g} algorithms. When $\omega_2 = 0$, the algorithm is MH^{ℓ} indeed. Different specifications of parameters are tested by varying ζ to be 1.052, 1.026, 1.013, and 1.009 (corresponding ω_1: -0.005, -0.01, -0.02, and -0.03), together with ω_2 to be 0, 1, 2, and 3.

We draw 100 random samples for each observation and each specification of each algorithm.

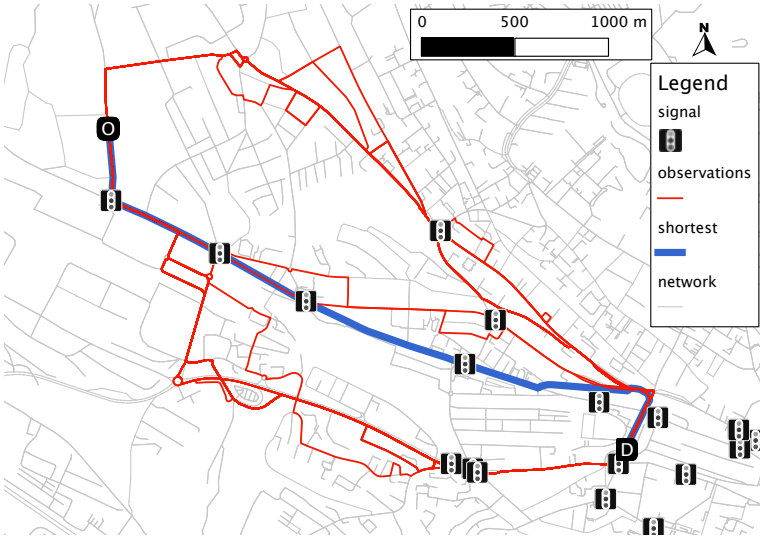


Figure 5.3: Synthetic choices and the (unchosen) shortest path

A set of alternatives $\mathcal{C}_n(s)$ is composed of a number of random samples, plus the chosen alternative. In order to test the effect of the the size of $\mathcal{C}_n(s)$, we draw different numbers of random samples: 5, 10, 20, 30, 50 and 100 respectively.

Step 5: Estimating the models

The model specification is the same as the postulated PSL model. The sampling correction term is added to the deterministic part of the utility function:

$$V_{pn} = \beta_\ell L_p + \beta_s S_p + \beta_{ps} \ln PS_p + \ln \frac{k_{pn}}{b(j)} \tag{5.11}$$

The parameters β_ℓ, β_s and β_{ps} are estimated. The sampling weight $b(p)$ is calculated according to the corresponding sampling algorithm. The results are analyzed below.

5.2.3 Result analysis

The overall performance is first illustrated with the best results that have been obtained from each algorithm. Detailed analyses on the proposed algorithm are then presented.

Overall performance

Table 5.1 provides an insight about the performance of different algorithms. Each $\mathcal{C}_n(s)$ takes 100 random draws. First, although different parameter settings are tested for

Table 5.1: Comparison among all algorithms with 100 random draws

	True	RW	MH^e	MH^t	$MH^\ell(1.026, 0)$[a]	MH^g (1.026, 2)
$\widehat{\beta}_{ps}$[b]	3.00	-[c]	-	3.1(0.64)[d]	2.2(2.2)*[e]	3.0(0.032)
$\widehat{\beta}_s$	−3.00	-	-	−3.3(1.5)	−2.5(0.94)	−3.1(0.14)
$\widehat{\beta}_l$	−0.03	-	-	−0.038(2.5)*	−0.027(0.35)	−0.033(0.77)

[a] The figures in parentheses are the values of (ζ, ω_2) for algorithms MH^ℓ and MH^g.
[b] $\widehat{\beta}$ denotes the estimate of parameter β.
[c] A cell with "-" indicates that the estimate is unavailable because the corresponding model is unidentifiable.
[d] The figure in parentheses is the t-test of the estimate against its true value.
[e] * is noted, if the parameter estimate is statistically different from its true value at 5% significance level (t-test > 1.96).

RW, the model is unidentifiable. In fact, the sampled paths contain a lot of loops with u-turns and detours, so are not realistic. This confirms the findings of Schüssler (2010). Algorithm MH^e generates uniformly random paths, most of which are too long to be relevant for the decision maker. Therefore, the corresponding model is not identifiable either. MH^t algorithm with the true choice probability yields one imprecise parameter estimate, i.e. β_ℓ. The reason is that the sampling distribution has a smaller variance, thus the sampled alternatives do not contain an adequate number of distinct paths for correctly identifying the parameters.

Several settings for MH^ℓ and MH^g are tested, the detailed results are reported in Table 5.2. Table 5.1 presents the best results of them. The best result obtained from MH^ℓ, with settings $(\zeta, \omega_2) = (1.026, 0)$, still contains imprecise parameter estimates. The proposed algorithm MH^g performs the best as all the parameter estimates are precise, with settings $(\zeta, \omega_2) = (1.026, 2)$.

Comparison between MH^ℓ and MH^g

We analyze the added value of using route choice data in alternative sampling, and compare the performance of MH^ℓ and MH^g with different settings. The effect of the sample size is also analyzed, by drawing different numbers of random paths in $C_n(s)$. Table 5.2 reports the parameter estimates and their t-test statistics.

We first find that when the magnitude of ζ is too low ($\zeta = 1.009$) or too high ($\zeta = 1.052$), it is difficult to get an identifiable model and precise parameter estimates, although different values of ω_2 are tried. Precise parameter estimates are achieved only when $\zeta = 1.026$ or $\omega_2 = 1.013$. ζ controls the spread of the sampling algorithm, hence the heterogeneity of the drawn alternatives. Figure 5.4 visualizes random paths produced

by MH^{ℓ} using different ζ settings, while fixing $\omega_2 = 0$. Although $\Psi_n = 100$ is set for each of them, different numbers of distinct paths are generated. When ζ is small, e.g. $\zeta = 1.009$, the sampled paths tend to concentrate around the shortest path, and they do not cover relevant paths in terms of other factors, e.g. less traffic lights. The larger the value of ζ, e.g. $\zeta = 1.052$, the wider the paths spread in the network. Consequently, the sampled alternatives include too many unattractively long paths that do not contribute to reveal the decision maker's trade-off. ζ should be in a reasonable range, in this case $\zeta \in [1.013, 1.026]$, such that $\mathcal{C}_n(s)$ includes an adequate number of distinct relevant alternatives that help to identify the compensatory decision process.

When $\zeta = 1.026$ or $\zeta = 1.013$, MH^g ($\omega_2 > 0$) in general performs not worse than MH^{ℓ} ($\omega_2 = 0$). Even if $\mathcal{C}_n(s)$ contains 100 random draws, MH^{ℓ} still gets imprecise parameter estimates for all settings. However, several MH^g settings yield precise parameter estimates, including (ζ, ω_2) being $(1.026, 1)$, $(1.026, 2)$, $(1.026, 3)$ and $(1.013, 1)$. The performances of both algorithms are sensitive to ζ, and setting $\omega_2 = 1$ in general yields better results. This is due to the intuitive motivation of ω_2, i.e. when $\omega_2 = 1$, the path with the highest *observation score* gets the same sampling weight as the shortest path.

The impact of the sample size has two aspects. First, the variance of a parameter estimate decreases as the sample size increases, as expected. Second, MH^g requires less samples in order to achieve an identifiable model and precise parameter estimates. We can conclude that MH^g has an overall better performance than MH^{ℓ} and other importance sampling based algorithms. We also tried to construct probabilistic observations by introducing detours to the deterministic choices, and the results are still robust. However, the performance is sensitive to the ζ parameter which controls the spread of the sampled alternatives. Hence, the specification of this parameter is discussed below.

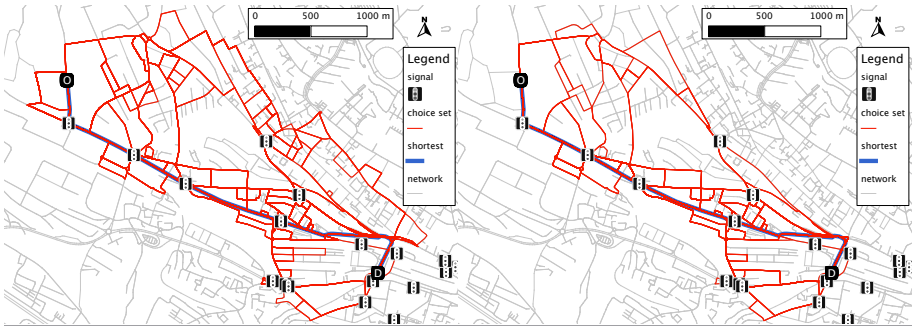
5.2.4 Discussions on the parameters' specification

In the above experiment, taking ζ between 1.013 and 1.026 yields best results. This conclusion is valid for another postulated route choice model (*Model B*) with the same specification except that the traffic light is not considered as a factor, i.e. $\beta_s = 0$. We have performed the same experiment for *Model B*, as we have done for *Model A*, and report the results in Table 5.3. We first notice that many settings yield precise parameter estimates, and less samples are needed than *Model A*. The results also suggest that the performances of MH^{ℓ} and MH^g algorithms are less sensitive to the ζ parameter. With a sufficient number of samples, e.g. 100, some MH^{ℓ} specifications also achieve precise parameter estimates, such as ζ being 1.026, 1.013 or 1.009. Especially when the value of ζ is very small ($\zeta = 1.009$), the parameter estimates are also precise, although higher variance is observed. In *Model B*, length is the only factor that affects the choice

Table 5.2: Estimation results for *Model A*

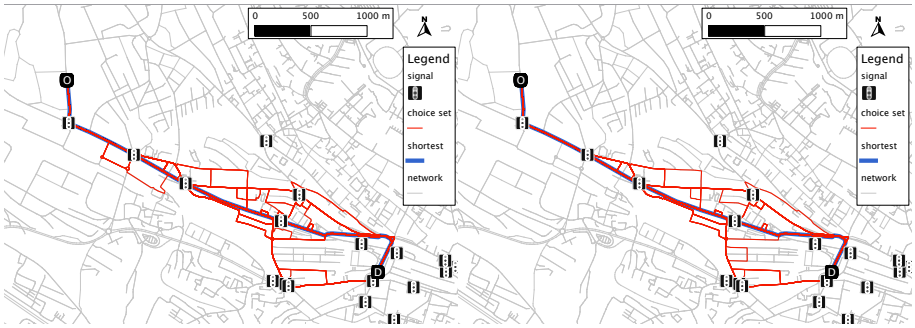
ζ,ω_2 [a]	$\widehat{\beta}$	5 [b]	10	20	30	50	100
1.052, 0		- [c]	-	-	-	-	-
1.052, 1		-	-	-	-	-	-
1.052, 2		-	-	-	-	-	-
1.052, 3		-	-	-	-	-	-
1.026, 0	$\widehat{\beta}_{ps}$	-	-	-	2.2(1.5)	2.3(1.3)	2.2(2.2)*
	$\widehat{\beta}_{s}$	-	-	-	-2.5(0.91)	-2.9(0.21)	-2.5(0.94)
	$\widehat{\beta}_{l}$	-	-	-	-0.029(0.12)	-0.032(0.16)	-0.027(0.35)
1.026, 1	$\widehat{\beta}_{ps}$	-	-	-	-	3.5(0.82)	2.4(1.6)
	$\widehat{\beta}_{s}$	-	-	-	-	-2.9(0.12)	-2.7(0.67)
	$\widehat{\beta}_{l}$	-	-	-	-	-0.03(0.077)	-0.03(0.069)
1.026, 2	$\widehat{\beta}_{ps}$	-	2.7(0.88)	2.4(3.0)*	2.4(3.5)*	2.6(2.1)*	3.0(0.032)
	$\widehat{\beta}_{s}$	-	-6.0(2.9)*	-4.2(1.3)	-2.9(0.27)	-2.7(1.3)	-3.1(0.14)
	$\widehat{\beta}_{l}$	-	-0.042(1.9)	-0.028(0.38)	-0.025(1.7)	-0.025(1.6)	-0.033(0.77)
1.026, 3	$\widehat{\beta}_{ps}$	2.8(0.87)	2.6(2.3)*	2.8(0.56)	2.9(0.45)	2.8(0.81)	2.8(0.78)
	$\widehat{\beta}_{s}$	-3.7(1.3)	-3.7(2.0)*	-2.7(0.66)	-2.9(0.27)	-3.6(0.56)	-3.6(0.9)
	$\widehat{\beta}_{l}$	-0.028(0.72)	-0.024(1.7)	-0.027(0.4)	-0.027(0.27)	-0.028(0.3)	-0.026(0.55)
1.013, 0	$\widehat{\beta}_{ps}$	-	-	-	-	-	5.4(1.8)
	$\widehat{\beta}_{s}$	-	-	-	-	-	-5.6(2.0)*
	$\widehat{\beta}_{l}$	-	-	-	-	-	-0.056(2.2)*
1.013, 1	$\widehat{\beta}_{ps}$	-	-	5.0(2.1)*	5.0(2.1)*	3.0(0.08)	3.1(0.41)
	$\widehat{\beta}_{s}$	-	-	-5.3(2.1)*	-5.5(2.0)*	-2.8(0.46)	-3.0(0.035)
	$\widehat{\beta}_{l}$	-	-	-0.057(2.3)*	-0.056(2.0)*	-0.032(0.23)	-0.03(0.098)
1.013, 2	$\widehat{\beta}_{ps}$	3.4(0.57)	0.36(1.1)	-	2.8(0.39)	2.8(0.21)	3.4(0.64)
	$\widehat{\beta}_{s}$	-9.4(8.1)*	-9.0(5.6)*	-	-7.3(4.8)*	-6.2(3.2)*	-6.0(2.6)*
	$\widehat{\beta}_{l}$	-0.043(1.2)	0.0093(0.86)	-	-0.03(0.0072)	-0.025(0.24)	-0.037(0.41)
1.013, 3		-	-	-	-	-	-
1.009, 0	$\widehat{\beta}_{ps}$	-	-	-	-	7.6(24.0)*	4.5(3.9)*
	$\widehat{\beta}_{s}$	-	-	-	-	-9.2(52.0)*	-6.1(10.0)*
	$\widehat{\beta}_{l}$	-	-	-	-	-0.07(10.0)*	-0.047(3.2)*
1.009, 1	$\widehat{\beta}_{ps}$	5.4(3.5)*	4.8(3.4)*	4.2(4.2)*	3.8(4.6)*	3.2(1.2)	-
	$\widehat{\beta}_{s}$	-6.1(3.4)*	-5.4(3.5)*	-5.1(3.9)*	-5.5(4.5)*	-4.6(3.1)*	-
	$\widehat{\beta}_{l}$	-0.065(5.4)*	-0.059(5.0)*	-0.054(4.5)*	-0.052(4.4)*	-0.04(1.4)	-
1.009, 2		-	-	-	-	-	-
1.009, 3		-	-	-	-	-	-

[a] If $\omega_2 = 0$, the algorithm is MH^{ℓ}; otherwise, it is MH^{g}.
[b] Number of random samples in $\mathcal{C}_n(s)$.
[c] A "-" cell indicates an unidentifiable model.



(a) $\zeta = 1.052$, 100 distinct paths

(b) $\zeta = 1.026$, 100 distinct paths



(c) $\zeta = 1.013$, 89 distinct paths

(d) $\zeta = 1.009$, 64 distinct paths

Figure 5.4: 100 random samples from MH^{ℓ} with different ζ. Different numbers of distinct paths are generated.

decisions, therefore, only short paths are relevant to the decision maker. As a result, even though $\zeta = 1.009$ results in sampled alternatives that concentrates around the shortest paths, it still yields precise parameter estimates. However, in *Model A* where traffic light is also an important factor, longer paths with less traffic lights may also be relevant to the decision maker. Thus, the value of ζ should be larger such that the sampling spreads wider in the network, and relevant longer paths, possibly having less traffic lights, have higher chance to be sampled.

Comparing the estimation results for *Model A* and *Model B*, we can find that ω_2 plays more important roles when the chosen paths deviate more from the shortest path. In *Model B*, only short paths are relevant to the traveler, so even without the usage of route choice data, MH^{ℓ} algorithm samples relevant short paths for the sampled alternatives, and yields precise parameter estimates. *Model A* relies on MH^{g} to sample relevant alternatives available from route choice data. In summary, we suggest that the specification of the algorithm should depend on the route choice behavior. If the observed route choice decisions deviate from the shortest path, then: the value of ζ should be small (low magnitude of ω_1 correspondingly); $\omega_2 > 0$ is suggested; a large number of sampled alternatives should be used.

5.3 Route choice modeling application on real GPS data

In this section, the proposed route choice modeling framework is applied to a set of real smartphone GPS route choice data, which are extracted from the Lausanne Data Collection Campaign database. The dataset contains traces about 19 trips performed by the same driver, over a period of 2 months, mostly commuting and shopping. The experiment is performed in three steps according to the proposed framework. First, the probabilistic unimodal MM method is used to generate probabilistic path observations from GPS data. Second, MH^{g} algorithm generates a sample of alternatives $\mathcal{C}_n(s)$ for each candidate path observation. The *burn-in* period is 500′000. After that, every 100′000th sample is drawn as an alternative. Each $\mathcal{C}_n(s)$ is composed of 100 draws. $\zeta = 1.026$ and $\omega_2 = 1$. Third, a route choice model is specified and estimated based on the "network-free" likelihood estimation. The transportation network is the Switzerland road network extracted from *OpenStreetMap.org* (OSM). Table 5.4 reports some statistics about the trips, as well as the probabilistic path observations and the sampled alternatives.

The deterministic part of the utility function of candidate path p is specified as:

$$V_{pn} = \beta_{ps} \ln PS_p + \beta_{ll} LL_p + \beta_{hl} HL_p + \ln \frac{k_{pn}}{b(p)}. \tag{5.12}$$

The set of paths for calculating the path size PS_p is the sampled alternatives $\mathcal{C}_n(s)$. The

Table 5.3: Estimation results for *Model B*

ζ,ω_2		5	10	20	30	50	100
1.052, 0		-	-	-	-	-	-
1.052, 1		-	-	-	-	-	-
1.052, 2		-	-	-	-	-	-
1.052, 3	$\widehat{\beta}_{ps}$	-	-	-	-	-	3.0(0.045)
	$\widehat{\beta}_{s}$	-	-	-	-	-	0.62(1.9)
	$\widehat{\beta}_{l}$	-	-	-	-	-	-0.037(1.0)
1.026, 0	$\widehat{\beta}_{ps}$	-	-	-	3.2(0.4)	3.4(0.88)	3.2(0.49)
	$\widehat{\beta}_{s}$	-	-	-	-0.37(2.7)*	-0.15(0.97)	-0.12(0.84)
	$\widehat{\beta}_{l}$	-	-	-	-0.034(0.97)	-0.036(1.6)	-0.033(0.91)
1.026, 1	$\widehat{\beta}_{ps}$	-	-	-	-	2.6(1.4)	2.7(1.3)
	$\widehat{\beta}_{s}$	-	-	-	-	-0.059(0.39)	-0.2(1.5)
	$\widehat{\beta}_{l}$	-	-	-	-	-0.027(1.2)	-0.028(0.86)
1.026, 2	$\widehat{\beta}_{ps}$	-	-	-	-	3.2(0.58)	3.6(1.5)
	$\widehat{\beta}_{s}$	-	-	-	-	-0.28(2.2)*	-0.22(1.8)
	$\widehat{\beta}_{l}$	-	-	-	-	-0.031(0.59)	-0.034(1.6)
1.026, 3	$\widehat{\beta}_{ps}$	-	3.7(0.56)	3.0(0.031)	2.8(0.45)	2.7(0.88)	3.0(0.08)
	$\widehat{\beta}_{s}$	-	0.26(0.64)	-0.02(0.09)	-0.011(0.054)	-0.015(0.085)	-0.11(0.7)
	$\widehat{\beta}_{l}$	-	-0.034(0.52)	-0.03(0.13)	-0.028(0.64)	-0.028(0.99)	-0.029(0.39)
1.013, 0	$\widehat{\beta}_{ps}$	7.0(1.1)	4.2(1.1)	3.4(0.89)	3.4(0.93)	3.7(1.4)	3.3(0.8)
	$\widehat{\beta}_{s}$	-2.7(1.4)	-0.75(1.2)	-0.063(0.21)	-0.12(0.39)	-0.15(0.61)	-0.16(0.81)
	$\widehat{\beta}_{l}$	-0.05(1.2)	-0.039(1.5)	-0.034(1.2)	-0.033(0.86)	-0.034(1.2)	-0.032(0.66)
1.013, 1	$\widehat{\beta}_{ps}$	-	4.5(1.2)	4.5(1.3)	3.5(1.1)	3.1(0.2)	3.1(0.37)
	$\widehat{\beta}_{s}$	-	0.6(1.8)	0.42(1.3)	0.07(0.17)	-0.091(0.29)	0.006(0.029)
	$\widehat{\beta}_{l}$	-	-0.04(1.2)	-0.038(1.2)	-0.031(0.23)	-0.028(0.61)	-0.03(0.12)
1.013, 2	$\widehat{\beta}_{ps}$	4.9(1.7)	3.4(0.59)	3.6(1.1)	3.5(1.2)	3.1(0.48)	3.1(0.38)
	$\widehat{\beta}_{s}$	-1.0(3.4)*	-0.45(1.4)	-0.54(2.0)*	-0.48(2.4)*	-0.28(2.0)	-0.2(1.3)
	$\widehat{\beta}_{l}$	-0.045(2.6)*	-0.034(0.65)	-0.035(1.0)	-0.034(1.3)	-0.031(0.29)	-0.029(0.31)
1.013, 3	$\widehat{\beta}_{ps}$	2.2(2.4)*	2.6(1.4)	3.4(0.41)	3.3(0.45)	3.5(0.78)	3.8(1.5)
	$\widehat{\beta}_{s}$	0.53(2.5)*	0.33(1.5)	-0.14(0.25)	-0.12(0.25)	-0.15(0.52)	-0.17(0.87)
	$\widehat{\beta}_{l}$	-0.018(2.8)*	-0.019(4.5)*	-0.03(0.013)	-0.029(0.15)	-0.031(0.18)	-0.035(1.3)
1.009, 0	$\widehat{\beta}_{ps}$	-	4.2(0.73)	4.7(1.1)	4.6(1.0)	4.9(1.3)	4.3(1.2)
	$\widehat{\beta}_{s}$	-	0.49(2.0)*	0.64(2.2)*	0.42(1.7)	0.46(1.8)	-0.88(1.2)
	$\widehat{\beta}_{l}$	-	-0.035(0.52)	-0.027(0.32)	-0.024(0.88)	-0.024(0.98)	-0.032(0.33)
1.009, 1	$\widehat{\beta}_{ps}$	10.0(3.5)*	-	6.5(3.0)*	4.0(0.79)	3.6(0.61)	2.4(1.5)
	$\widehat{\beta}_{s}$	3.6(4.1)*	-	1.9(3.4)*	1.1(3.5)*	0.062(0.098)	0.016(0.054)
	$\widehat{\beta}_{l}$	-0.029(0.095)	-	-0.018(1.4)	-0.02(2.3)*	-0.019(2.3)*	-0.023(1.6)
1.009, 2	$\widehat{\beta}_{ps}$	-	4.0(0.8)	4.0(0.73)	3.1(0.075)	3.1(0.15)	2.7(0.74)
	$\widehat{\beta}_{s}$	-	1.4(4.0)*	0.57(1.8)	0.26(1.0)	0.09(0.29)	-0.32(1.0)
	$\widehat{\beta}_{l}$	-	-0.021(1.6)	-0.034(0.5)	-0.032(0.32)	-0.031(0.31)	-0.027(0.72)
1.009, 3	$\widehat{\beta}_{ps}$	2.2(1.5)	2.0(2.1)*	2.4(1.2)	2.6(1.3)	3.2(0.22)	3.5(0.78)
	$\widehat{\beta}_{s}$	1.2(2.8)*	0.96(2.9)*	0.44(1.4)	0.39(1.4)	-0.044(0.079)	-0.25(0.55)
	$\widehat{\beta}_{l}$	-0.025(0.57)	-0.025(0.72)	-0.025(0.87)	-0.025(1.3)	-0.029(0.13)	-0.03(0.059)

Table 5.4: Statistics of the 19 trips

	Min	Average	Max
Number of GPS points per trip	16	36	58
Approximate[a] travel time per trip [second]	179	397	795
Number of candidate path observations	5	22	50
Number of relevant OD pairs	1	9	19
Average[b] length of candidate path observations [m]	1'930	3'980	6'420
Number of distinct sampled paths	1	62	100

[a] The travel time is approximated by the difference between the timestamps of the first and the last GPS points.
[b] Weighted by the candidate path probability.

Table 5.5: Estimation result

Coefficient	Value	Robust Std. Error	Robust t-test	p value
β_{ps}	2.41	0.780	3.10	0.00
β_{ℓ}	-0.0293	0.00780	-3.76	0.00
β_{h}	-0.0268	0.00701	-3.83	0.00

Number of observations: 19
Null log likelihood: -774.976
Final log likelihood: -733.951
Adjusted rho-square: 0.049
Model estimated by BIOGEME (Bierlaire 2003)

roads in the transportation network are classified into two categories. The high class is fast roads, including the following types that are described by OSM (2012): motorway, trunk, and primary roads. The low class includes the rest. LL_p and HL_p denotes the length (in meter) of low class and high class roads on path p respectively. $\ln \frac{k_{pn}}{b(p)}$ is the sampling correction term.

Table 5.5 reports the coefficient estimates. All coefficients have their expected signs. Positive path size coefficient is consistent with the established route choice theory. The magnitude of $\beta_{\ell\ell}$ is higher than $\beta_{h\ell}$, indicating that people are more sensitive to the length of the lower class roads. In order to test the significance of the difference, another model is specified with a null hypothesis $\beta_{\ell} = \beta_{h\ell}$. The likelihood ratio test suggests that the null hypothesis should be rejected at 90% significance level, thus the difference is statistically significant.

The data set in this experiment is relatively small, due to the limited computational capability of the implementations of the probabilistic MM algorithm and the MHPS algorithm. Indeed, it takes 25 hours to sample alternatives for the total 414 candidate chosen paths of 19 trips, on a computing server using 24 CPUs (3.33GHz). Moreover,

although we have access to a large smartphone data set, we have little information about the true transport mode of the data. It is the only dataset that is known to have been produced from driving of a person. However, we have illustrated the feasibility of the overall approach on a real data set. Better computational efficiency will be targeted in the future.

5.4 Conclusions and future works

In this chapter, we have presented a comprehensive and operational route choice modeling framework for RP GPS data. This framework integrates three major components, and includes necessary modifications to each such that they are applicable to GPS data. First, the probabilistic unimodal MM method is used to process the GPS data. Second, the "network-free" data approach defines a model estimation framework. Third, a new algorithm for sampling path alternatives is proposed to exploit RP route choice data.

We have performed experiments for analyzing the performance of different importance sampling algorithms using a real transportation network. We have empirically analyzed the precision of the parameter estimates resulted from each algorithm, and compared the number of samples that is needed for achieving precise parameter estimates. The conclusion is that the proposed algorithm exploiting GPS data performs the best, especially when the route choice decisions often deviate from the shortest path. We have also discussed the calibration of the parameters of the proposed algorithm. The proposed route choice modeling framework is applied to a set of real smartphone GPS data. It shows the viability of applying the proposed methods to real GPS data.

Future work includes more investigations on the ω_1 or ζ parameter which controls the spread of the sampled alternatives. The proposed algorithm gives higher probability only to chosen paths, however, similar paths in terms of physical overlaps will also be considered in the future. The proposed algorithm will be tested on larger problems with more observations. The computational speed of the prototype software must first be improved.

6 Conclusions and discussions

This section concludes the thesis, followed by discussions on future directions in research and application.

6.1 Conclusions

Through a literature review, we identify the challenges and opportunities in the field of using smartphone data for route choice behavior modeling. They motivate the research objective of this thesis, which is exploiting smartphone data for route choice models.

Path observations are essential input for discrete route choice models. Thus, the traveled paths need to be inferred from the GPS data, by matching it to the transportation network. In order to account for the inaccuracy and sparsity of smartphone GPS data, a probabilistic unimodal map-matching method is proposed to generate probabilistic path observations. Each probabilistic path observation is produced from a sequence of GPS data recorded during a unimodal travel, and is composed of a set of candidate paths and a measurement likelihood for each path. The measurement likelihood is realistic, and can be applied with "network-free"' route choice modeling approach. The numerical experiments show that this method is effective and robust in dealing with sparse and inaccurate smartphone GPS data.

The unimodal map-matching assumes that the corresponding travel is unimodal and the mode is known. In reality, however, the travel can be multimodal and the modes are usually unknown. Therefore, the unimodal map-matching method is extended to multimodal context. The multimodal map-matching method infers the traveled path and the mode of each road simultaneously from various kinds of data. The proposed framework is capable of dealing with any type of data as long as it provides relevant location or transport mode information. This thesis implements GPS, BT and ACCEL

data. Because the path and the mode are inferred simultaneously, the structure of the network also helps to identify the transport mode. Several examples are visualized to illustrate the effectiveness of the proposed method in dealing with multimodal trips. Empirical analyses show the capability of the proposed method in correctly identifying the transport mode, and the contributions of rich smartphone data.

The path observations inferred from the proposed map-matching methods are applied in route choice modeling. We further exploit GPS data to specify the route choice set. We do not attempt to generate a "consideration set", but instead to employ importance sampling of alternatives for the model estimation. Based on the MH path sampling technique, relevant path alternatives are sampled by exploiting GPS data. Empirical analysis shows that this algorithm yields more precise parameter estimates than other importance sampling based algorithms. Probabilistic unimodal map-matching method, "network-free" route choice modeling approach, and the new importance sampling algorithm are integrated to build a complete route choice modeling framework for GPS data. This framework is operational, as we show a route choice model estimated from a set of real smartphone GPS data.

6.2 Future directions

The methods proposed in this thesis link the smartphone data and route choice models. This section discusses future topics that can further exploit rich smartphone data and the proposed methods in the research and applications of route choice models.

Information inference

In this thesis, we infer route choice decisions from various kinds of smartphone data, including GPS, BT and ACCEL. The inference methods also exploit transport networks. The proposed inference methods should be extended so as to account for more external data sources. For example, instead of specifying the uniform distribution for the prior probability in the candidate path generation algorithms, simple route choice models estimated from historical or external data sources will improve the results. In many cities, public transport schedules and real time public transport information are available. Such information is particularly useful in recognizing the public transport.

Rich smartphone data can also be used to understand the route choice context and the smartphone user's socio-economic characteristics. For example, the social network interaction can be learned from call logs and SMS logs, and this information is useful in predicting future destinations (De Domenico, Lima & Musolesi 2012). The location

data can be combined with the point of interest data so as to reveal the smartphone user's activity at the destination. Calendar entries even provide more precise information about activities. The personality and the emotional status of the smartphone owner can be learned from the played songs (North & Hargreaves 2008). The BT sensor can detect accompanying persons during travels, via their BT capable smartphones. These factors may affect the route choice decisions, and should be considered for building more precise route choice models. Systematic methods must be developed for inferring them from the raw data.

Route choice models

This thesis presents a complete route choice modeling framework with a rather simple example of the model specification. Future work should aim at more precise route choice models by including more path attributes, route choice context information, and the traveler's socio-economic characteristics. Some of them can be added to the utility function directly, such as path attributes. But as the same problem that we encounter in inferring route choice decisions, some information is not explicitly observable, and the measurements are recorded with errors. For example, in discrete choice modeling, we are also interested in psychometric variables that affect the choice decisions. Latent variable models should be specified using smartphone data to be indicators of unobservable psychometric variables.

With the multimodal MM method proposed in Chapter 3, route choice decisions for multimodal route choice models are available. But multimodal route choice models capture more complex behavior, which requires more variables for the model specification. The feasibility of inferring adequate information from smartphone data needs to be investigated.

Application

Based on the smartphone platform, the proposed methods can be used in various applications. First, people update their route choice decisions under the influence of real time traffic information provided by smartphone APPs. This dynamic route choice decision can be observed by recording the smartphone user's access to real time traffic information, and her corresponding route choice decisions. Second, navigation APPs can incorporate route choice models to provide customized route recommendations. The smartphone user's preferred route can be predicted by the route choice model estimated using data from the phone. The predicted route incorporates both the real time traffic information and the smartphone user's preferences, and is recommended to the user

instead of the simple fastest or shortest route recommendations.

Privacy is a major concern when we collect data from smartphones. By just visualizing the rich smartphone data, a lot of private information becomes straightforward. For example, where are the smartphone owner's home and office; what are her most visited shopping places; when did she go to ski last time, etc. Nonetheless, since smartphones have more and more powerful computing units, the behavior learning process will be able to be performed locally, without uploading sensitive raw data to a centralized server. Modelers only need to acquire abstract route choice models which have already been estimated on the smartphones.

A List of available smartphone data

The following lists the data recorded by *EPFLScope* in the Lausanne Data Collection Campaign. The descriptions involve standard terminologies, and we suggest the readers to look up the detailed explanations from the Internet.

- gps
  - coordinates
  - speed, heading
  - accuracy indicators
  - time since the gps was booted
- wlan: wifi access points in range
  - SSID
  - mac address
  - channel
  - security
  - opmode
- gpswlan: fake gps records which is actually the gps of nearby wifi
  - coordinates
  - mac: wifi used for the fake gps record
- gsm: cell tower which the phone is connected to
  - cell id, network code, area code, country code
  - signaldbm: real reception signal strength
  - signal: displayed signal level level
- bt: bluetooth devices in range
  - mac
  - name
- ambient sound
- accelerometer: acceleration recorded from the 3-axis acceleromter
- calllog: phone calls and messages
  - status: only for messages (sent or not)
  - direction: incoming or outgoing
  - description
  - number
  - name
  - contact: pointer to contacts data

  - duration
- sms: sms specific information
  - box: folder in which the sms is located
  - status
  - total length
  - letternbr: statistics on word length
  - address: the receipient or the receiver number
- calendar
  - begin of the event
  - title
  - location
  - status: confirmed or not
  - type, class
  - last_mod: last modification time of the entry
  - history of modification
- contact
  - first name, last name
  - mobile, tel: list of phone numbers
  - last_mod: last modification time of the entry
  - history
- media: created media files (pictures, movies)
  - time: creation time
  - file name
  - file size
- mediaplay: played songs and videos
  - Album, artist, title, tracks, tags
  - Uri: file location
  - State: play, pause, etc
  - Duration
- users: list of participants
  - username, userid
  - source: IMEI of phone
  - pointer to phone number and MAC address
  - consumer segment
- process: current processes
  - application: information about front application
  - event: started, closed, view, foreground
  - uid: unique application id assigned by Symbian
  - name: application name
- state: collection client internal state
  - state: client internal state
  - reason: event that triggered state changes
- sys: information from operating system
  - profile: general, silent, etc
  - battery: percentage of available battery
  - charging: yes or no
  - c,d,e,y,z: free space on different long-term storage space
  - inactive: time since last interaction
  - ring: ringer type
  - freeram

B Unimodal map matching examples

Figures B.1 to B.6 show 6 more probabilistic MM results, in addition to the 4 shown in Section 3.3.

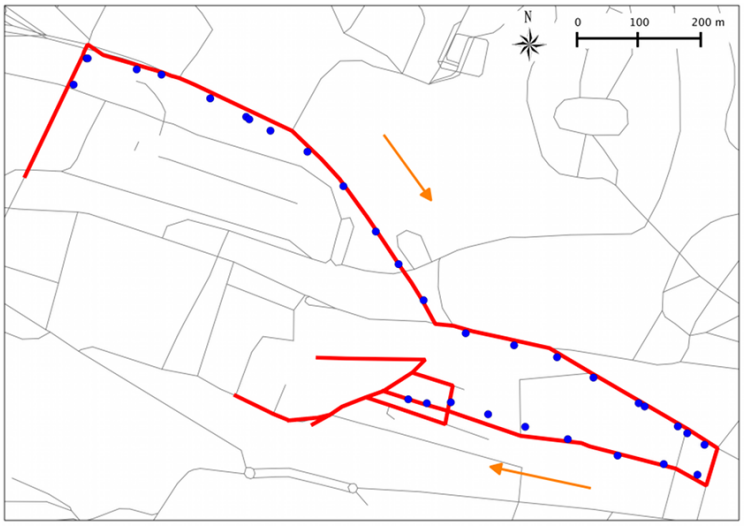


Figure B.1: trip 3 (29 paths)

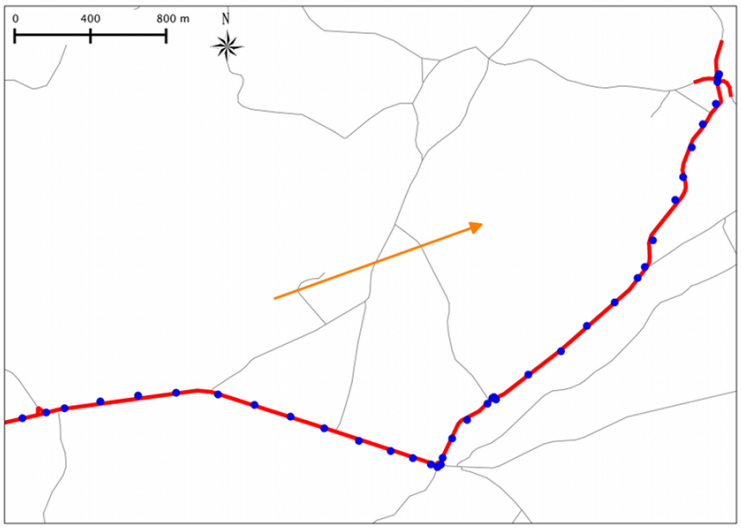


Figure B.2: trip 5 (22 paths)

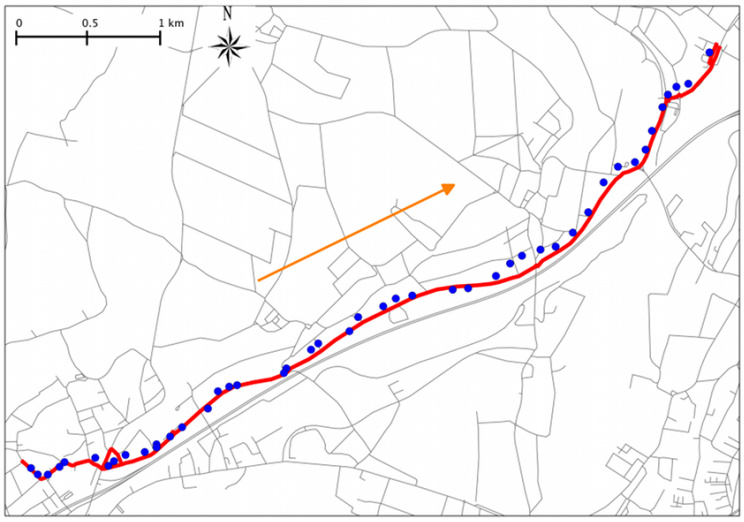


Figure B.3: trip 6 (6 paths)

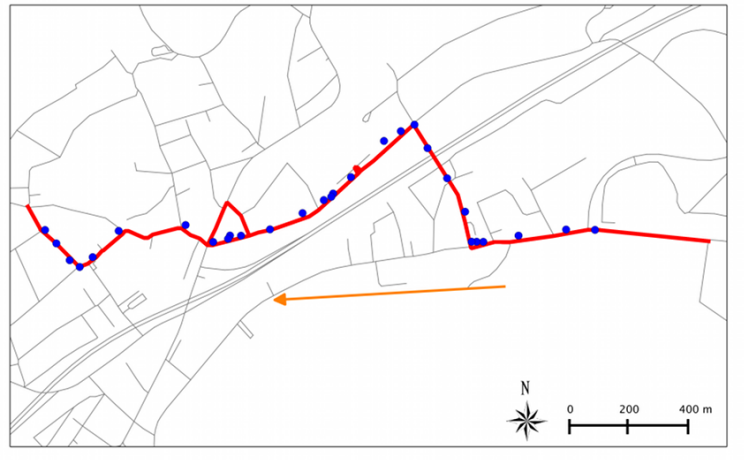


Figure B.4: trip 7(12 paths)

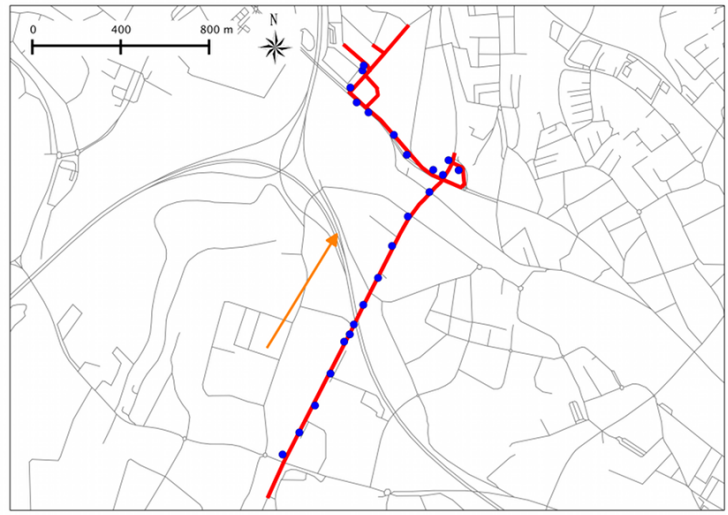


Figure B.5: trip 8 (13 paths)

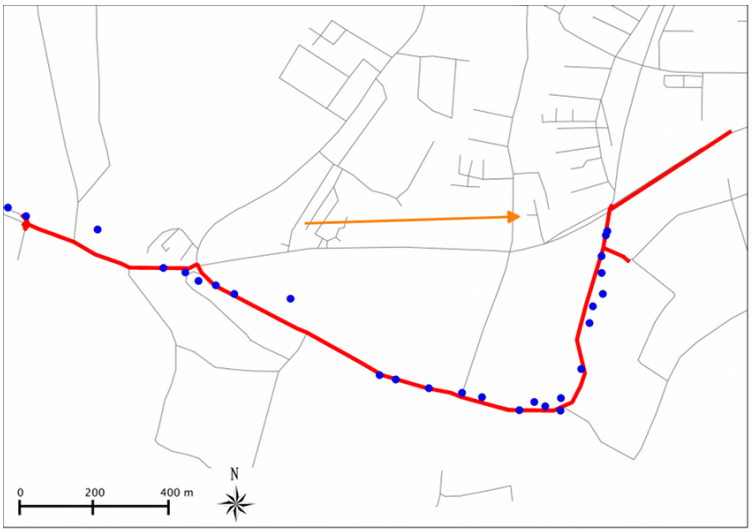


Figure B.6: trip 9 (36 paths)

Bibliography

Abdel-aty, M. A. & Abdalla, M. F. (2006). Examination of multiple mode / route choice paradigms under atis, *IEEE Transcation on intelligent transportation systems.* **7**(3): 332–348.

Abdel-Aty, M. a., Kitamura, R. & Jovanis, P. P. (1997). Using stated preference data for studying the effect of advanced traffic information on drivers' route choice, *Transportation Research Part C: Emerging Technologies* **5**(1): 39–50.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0968090X9600023X*

Azevedo, J., Santos Costa, M., Silvestre Madeira, J. & Vieira Martins, E. (1993). An algorithm for the ranking of shortest paths, *European Journal of Operational Research* **69**(1): 97–106.

Barceló, J. & Casas, J. (2005). Dynamic network simulation with AIMSUN, *Simulation Approaches in Transportation Analysis* pp. 57–98.
URL: *http://dx.doi.org/10.1007/0-387-24109-4_3*

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

Bekhor, S. & Shiftan, Y. (2010). Specification and estimation of mode choice model capturing similarity between mixed auto and transit alternatives, *Journal of Choice Modelling* **3**(2): 29–49.
URL: *http://jocm.org.uk/index.php/JOCM/article/view/90*

Ben-Akiva, M. (1993). Lecture notes on large set of alternatives.

Ben-Akiva, M., Bergman, M. & Daly, A. (1984). Modeling inter urban route choice behaviour, *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, pp. 299–330.

Ben-Akiva, M., Bergman, M. J., Daly, A. & Ramaswamy, V. (1984). Modeling interurban route choice behavior, *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, Utrecht, The Netherlands, pp. 299–330.

Ben-Akiva, M. & Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions, *in* R. Hall (ed.), *Handbook of Transportation Science*, Kluwer Academic Publishers, pp. 5–34.

Ben-Akiva, M. & Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice, *in* R. Hall (ed.), *Handbook of Transportation Science, Second Edition*, Kluwer Academic Publishers, pp. 7–37.

Ben-Akiva, M., Bierlaire, M., Burton, D., Koutsopoulos, H. N. & Mishalani, R. (2001). Network state estimation and prediction for real-time transportation management applications, *Networks and Spatial Economics* **1**(3-4): 293–318.

Ben-Akiva, M. & Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press Series in Transportation Studies, The MIT Press, Cambridge, MA.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *3rd Swiss Transportation Research Conference*, Ascona, Switzerland.

Bierlaire, M. & Frejinger, E. (2008). Route choice modeling with network-free data, *Transportation Research Part C: Emerging Technologies* **16**(2): 187–198.

Blewitt, G., Heflin, M. B., Webb, F. H., Lindqwister, U. J. & Malla, R. P. (1992). Global coordinates with centimeter accuracy in the international terrestrial reference frame using GPS, *Geophysical Research Letters* **19**(9): 853–856.

Blunck, H., Kjæ rgaard, M. & Toftegaard, T. (2011). Sensing and classifying impairments of GPS reception on mobile devices, *Pervasive Computing* pp. 350–367.
URL: *http://www.springerlink.com/index/PM58864W5542U636.pdf*

Bohte, W. & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands, *Transportation Research Part C: Emerging Technologies* **17**(3): 285–297.

Bovy, P. & Fiorenzo Catalano, S. (2007). Stochastic route choice set generation: behavioral and probabilistic foundations, *Transportmetrica* **3**(3): 173–189.

Bovy, P. H. L. (2007). Modeling route choice sets in transportation networks: a preliminary synthesis.
URL: *http://transp-or2.epfl.ch/tristan/FullPapers/187Bovy.pdf*

Bovy, P. H. L. (2009). On modelling route choice sets in transportation networks: A synthesis, *Transport Reviews* **29**(1): 43–68.

Bovy, P. H. L. & Bradley, M. A. (1985). Route choice analyzed with stated-preference approaches, *Transportation Research Record* (1037): 11–20.
URL: *http://trid.trb.org/view.aspx?id=272229*

Bovy, P. & Hoogendoorn-Lanser, S. (2005). Modelling route choice behaviour in multi-modal transport networks, *Transportation* **32**(4): 341–368.

Broach, J., Dill, J. & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data, *Transportation Research Part A: Policy and Practice* pp. 1–11.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0965856412001164*

Burrell, J. (1968). Multipath route assignment and its application to capacity restraint, *Proceedings of the 4th International Symposium on the Theory of Road and Traffic Flow*.

Carlier, K., Fiorenzo-Catalano, S., Lindveld, C. & Bovy, P. (2003). A supernetwork approach towards multimodal travel modeling, *Proceedings of the 82th Annual Meeting of Transportation Research Board*, Washington, D.C., USA.

Cascetta, E., Nuzzolo, A., Russo, F. & Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks, *Proceedings of the Thirteenth International Symposium on Transportation and Traffic Theory*, Lyon, France, pp. 697–711.

Chu, C. (1989). A paired combinatorial logit model for travel demand analysis, *Proceedings of the 5th World Conference on Transportation Research*, Ventura, USA, pp. 295–309.

Chung, E.-H. & Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys, *Transportation Planning and Technology* **28**(5): 381–401.

Daganzo, C. F. & Sheffi, Y. (1977). On stochastic models of traffic assignment, *Transportation Science* **11**: 253–274.

De Domenico, M., Lima, A. & Musolesi, M. (2012). Interdependence and predictability of human mobility and social interaction, *Proceedings of the Nokia Mobile Data Challenge Workshop*, Newcastle, United Kingdom.

de la Barra, T., Perez, B. & Anez, J. (1993). Multidimensional path search and assignment., *Proceedings of the 21st PTRC Summer Annual Meeting*, Manchester, England.

Dia, H. (2002). An agent-based approach to modelling driver route choice behaviour under the influence of real-time information, *Transportation Research Part C: Emerging Technologies* **10**(5-6): 331–349.
URL: *http://dx.doi.org/10.1016/S0968-090X(02)00025-6*

Ding, F., Zhang, J. & Wang, J. (2010). Accelerometer based transportation mode recognition on mobile phones, *2010 Asia-Pacific Conference on Wearable Computing Systems*, Ieee, pp. 47–50.
URL: *http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5481106*

Doirado, E., van den Berg, M., van Lint, H., Hoogendoorn, S. & Prendinger, H. (2012). Everscape: the making of a disaster evacuation experience, pp. 2285–2285–2290–2290.
URL: *http://dl.acm.org/citation.cfm?id=2223656.2223790*

Ebendt, R., Sohr, A., Tcheumadjeu, L. C. T. & Wagner, P. (2010). Utilizing historical and current travel times based on floating car data for management of an express truck fleet, *5th International Scientific Conference: Theoretical and Practical Issues in Transport*.

Ehsani, R., Buchanon, S. & Salyani, M. (2009). GPS accuracy for tree scouting and other horticultural uses, *Technical report*, University of Florida.

Fiorenzo Catalano, S. (2007). *Choice set generation in multi-modal transportation networks*, PhD thesis, Delft University of Technology, TRAIL Research School.

Flamm, M., Jemelin, C. & Kaufmann, V. (2007). Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions, *7th Swiss Transport Research Conference*, Ascona, Switzerland.

Flötteröd, G. & Bierlaire, M. (2013). Metropolis-hastings sampling of paths, *Transportation Research Part B: Methodological* **48**: 53–66.

Frejinger, E. & Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models, *Transportation Research Part B: Methodological* **41**(3): 363–378.

Frejinger, E., Bierlaire, M. & Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling, *Transportation Research Part B: Methodological* **43**(10): 984–994.

Friedrich, M., Hofsaess, I. & Wekeck, S. (2001). Timetable-based transit assignment using branch and bound techniques, *Transportation Research Record: Journal of the Transportation Research Board* pp. 100–107.

González, M. C., Hidalgo, C. a. & Barabási, A.-L. (2008). Understanding individual human mobility patterns., *Nature* **453**(7196): 779–82.

Greenfeld, J. S. (2002). Matching GPS observations to locations on a digital map, *Proceedings of the 81th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA.

Hess, S., Rose, J. M. & Hensher, D. A. (2008). Asymmetric preference formation in willingness to pay estimates in discrete choice models, *Transportation Research Part E: Logistics and Transportation Review* **44**(5): 847–863.
URL: *http://www.sciencedirect.com/science/article/pii/S1366554507000786*

Hoogendoorn-Lanser, S. (2005). *Modelling Travel Behavior in Multi-modal Networks*, PhD thesis, Technology University of Delft.

Hoogendoorn-Lanser, S. & Bovy, P. (2007). Modeling overlap in multimodal route choice by including trip part-specific path size factors, *Transportation Research Record: Journal of the Transportation Research Board* pp. 74–83.

Hoogendoorn-Lanser, S., van Nes, R. & Bovy, P. (2005). Path size modeling in multi-modal route choice analysis, *Transportation Research Record: Journal of the Transportation Research Board* **1921**(-1): 27–34.

Hoogendoorn-Lanser, S., van Nes, R. & Hoogendoorn, S. (2006). Modeling transfers in multimodal trips: Explaining correlations, *Transportation Research Record: Journal of the Transportation Research Board* **1985**: 144–153.

Hunter, T., Abbeel, P. & Bayen, A. (2012). The path inference filter: model-based low-latency map matching of probe vehicle data, *10th International Workshop on the Algorithmic Foundations of Robotics*.

Inrix (2012). http://www.inrix.com/.

Jan, O., Horowitz, A. J. & Peng, Z.-R. (2000). Using gps data to understand variations in path choice, *Transportation Research Record* (1725): 37–44.

Jenelius, E., Rahmani, M. & Koutsopoulos, H. N. (2012). Travel time estimation for urban road networks using low frequency probe vehicle data, *Transportation Research Board 2012 Annual Meeting*, Washington DC, USA.

Knoblauch, R., Pietrucha, M. & Nitzburg, M. (1996). Field studies of pedestrian walking speed and start-up time, *Transportation Research Record: Journal of the Transportation Research Board* (1538): 27–38.

Komárek, A. (2009). A new R package for bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data, *Computational Statistics & Data Analysis* **53**(12): 3932–3947.

Kwapisz, J., Weiss, G. & Moore, S. (2010). Activity recognition using cell phone accelerometers, *Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data*, pp. 10–18.

Lemp, J. D., Ridge, M. & Kockelman, K. M. (2011). Strategic sampling for large choice sets in estimation and application, *Transportation Research Part A* .

Li, H., Guensler, R. & Ogle, J. (2005). Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data, *Transportation Research Record: Journal of the Transportation Research Board* **1926**: 162–170.

Liao, L., Patterson, D. J., Fox, D. & Kautz, H. (2007). Learning and inferring transportation routines, *Artificial intelligence* **171**(5-6): 311–331.

Mahmassani, H. (2001). Dynamic network traffic assignment and simulation methodology for advanced system management applications, *Networks and Spatial Economics* **1**(3): 267–292.
URL: *http://dx.doi.org/10.1023/A:1012831808926*

Mahmassani, H., Joseph, T. & Jou, R.-C. (1993). Survey approach for study of urban commuter choice dynamics, *Transportation research record* (1412): 80–89.

Marchal, F., Hackney, J. & Axhausen, K. W. (2005). Efficient map matching of large global positioning system data sets: Tests on speed-monitoring experiment in Zurich, *Transportation Research Record: Journal of the Transportation Research Board* **1935**: 93–100.

McFadden, D. (1978). Modeling the choice of residential location, *in* A. Karlqvist (ed.), *Spatial Interaction Theory and Residential Location*, North-Holland, Amsterdam, pp. 75–96.

Murakami, E. & Wagner, D. P. (1999). Can using Global Positioning System (GPS) improve trip reporting?, *Transportation Research Part C: Emerging Technologies* **7**(2-3): 149–165.

Naya, F., Noma, H. & Kogure, K. (2005). Bluetooth-based indoor proximity sensing for nursing context awareness, *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*, Ieee, pp. 212–213.

Nham, B., Siangliulue, K. & Yeung, S. (2008). Predicting mode of transport from iPhone accelerometer data, *Technical report*, Stanford University.

North, A. & Hargreaves, D. (2008). *The social and applied psychology of music*, Oxford University Press.

Ochieng, W. Y., Quddus, M. & Noland, R. B. (2003). Map-matching in complex urban road networks, *Brazilian Journal of Cartography (Revista Brasileira de Cartografia)* **55**(2): 1–18.

OFROU (2011). Route et trafic - chiffres et faits 2010, *Technical report*, Office fédéral des routes (OFROU).

Oshyani, M. F., Sundberg, M. & Karlström, A. (2012). Estimating flexible route choice models using sparse data, *15th International IEEE Conference on Intelligent Transportation Systems*, pp. 1215–1220.
URL: *http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6338676*

OSM (2012). http://wiki.openstreetmap.org/wiki/Switzerland/Map_Features.

Park, B.-J., Zhang, Y. & Lord, D. (2010). Bayesian mixture modeling approach to account for heterogeneity in speed data, *Transportation Research Part B: Methodological* **44**(5): 662–673.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions, *Journal of Choice Modelling* **2**(1): 65–100.

Prato, C. G. & Bekhor, S. (2006). Applying branch-and-bound technique to route choice set generation, *Transportation Research Record: Journal of the Transportation Research Board* **1985**: 19–28.

Prato, C. G., Bekhor, S. & Pronello, C. (2011). Latent variables and route choice behavior, *Transportation* **39**(2): 299–319.
URL: *http://www.springerlink.com/index/10.1007/s11116-011-9344-y*

Pyo, J., Shin, D. & Tae-Kyung, S. (2001). Development of a map matching method using the multiple hypothesis technique, *IEEE Proceedings on Intelligent Transportation Systems* pp. 23–27.

Quddus, M. a., Noland, R. B. & Ochieng, W. Y. (2005). Validation of map matching algorithms using high precision positioning with GPS, *Journal of Navigation* **58**(2): 257–271.
URL: *http://www.journals.cambridge.org/abstract_S0373463305003231*

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

Ramming, M. S. (2002). *Network knowledge and route choice.*, PhD thesis, Massachusetts Institute of Technology, Cambridge, USA.

Ravi, N., Dandekar, N., Mysore, P. & Littman, M. L. (2005). Activity recognition from accelerometer data, *the Seventeenth Conference on Innovative Applications of Artificial Intelligence(IAAI)*, pp. 1541–1546.

Reddy, S., Burke, J., Estrin, D., Hansen, M. & Srivastava, M. (2009). Determining transportation mode on mobile phones, *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on Wearable Computers*, IEEE, pp. 25–28.

Schuessler, N. & Axhausen, K. (2009a). Processing raw data from global positioning systems without additional information, *Transportation Research Record: Journal of the Transportation Research Board* **2105**: 28–36.

Schuessler, N. & Axhausen, K. W. (2009b). Map-matching of GPS traces on high-resolution navigation networks using the multiple hypothesis technique, *Working paper* .

Schüssler, N. (2010). *Accounting for similarities between alternatives in discrete choice models based on high-resolution observations of transport behaviour*, Phd thesis, ETHZ.

Stenneth, L., Wolfson, O., Yu, P. S. & Xu, B. (2011). Transportation mode detection using mobile phones and GIS information, *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, ACM, New York, NY, USA, pp. 54–63.

Stopher, P., Clifford, E., Zhang, J. & Fitzgerald, C. (2008). Deducing mode and purpose from GPS data.

Swisstopo (2011). *Cartes nationales de la Suisse: Symboles de nos cartes Symboles de nos cartes*, Office fédéral de topographie swisstopo.

Thompson, D. C., Rebolledo, V., Thompson, R. S., Kaufman, a. & Rivara, F. P. (1997). Bike speed measurements in a recreational population: validity of self reported speed., *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention* **3**(1): 43–5.
URL: *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1067763&tool=pmcentrez&render*

van Diggelen, F. (1998). GPS accuracy: lies, damn lies, and statistics, *GPS World* **9**(1): 41–45.
URL: *http://www.gpsworld.com/gps/gps-accuracy-lies-damn-lies-and-statistics-1127*

Vovsha, P. (1997). The cross-nested logit model: application to mode choice in the {T}el {A}viv metropolitan area, *Transportation Research Record: Journal of the Transportation Research Board* **1607**: 13–20.

Vrtic, M., Schüssler, N., Erath, A., Axhausen, K., Frejinger, E., Bierlaire, M., Stojanovic, S., Rudel, R. & Maggi, R. (2006). Including travelling costs in the modelling of mobility behaviour., *Technical report*, Final report for SVI research program Mobility Pricing: Project B1, on behalf of the Swiss Federal Department of the Environment, Transport, Energy and Com- munications, IVT ETH Zurich, ROSO EPF Lausanne and USI Lugano.

Wing, M. G., Eklund, A. & Kellogg, L. D. (2005). Consumer-grade global positioning system (gps) accuracy and reliability, *Journal of Forestry* **103**(4): 169–173.
URL: *http://www.ingentaconnect.com/content/saf/jof/2005/00000103/00000004/art00004*

Wolf, J., Oliveira, M. & Thompson, M. (2003). Impact of underreporting on mileage and travel time estimates - results from Global Positioning System-enhanced household travel survey, *Transportation Research Record: Journal of the Transportation Research Board* **1854**: 189–198.

Yang, Q. I. & Koutsopoulos, H. N. (1996). A microscopic traffic simulator for evaluation of dynamic traffic management systems, *Transportation Research Part C: Emerging Technologies* **4**(3): 113–129.
URL: *http://www.sciencedirect.com/science/article/B6VGJ-3VWT8KB-1/2/4bfda91270dcf22f26439123b886be90*

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G. & Huang, Y. (2010). T-Drive: Driving directions based on taxi trajectories, *ACM SIGSPATIAL GIS 2010*.

Zhang, L., Dalyot, S., Eggert, D. & Sester, M. (2008). Multi-stage approach to travel-mode segmentation and classification of GPS traces, *ISPRS Workshop on Geospatial Data Infrastructure:*, pp. 87–93.

Zheng, Y., Li, Q., Chen, Y., Xie, X. & Ma, W.-y. (2008). Understanding mobility based on gps data, *Tenth International Conference on Ubiquitous Computing*, number 49, pp. 312–321.

Zheng, Y., Liu, L., Wang, L. & Xie, X. (2008). Learning transportation mode from raw gps data for geographic applications on the web, *Proceeding of the 17th international conference on World Wide Web*, Beijing, China.

Jingmin Chen

Avenue de Tivoli 8
CH-1007 Lausanne, Switzerland
Phone: +41 21 693 2532
Fax: +41 21 693 8060
Web: http://transp-or.epfl.ch/personnal.php?Person=CHEN

Birth date: April 12, 1985
Nationality: Chinese
jingmin.chen@epfl.ch
chen.jingmin.cn@gmail.com

Research and Teaching Assistant 2008-2012

Transport and Mobility Laboratory, Swiss Institute of Technology, Lausanne

Research Projects:

- **Route choice models and smart phone data. Oct. 2010 – present**
  Project Manager. Sponsor: Swiss National Science Foundation. This project aims at learning mobility information from smartphone data, and estimating route choice behavior from the data.
- **Disaggregate behavioral models exploiting data from Nokia devices. Dec. 2008 – Jun. 2010**
  Sponsor: Nokia Research Center. This project aims at investigating how to perform an appropriate fusion of data from smartphones in order to predict the mobility behavior of a given individual.

Teaching Assistant for Courses:

- Decision-aid methodologies in transportation (2011 and 2012)
- Mathematical Modeling of Behavior (2009)
- Recherche opérationnelle (2009)

Supervised Master Project:

- Mobility in the Middle East. Laurene Aigrain, Swiss Institute of Technology, Lausanne and Imperial College, London, UK, June 2011

Supervised Semester Projects:

- Mobility identification from smartphone GPS data, Denis Garcia (SMA), January 11, 2012
- Visualization of Cell Phone Data on Google Earth, Raoul Neu (SIN), January 15, 2010
- Testing the algorithm for generating path observation from GPS data, Jensen Anders Fjendbo (SGC), January 15, 2010

Education

2012 **Doctor of Philosophy in Mathematics**
Swiss Federal Institute of Technology, Lausanne.
Thesis: Modeling route choice behaviors using rich smartphone data
Thesis supervisor: Prof. Michel Bierlaire

2008 **Master of Science in Planning and Management of Traffic and Transportation**
Southeast University, China
Thesis: Considering travel time reliability in route choice behavior under uncertain demand
Thesis supervisors: Prof. Wei Wang and Prof. Lin Cheng

2005 **Becholar of Engineering in Transportation**
Sun Yat-sen University, China
First two years in Mathematics, then transferred to Transportation Engineering major.

Skills

Computer:	C++, Python, QT, Javascript, SQL, MatLab, R, PHP, Java, Latex
	Linux, GIS (PostGreSQL + PostGIS + qGIS), OpenStreetMap, GAE
Language:	Mandarin (mother tongue), English (Full proficiency), Cantonese (limited proficiency)

Papers in International Journals

- Chen, J., and Bierlaire, M. (to appear). Probabilistic multimodal map-matching with rich smartphone data,Journal of Intelligent Transportation Systems (accepted for publication on October 22, 2012)
- Bierlaire, M., Chen, J., and Newman, J. P (2013). A probabilistic map matching method for smartphone GPS data, Transportation Research Part C: Emerging Technologies 26:78–98.

Papers in Conference Proceedings

- Chen, J., Bierlaire, M., and Flötteröd, G. (2011). Probabilistic multi-modal map matching with rich smartphone data. Proceedings of the Swiss Transportation Research Conference (STRC) May 11-13, 2011 , 2011.
- Bierlaire, M., Chen, J., and Newman, J. P (2010). Using location observations to observe routing for choice models. Proceedings of the 89th Transportation Research Board Annual Meeting (TRB) January 10-14, 2010.
- Newman, J. P, Chen, J., and Bierlaire, M. (2009). Generating probabilistic path observation from GPS data for route choice modeling. Proceedings of the European Transport Conference (ETC) 5-7 October 2009, 2009.
- Chen, J., Newman, J. P, and Bierlaire, M. (2009). Modeling route choice behavior from smart-phone GPS data. Proceedings of the The 12th International Conference on Travel Behaviour Research (IATBR) December 13-18, 2009, 2009.
- Bierlaire, M., Newman, J. P, and Chen, J. (2009). A method of probabilistic map distribution of path likelihood. Proceedings of the 9th Swiss Transport Research Conference (STRC) September 9 - 11, 2009.
- Chen, J. and Cheng, L. (2007). Considering travel time reliability in route choice behavior under uncertain demand. Proceedings of the 3th International Symposium on Transportation Network Reliability (INSTR). The Netherlands,

Speaker in Seminars and Conferences

2011
- 11th Swiss Transportation Research Conference, Ascona, Switzerland.

2010
- 10th Swiss Transportation Research Conference, Ascona, Switzerland.
- integrated Transportation and Energy Activity-based Model Workshop (MIT-Portunal), Lisbon, Portugal.
- World Conference on Transport Research 2010, Lisbon, Portugal.
- NRC-Lausanne bi-annual seminars, Lausanne, Switzerland.

2009
- 12th International Conference on Travel Behavior Research, Jaipur, India.
- European Transportation Conference 2009, Leeuwenhorst, The Netherlands.
- 6th Workshop on Discrete Choice Models, Lausanne, Switzerland.
- 9th Swiss Transportation Research Conference, Ascona, Switzerland.
- Nokia Research Center Data Collection Workshop, Lausanne, Switzerland.

2007
- 3rd International Symposium on Transportation Network Reliability, The Hague, The Netherlands.