Deriving trip's modes and trip's purposes from GPS-based travel surveys

Minh Hieu Nguyen

UNIVERSITÉ PARIS – EST

École doctorale Organisations, Marchés, Institutions (OMI)

A thesis for the degree of Doctor of Sciences is entitled

Deriving Trip's Modes and Trip's Purposes from GPS-Based Travel Surveys

and submitted by **MINH HIEU NGUYEN**

Being supervised by

Jimmy ARMOOGUM, Chargé de Recherche**,** HDR at Université Paris-Est; IFSTTAR/AME/DEST

Sy Sua TU, Professor at University of Transport and Communications (Vietnam)

The thesis is defended on 28th January, 2020

Jury composition:

Kay AXHAUSEN,	Professor at ETH Zurich	*Reviewer*
Ariane DUPONT-KIEFFER,	Dean of the Sorbonne School of Economics, HDR, Université Paris 1 Panthéon-Sorbonne	*Reviewer*
Aruna SIVAKUMAR,	Senior Lecturer at Imperial College London	
Latifa OUKHELLOU,	Directrice de Recherche at Université Paris-Est / IFSTTAR	
Jean-Loup MADRE,	Directeur de Recherche at Université Paris-Est / IFSTTAR	
Jimmy ARMOOGUM,	Chargé de Recherche, HDR at Université Paris-Est / IFSTTAR	

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DECLARATION

I, Minh Hieu Nguyen, confirm that the work presented in this dissertation has been mainly conducted and implemented by myself.

I confirm information from other sources is cited adequately.

I am aware of that this work may be available on online free-access libraries.

Champs-sur-Marne, 1st November 2019

Minh Hieu Nguyen

ACKNOWLEDGEMENT

To complete the final version of the dissertation you are reading, I have received a great number of assistances. Therefore, I would like to devote one of the first parts to acknowledgement.

First and foremost, I would like to give such profuse thanks to HDR. Jimmy Armoogum, my advisor for his invariable assistance, guidance, instructions and encouragement during the three-year period of my PhD work. I have learnt not only academic knowledge but also professional skills in research and collaboration with colleagues, which I had hardly found out at university and school before. Words cannot reveal how thankful and grateful I am.

I would like to be indebted Prof. Sy Sua Tu, the co-advisor of my thesis together with Dr. Jean-Loup Madre and Dr. Christophe Rizet who are members of the committee responsible for monitoring and boosting the advancement of my thesis. I have been deeply impressed by very detailed remarks, which Jean-Loup made to my drafts. I feel lucky and happy because of having opportunities for working with him.

I would like to thank Mr. Cedric Garcia so much for his valuable advice and active support whilst I were struggling with the language barrier and statistics-related issues. I wish to send my sincere appreciation to the direction board of Economic and Social Dynamics of Transport Laboratory (DEST), including Dr. Francis Papon, Dr. Laurent Hivert and Ms. Christine Rouillon for accepting and helping me to complete academic registrations and engage in enlightening transport conferences.

In addition, I would like to express my sincere gratitude to doctoral students, that is, Mr. Fabio Rendina, Mr. Clement Dusong, Ms. Mai Hue Nguyen together with Ms. Yasmine Haddad. Interesting experience in conferences and the doctoral days will be unforgettable events in my life. I wish to thank colleagues and senior researchers in DEST (i.e. Dr. Leslie Belton-Chevallier, Dr. Akli Berri, Dr. Laurent Carnis, Dr. Jean-Michel Fourniau, Dr. Jean-Paul Hubert, Dr. Benjamin Motte-Baumvol and Dr. Amakoe Adolehoume), staff at IFSTTAR and Organization, Mobility, Institution (OMI) doctoral school for their kind support and/or handy discussions.

During my PhD student's period, I have confronted many difficulties regarding language, administrative procedures and the pressure of research. Many Vietnamese friends and colleagues including Mr. Ba Thanh Vu, Mr. Tung Lam Nguyen, Mr. Le Hung Tran, Mr. Quang Huy Dang, Mr. Vinh Hoang Tan Le, and Mr. Van Thanh Ho encouraged and helped me to overcome these challenges. Thus, I wish to say "thank you a lot" to them.

Besides, I would like to send special thanks to Ms. Hoai Thu Tu Thi who is both my talented colleague and my close friend.

One of two main datasets used in this thesis was collected in a smartphone-based survey in Hanoi (Vietnam) thanks to the kind support and volunteer participation of my colleagues at University of Transport and Communications and their students, relatives along with friends. Besides, Trivector Traffic AB, a Swedish company not only offered my laboratory (DEST) discount for renting the app (TravelVU) to collect GPS data in Hanoi but also supported technically me in a prompted way. Therefore, I would like to give my sincere appreciation to them.

I would like to appreciate the monthly financial aids, which the Vietnamese Ministry of Education and Training has given to me during my PhD work's time.

Last, my greatest appreciation is for my wife Thuy Linh Nguyen and my son Nam Anh Nguyen together with my paternal and maternal families. Their patience and sacrifice have been the major motivation for me to complete this thesis. ***This work is dedicated for them***!

PUBLICATIONS OF THESIS

Conference presentations/posters:

Nguyen, M.H., Armoogum, J., (2020). Feature Selection for Enhancing Purpose Imputation from GPS Data without GIS Data. Presented at the TRB 99th Annual Meeting, Washington D.C.

Nguyen, M.H., Armoogum, J., Garcia, C., (2019a). Experiment on mobility survey using smartphone in Hanoi, Vietnam. Presented at the Transportation for A Better Life: Smart Mobility for Now and Then, Bangkok, Thailand.

Nguyen, M.H., Armoogum, J., Garcia, C., (2019b). Mode-Based Comparison of Data in Mobility Surveys using GPS and Telephone. Presented at the 98th TRB Annual Meeting, Washington, D.C.

Accepted paper:

Nguyen, M.H., Armoogum, J., Madre, J-L., Garcia, C., (2020). Reviewing Trip Purpose Imputation in GPS-based Travel Surveys. *Journal of Traffic and Transportation Engineering (English Edition).*

Submissions under review:

Nguyen, M.H., Armoogum, J., (____). Hierarchical Process of Travel Mode Imputation from GPS Data in a Motorcycle-Dependent Area. Submitted to *Travel Behaviour and Society*.

Nguyen, M.H., Armoogum, J., (____). Feature Selection for Enhancing Purpose Imputation from GPS Data without GIS Data. Submitted to *Transportation Research Record: Journal of the Transportation Research Board*.

ABSTRACT

Mobility data play a crucial role in travel behavior research and demand forecast. The complete reliance on conventional datum collection techniques, that is, face-to-face interview, computer-assisted telephone/web/personal interview, postal survey, and email has a number of big drawbacks, including (1) the high burden on respondents, thus high non-response rate, (2) inclusion of one-day data per person, (3) lack of reliability due to human memory limits and habit of rounding travel time, (4) high cost with intensive labor and (5) big time gaps between periodic household travel surveys not to mention the difficulties in combining and harmonizing data of surveys in different regions or countries. The unlimited use of Global Positioning System (GPS) has opened up great opportunities for dealing with the problem of poor data. GPS logs are objective, numerous, continuous, detailed and accurate spatiotemporally. Yet, positioning information itself is not eligible for analysis due to the lack of trip characteristics. This deficiency has induced the substantial development of two new research fields that are involved in imputing transportation modes and trip purposes from GPS data, respectively. On-board devices were initially utilized to take advantage of electricity. Afterward, lightweight, small and wearable personal devices have been developed to collect person-based data, which emphasized the need for detecting trip modes. Currently, smartphones are the most preferred devices to gather both logs and their corresponding so-called ground truth.

Detections of mode and purpose are essential steps prior to doing any travel behavior analyses (e.g. mode choice or time spending in activities). In this sense, the performances of mode and purpose inference algorithms determine the potential of employing GPS-based surveys as a supplement and even an entire alternative to conventional techniques. In the literature, there are three research gaps related to imputation algorithms and GPS-assisted surveys. The first is the great focus of investigations in well-structured urban areas of developed countries and occasionally in China. Therefore, the use of GPS in mobility surveys in cities of developing countries has been questionable. The others are consequences of the first limitation. Second, the list of mode detection encompasses the modes of walk, bike, transit, and car but not motorcycle that is one of the main means in emerging countries. Last, purpose imputation has been implemented very frequently with the support of GIS data; however, GIS data are not available and good enough everywhere. Lack of reasonable solutions to derive purposes from GPS data without GIS data is a gap.

This thesis aims at seeking answers to three mentioned-above existing questions by building both mode and purpose inference models. Two data sets were used. The first was collected in Rhone-Alpes, France by dedicated device whilst the second was gathered in Hanoi, Vietnam by smartphone. Based on prediction results, discussions and recommendations for enhancing the quality of GPS-based surveys in general and for developing countries in particular have been proposed.

Keywords: GPS, mobility survey, purpose imputation, trip detection, mode detection, random forest, GIS, fuzzy logic, smartphone

RÉSUMÉ

Les données sur la mobilité jouent un rôle crucial dans la recherche sur le comportement des voyageurs et la prévision de la demande. Le recours exclusif aux techniques de collecte de données conventionnelles, à savoir entretien individuel, entretien téléphonique / Web / personnel assisté par ordinateur, sondage postal et courrier électronique présente de nombreux inconvénients, notamment une lourde charge pour les répondants, d'où un taux de réponse faible, description de la mobilité sur une seule journée par personne, manque de fiabilité en raison des limites de la mémoire humaine, coût élevé vue la nécessité d'un travail intensif, écart important entre les enquêtes périodiques sur la mobilité des ménages, sans oublier les difficultés à combiner et à harmoniser les données des enquêtes dans différents pays ou régions. L'utilisation maintenant sans contrainte du système de positionnement global (GPS) a ouvert de grandes possibilités pour améliorer la qualité de ces données. Les journaux relevés GPS sont objectifs, nombreux abondants, continus, détaillés et précis d'un point de vue spatio-temporel. Cependant, les informations de positionnement en elles-mêmes ne sont pas éligibles suffisantes pour l'analyse de la mobilité en raison de l'absence de caractéristiques des déplacements. Cette lacune a entraîné un développement substantiel de deux nouveaux domaines de recherche: l'imputation des modes de transport déplacement et du motif du voyage à partir des données GPS, respectivement. À l'heure actuelle, les téléphones intelligents sont les appareils privilégiés pour collecter les traces GPS et la description des déplacements correspondants.

La détection du mode et du motif est une étape essentielle avant de procéder à une analyse du comportement de déplacement (choix du mode ou temps passé dans les activités, par exemple). En ce sens, les performances des algorithmes d'inférence du mode et du motif déterminent le potentiel de l'utilisation d'enquêtes basées sur GPS en tant que complément, voire même alternative complète aux techniques conventionnelles. Dans la littérature, il existe trois lacunes dans la recherche concernant les algorithmes d'imputation et pour les enquêtes assistées par GPS. La première provient du fait que, jusqu'à présent, les enquêtes GPS ont été menées essentiellement dans les zones urbaines des pays développés et parfois aussi en Chine. Par conséquent, l'utilisation du GPS dans les enquêtes sur la mobilité dans les villes des pays en développement reste à développer. La seconde et la troisième lacunes sont les conséquences de cette première limitation. La liste des modes de détection comprend la marche, le vélo, le transport en commun et la voiture, mais pas la moto, l'un des principaux moyens utilisés dans les pays émergents. Enfin, les imputations ont été mises en œuvre très fréquemment avec le support des Systèmes d'Information Géographique (SIG); Cependant, les données SIG ne sont pas partout disponibles et d'une qualité suffisante. L'absence de solutions raisonnables pour imputer les caractéristiques des déplacements à partir des données GPS sans l'apport d'un SIG constitue une lacune de la recherche.

Cette thèse a pour objectif de rechercher des réponses aux trois questions susmentionnées en construisant à la fois des modèles d'inférence de mode et de motif. Deux ensembles de données ont été utilisés. Le premier a été collecté en Rhône-Alpes, en France, par un appareil dédié, tandis que le second a été rassemblé à Hanoï, au Vietnam, par smartphone. Sur la base des performances des classificateurs des recommandations visant à améliorer la qualité des enquêtes par GPS en général et pour les pays en développement en particulier ont été proposées.

Mots-clés: GPS, enquête sur la mobilité, imputation des modes et motifs, détection de voyage, segmentation des déplacements, détection de mode, forêt aléatoire, SIG, logique floue, smartphone

TERMINOLOGIES AND DEFINITIONS

- *Fix, record, log, GPS point*

Refer to an object's position that is at a specific time on the ground. It is captured from satellites' signals/messages and processed by a receiver mounted on the object.

- *Coordinate*

Specifies longitude, latitude and altitude of a GPS point. In this study, it refers to latitude and longitude only.

- *Feature, variable, dimension, indicator*

Feature is a common and important concept in the machine learning field (Forman, 2003). A feature is an individual measurable property or characteristic of a phenomenon being observed (Bishop, 2006). Feature is an equivalent concept to independent/exploratory/input variable in the statistics and modeling fields because values of both variables and features vary across samples. As for a specific sample, their values estimated from data enable to address either regression or classification problems.

- *Supervised machine learning*

Machine learning is the scientific study that gives computers the ability to learn without being explicitly programmed, according to Arthur Samuel in 1959. A supervised machine learning model is a mathematical model learnt from training data with known labels in order to predict unseen data.

- *Purpose, activity*

Refer to the main business undertaken in a trip's destination that is a meaningful location such as university, home, workplace, malls and restaurants.

- *Trip*

Refers to a one-way course of travel to do an activity. A trip is made by only one or multi modes.

- *Segment, stage, trip leg*

Refer to a continuous movement by only one mode. A segment is a part of a trip.

- *Trip end*

Are involved in place and time at which an activity takes place. Hence, a trip end is between two consecutive trips.

- *Transition point, transfer point*

Are involved in place and time at which there is a shift from one mode to another mode.

- *Velocity, speed*

The measurement of the change in the position of an object during a time unit.

- *Heading, bearing*

Refer to the angle that is measured in degrees in clockwise direction from true north. Heading between two GPS points can be estimated from their coordinates

- *Participant, respondent, user*

Refer to persons who participate in surveys to provide data for research. If data are collected by smartphone, participants are called as users.

- *Class, label*

Refers to modes in mode detection and purposes/activities in purpose imputation.

- *Imbalanced data*

Include the classes that are not represented (nearly) equally. For example, the percentages of home and work activities are much higher than those of other purposes.

- *Ground truth*

Is an cartographic terminology that refers to validation by measurement on the ground for data collected by satellite (Shen and Stopher, 2014a). Simply, ground truth is information reported by participants and corresponding to the GPS data collected.

- *Tour*

Is defined as any sequence of trips and activities a person makes from leaving from a particular place to returning it. In travel behavior analysis, tour is frequently in relation to home or work, thus indicated as home-based tour and work-based tour, respectively.

- *Trajectory*

Is a series of GPS points considered in a particular period (e.g. a day) and arranged chronologically. A trajectory contains points of both trips and activities.

ACRONYMS AND ABBREVIATIONS

AFC	Automated Fare Collection
CATI	Computer-Assisted Telephone Interviews
DT	Decision Tree
EDR-RA	L'Enquête Déplacements Régionale Rhône-Alpes
GIS	Geographical Information System
GPS	Global Positioning System
HDOP	Horizontal Dilution Of Precision
HG	Human Geography
HP	Hyper-Parameter
MF	Membership Function
NSAT	Number of SATtellite in view
POI	Point Of Interest
OFB	Out-Of-Bag
RF	Random Forest
TOMOS	Test de l'Observation de la Mobilité par Suivi GPS
TS	Transportation Science
TTFF	Time To First Fix

Chapter 1: INTRODUCTION

1.1. MOTIVATION

The quality of data is the critical factor in analyzing travel behavior and forecasting travel demand (Chen et al., 2016). With samples' data being comprised of lack of adequate information of trips in terms of space and time, estimating the population's future travel patterns would be highly likely to be unreliable, leading to failures and mistakes in developing infrastructure and providing public transport services. Therefore, improving the quality of data has been necessary but demanding too. GPS is a promising tool to collect continuously and passively big spatiotemporal data, which enables to address the long-lasting poor data problem of conventional techniques like face-to-face interview or computer-assisted telephone interview (Auld et al., 2009; Wolf et al., 2014b). However, there are two main challenges to make the use of GPS data beyond experimental investigations. The first is technical issues pertaining to the signal loss and low-quality logs as a result of being affected by both man-made and natural obstacles such as tunnels, bus shelters, skyscrapers, or trees (Biljecki et al., 2013; Gong et al., 2012). This type of problem commonly occurs in almost all GPS-based surveys but possibly addressed well by solutions proposed (Biljecki et al., 2013; Chen et al., 2010; Chung and Shalaby, 2005; Schuessler and Axhausen, 2009; P. Stopher et al., 2008a). The second challenge is involved in algorithms to impute travel modes and trip purposes. On contrary to the high transferability of methods to deal with technical matters, a construction of an inference model depends largely on research areas' specific traveling, working and living conditions. The variety of these conditions have contributed to a great number of scientific efforts, which have been invested in detecting mode and purpose from GPS data (Gong et al., 2014; Prelipcean et al., 2017; Shen and Stopher, 2014b; Yang et al., 2018).

The main venues of mode detection are urban areas of developed countries and well-structured cities of China, which leads the modes considered to be walk, bike, car, bus/tram, and train but not motorcycle that is a primary mode in many developing countries (Stead and Pojani, 2017). Similarly, a vast majority of existing purpose imputation studies have been undertaken in cities where GIS data are sufficient for predicting activities at high accuracy while there is no publication carried out in regions not well-geocoded of emerging countries.

Accordingly, the implementation of GPS-based surveys and forming inference models in the developing countries are obvious shortcomings of the dissemination of mobility survey using GPS. This is such a strong motivation for conducting surveys in both developed and under-developed countries to make comparison and propose solutions to deploy GPS-assisted surveys.

This thesis is the fourth one conducted at laboratory DEST under IFSTTAR (France) and aiming at translating GPS streams collected in mobility surveys. The previous works of (Nguyen, 2013; Pham, 2016; Yuan, 2010) coped with the challenge of detecting trips from trajectories in order to analyze the magnitude of the missing trip problems between GPS-based and CATI surveys. This dissertation is their continuity by means of inferring trip characteristics, that is, modes and purposes.

1.2. RESEARCH OBJECTIVES

- Without direct ground truth, developing a mode detection algorithm to detect transport means from data collected by dedicated devices in Rhone-Alpes, France, a developed country witnessing a number of GPS-based surveys.

- With ground truth gathered from users, developing a mode detection algorithm to detect transport means from data collected by smartphones in Hanoi, Vietnam, a developing country with the dominant role of motorcycles.

- With ground truth gathered from users, developing a purpose detection algorithm to detect activities from data collected by smartphones in Hanoi, where GIS data of points of interest and land use are insufficient or inaccessible.

- Evaluating the performances of detecting mode and purpose in Hanoi to clarify the challenges to applying GPS-based surveys for cities of developing countries, particularly in comparison with the case of Rhone-Alpes of France.

- Proposing solutions to enhance GPS-based surveys in developing countries and developed countries.

1.3. THESIS STRUCTURE

The reminder of this work is structured as follows.

Chapter 2 is devoted to a review of the related literature, that is, (1) mobility data collection methods, (2) the use of GPS in mobility surveys, (3) processing GPS data to be aware of trips, travel modes and purposes.

Chapter 3 presents the two datasets used in this thesis. They are GPS data collected by personal device in Rhone-Alpes (France) and those gathered by smartphone in Hanoi (Vietnam).

Chapter 4 describes the development of a fuzzy logic algorithm to detect travel modes from GPS data in Rhone-Alpes without ground truth. The detection results were validated by the data of the regional household travel survey conducted at the corresponding temporal and spatial scope.

Chapter 5 documents a hierarchical process formed by the combination of fuzzy logic theory, rules and a Random Forest (RF) algorithm to predict five modes (i.e. walk, bike, bus, motorcycle, and car) in case of Hanoi.

The main content of Chapter 6 is the creation of a RF model to identify trip purposes without the use of GIS data.

Chapter 7 based on the outcomes of Chapters 3, 4, 5 and 6 highlights challenges to and recommendations for enhancing the quality of GPS-assisted surveys and developing models to impute trip characteristics (i.e. travel modes and trip purposes) from GPS data in both developed and developing countries.

The last chapter makes conclusion and potential future research directions.

1.4. CONTRIBUTIONS

In this PhD work, contributions to the literature of GPS-based surveys have been made and are presented in order of occurrence as follows:

- The development of fuzzy logic to detect transportation mode from GPS data with the use of heading change rate and the validation by data of the corresponding household travel survey.

- The development of a hierarchical process that takes advantages of deterministic, probabilistic and machine learning methods to detect travel modes in case of imbalanced data being comprised of obviously disproportional ratios of modes.

- The development of a rule-based model to detect effectively bus segments by considering both stopping at and passing slowly bus stops together with the threshold of distance between two consecutive stops in the whole transit network.

- The development of a RF model to detect trip purpose in case of the absence of GIS data.

- Discussions about challenges and recommendations for improving the implementation of GPS-enabled travel surveys.

Chapter 2: LITERATURE REVIEW

2.1. MOBILITY DATA COLLECTION

2.1.1. Traditional and conventional data collection methods

To picture travel patterns in regions or nations, a wide range of datum collection methods have been introduced, applied and categorized into three groups.

The first encompasses *traditional methods* based on both direct and indirect communications between surveyors and participants without the support of modern technologies such as computers, Global Positioning Systems (GPS). The most common technique in the past was face-to-face interviews where participants who accepted invitations are asked about their past travel diaries. Thanks to the proliferation of telephone in households, telephone survey has become greatly popular. Telephone here mainly refers to landline telephone but also including cellular phones. Compared with telephone, the availability of cell phone numbers is low in the public directories. Whilst the two mentioned-above types undertake direct communications with participants, postal surveys wait passively for respondents' volunteer feedbacks in letters, leading to such a very low-response rate.

The second is *conventional methods* that are improved versions of traditional methods by the ubiquitous prevalence of the internet and computer. Emails take place of letters to reduce efforts participants have to make to provide their answers. Based on the idea of giving more freedom to participants in terms of answering questionnaires to gather more responses, Computer-Assisted Web Interview (CAWI) has been introduced. A participant is sent a password logs into his/her account on the internet for providing his/her travel information. Questionnaires in web surveys are designed more sophisticated, colorful and attractive with pictures and guidelines on modernized and user-friendly interfaces. CAWI has been the main collection technique in the 2017 U.S. National Travel Household Survey. With the integration of computer rather than using paper and pen, face-to-face interviews can be carried out more effectively; it is frequently known as Computer-Assisted Personal Interviews (CAPI). In this technique, a portable electronic device is used to store answers, thus it is usually preferred to others in case of wishing to collect both verbal and non-verbal feedbacks of participants with long questionnaires. Computer-Assisted Telephone Interviews (CATI), one of the most widely applied methods, is the telephone-based investigation with the help of software. An interviewer is in charge of both reading questionnaires and recording a respondent's answers directly in a computer system equipped with software dedicated to the survey. CATI is superior compared with others, because it allows capturing trips of multiple members of a household, leading to save efforts to recruit and, more importantly, information of different members can be used to confirm their travel diaries and identify errors.

Each of methods in either traditional and conventional method groups has its own advantages and disadvantages that are synthesized rigorously in (Armoogum et al., 2014). Therefore, some of them are regularly combined to obtain data as effectively and efficiently as possible. For example, in the Netherlands, Germany, Norway, Switzerland and the U.S., interviews by telephone and web are deployed in concert (Bassett et al., 2008). A potential participant is first requested to engage in a web-based survey by letter. In case there is no reply during a particular period, a phone interview is undertaken. CAWI plays the main role and supplemented by CATI in the last U.S. national travel survey to increase the coverage that reaches households without landlines.

Traditional and conventional methods are based on self-reported travel diaries, which leads data to be prone to serious problems, including:

***- High and increasing non-response rate*:**

Non-response is involved in the failure to measure all the units of sample or all variables of interest (Armoogum et al., 2014). Simply, it is a consequence of the omission of answers to either the whole or a part of questions in a survey. The loss of information affects negatively the accuracy of estimators and causes a bias to the survey. A good CATI survey in North America can gain the 60% recruitment rate and 60% completion rate of successfully recruited households, resulting in the overall response rate of roughly 36% (Zimowski et al., 1997). In a synthesis of Ortúzar et al., (2011), the highest response rate is at 83% in the 1998 Denmark survey whilst the lowest is between 5% and 10% in German Mobility Panel. Between 2001/2002 and 2008/2009, the response rates in both the U.S. national household travel survey and the German national survey, frequently known as MiD (Mobilität in Deutschland) dropped from approximately 40% to about 20% (Buehler et al., 2011). The main reason for the non-response issue is the heavy burden on participants. Burden is involved in making efforts to give answers to questions. The biggest contributor to burden may be the cognitive attempts to remember the past travel diaries as detailed as possible. A person traveling more tends to be subject to bigger pressure because of remembering more information. Burden would be proportional to the length of questionnaires and the survey period (Golob and Meurs, 1986). Besides, tasks to access the survey, provide and modify responses contribute to participants' unpleasant senses.

***- Declining sample sizes*:**

Sample size plays an important role in creating models to analyze travel behaviors and project travel demand. However, a decline in size has witnessed. For example, initially, face-to-face home-based interviews could reach 1-3% of the population but dropping at less than 1% of the total households in the same area (Stopher and Greaves, 2007).

***- Lack of reliability*:**

Surveys that hinge upon the human memory are highly likely to be subject to unreliable travel information. Significant numbers of trips in CATI surveys are omitted (Forrest and Pearson, 2005; Wolf et al., 2003b). Missed trips are either unimportant or short (Forrest and Pearson, 2005; Richardson et al., 1996; Wolf et al., 2003b). On the other hand, the common habit of numerous respondents is to round starting and/or end times of each trip (Kelly et al., 2013), hindering surveyors from collect precise data. In a self-reported survey, a research area (i.e. a country or a region) is geographically divided into areas and units at the smallest level (Armoogum et al., 2018b). Trip length would be an approximation based on the shortest path between the unit of the origin and the unit of the destination rather than the actual routes between the origin and the destination.

***- Out-dated data*:**

The interval between surveys causes adverse impacts on data quality. National Travel Household Surveys are frequently carried out every 10 years (Armoogum et al., 2014). In France, the five most recent ones were in 1973-1974, 1981-1982, 1993-1994, 2007-2008, and 2018-2019. The frequency of regional surveys is even sparser. Data collection was conducted at an interval of almost 15 years in the Puget Sound region, US (Murakami and Watterson, 1990). During long time gaps, the changes in demographic characteristics and transportation policies may cause significant changes in travel behaviors. Serious problems and concerns like the proliferation of mobility together with motorized vehicles, changes in the share of transport modes due to energy crisis have not waited 10 or 15 years to be forecast and responded. Thus, data should be continuously gathered to capture changes in mobility patterns (Ortúzar et al., 2011). Another shortcoming of self-administered survey is the short coverage of only one reference day (e.g. in Denmark, Sweden, France), several days per person (e.g. in Italy, New Zealand) or 7 days (the Netherlands,

Great Britain, Germany). It is in nature because it is too difficult and burdensome for a respondent to recall a great amount of information that may be very old and/or during a very long period. Golob and Meurs, (1986) emphasized the considerable "trip reporting fatigue" respondents ended up in the Dutch National Mobility one-week survey.

***- Non-representative sample*:**

This is the direct result of high non-response rate and mainly related to the poor recruitment due to geographical and technological barriers along with privacy concern. When it comes to the face-to-face interview, reaching participants' home is challenge for those living in remote areas. For example, it is too dangerous to arrive in some areas of North America (Stopher and Greaves, 2007). In case technological devices are employed, the poor availability or non-ownership of telephone or/and computer or/and internet connection make the surveys fail to cover some of population. In fact, the elderly tend to be unfamiliar with modern devices and thus afraid of communicating via them. In many countries, the incomplete databases of telephone numbers prevent the survey from being more representative (Bayart and Bonnel, 2012). To describe groups whose under-representation and absence in surveys are frequently reported, the concept "hard-to-reach group" was introduced (Behrens et al., 2009). Generally, there are two main groups it is difficult for surveyors to contact. The first is those who are regularly omitted from sampling frames like the homeless, inhabitants with informal housing and without fixed utility servicing networks, residents recently changing home but not updating their new addresses. The second is those who are persistent non-respondents due to regulations or personal limitations related to language and illiteracy. Some are unwilling to engage in surveys because of religious barriers or restrictions of their workplaces. Non-response and representativeness in conventional surveys have been long-lasting issues and attracted great attentions in recent in-depth discussions about survey methods (Armoogum et al., 2018a; Armoogum and Dill, 2015).

***- Intensive labor and costly*:**

Traditional and conventional travel surveys require great participation of staff and surveyors during the processes of preparation, communication with respondents, saving and interpreting data in computers. The cost per household is in a range between 150$ and 225$ in US (Wolf et al., 2014b) or about 350$ in Sydney Travel Household Survey (Stopher and Greaves, 2007). One of big questions for every conventional survey is the trade-off between expectation of datum quality and cost. With the desire to have better data of travel behavior, the increased cost is unavoidable for nationwide and region-wide surveys.

These issues are interacted with others rather than exist independently. For example, due to the high cost, the frequency of national periodic surveys is sparse at around 10 years with one-day data per person. Or the increasing non-response rate and the falling sample sizes are responsible for a low representability of the sample because less information, particularly of several groups (i.e. hard-to-reach pools) would be successfully obtained. Mobility data need to be precise because they are essential for travel demand forecast, issuing and/or adapting urban policies towards sustainable development. Hence, with drawbacks indicated above, conventional techniques may not be sufficient for obtaining knowledge on the population's travel.

2.1.2. Evolution and characteristics of GPS-based travel surveys

Global Positioning System (GPS) is a space navigation system being developed by the US military and officially operational in 1995. It is able to track well movement of objects equipped with a receiver in terms of both time and space by employing the 32-satellite system. In the fancy, the use of GPS was mainly for military objectives whilst its uses for civilian purposes were under the effect of Selective Availability intentionally degrading GPS points' accuracy. Despite of this prevention, pilot studies such as the 1996

FHWA Lexington Pilot Study, the 1997 Austin Household Travel Survey and the 2000 Georgia Tech survey demonstrated the obvious promises of GPS in terms of collecting and detecting precisely details of travel patterns. Wolf et al., (2001) provide a typical evidence. Their purpose imputation analysis on the GPS data of 13 participants in Atlanta reached the 79% accuracy. Data vary within 30 to 100m of the true positions due to the impact of Selective Availability.

The removal of Selective Availability on May 1, 2000 has opened up great opportunity for conducting GPS-based applications. In US, GPS data are used for determining trip rate correction factors, detecting activity locations, identifying and analyzing route choice and mode choice, evaluating active transport behavior and identifying trip purpose (Wolf et al., 2014a). Currently, GPS is an indispensable component of daily life and a strong base for studies of mining human daily activity, mode choice, route detection, route planning and Point Of Interest (POI) recommendation, estimating fuel consumption, monitoring health (Byon et al., 2009; Chaix et al., 2013; Ermagun et al., 2017; Furletti et al., 2013; Hashemi and Karimi, 2016; Hu et al., 2018; Quddus and Washington, 2015; Schüssler, 2010; Thomas et al., 2018; van Hassel et al., 2017). The following paragraphs discuss characteristics of GPS-based surveys as a solution to supplement/replacement to traditional/conventional survey methods.

**** Advantages of GPS data***

The advantages of GPS data over self-reported data are obvious. Positional data enable to draw very high-resolution pictures of behaviors in that they are collected at high frequency (e.g. every second) with the high spatial accuracy ranging between 3 and 10m (Feng and Timmermans, 2015; Nour et al., 2016). Because of passively being gathered and independent on human beings' memory, spatiotemporal data include adequately and reliably participants' behaviors. Equally important is the alleviation of burden on participants. In case prompted recall surveys are conducted to collect actual travel diaries of logs, termed "ground truth", tasks for respondents are less complex and less time-consuming since GPS data play as recommendations and clues to remind them of their trip and their activities (Cottrill et al., 2013; Xiao et al., 2016). In many cases, data collection is close to be completely passive and silent without asking participants to confirm trip characteristics of logs (Patterson and Fitzsimmons, 2016; Schuessler and Axhausen, 2009). The result of relieving strain on respondents is an extension of survey period being from several days to several weeks, even multiple years. Various information (i.e. longitude, latitude, altitude, time, the number of satellite, instantaneous speed), which a GPS device employs to estimate its position, horizontal dilution of precision, heading, is useful for making travel-related inference models (Feng and Timmermans, 2019, 2016). Depending on the sampling interval and the length, a trip includes from a number of to numerous points; therefore, GPS data is a type of big spatial data that enable and encourage the application and development of machine learning algorithms on mobility data (Feng and Timmermans, 2016). GPS sensor is an important component of a smartphone thus the ubiquitous prevalence of mobile phone has fostered GPS data collection naturally. In this sense, more or less difficulties in recruitment may be lessened, because persons have potential devices for data collection. Decrease in sample size would be a possible merit of GPS data. Considering sampling error, the sample of multi-day GPS-assisted survey is smaller at 20-30% of than that of one-day survey (P. R. Stopher et al., 2008). However, there is no more scientific support to this view.

**** Disadvantages of GPS data***

As a coin has two sides, GPS technology has its own drawbacks. First and foremost is the shortage of trip attributes. Surveyors can be aware of the position of a receiver at a specific time but not how and why the respondent travels. This deficiency is the strong motivation for developing algorithms to infer transportation modes and trip purposes (Gong et al., 2014; Prelipcean et al., 2017; Shen and Stopher,

2014b). Similarly, logs do not have information of participants (e.g. gender, age and educational degree), passengers' assessment (e.g. service quality, attitudes to public transport utilization, preference) and vehicle occupancy, leading to the implementation of both pre-surveys and post-surveys. The former is used to collect personal information whilst the latter, frequently indicated as prompted recall surveys, aims at collecting the actual travel diaries to evaluate the results of inference models. Pre- and post-surveys principally are conventional investigations that increase respondent's burden. The more detailed prompted recall survey is, the heavier strain is. The pertinent limitations of conventional surveys in terms of quality of self-reported data would affect negatively the evaluation of and the participation in GPS-based surveys more or less.

It is undeniable that GPS can provide very accurate data; however, in many cases technological problems occur. A poor-quality point comes from the small Number of SATellites in view (NSAT) and/or high value of Horizontal Dilution Of Precision (HDOP). More discussions about NSAT and HDOP will be presented later. The loss and the degradation of GPS signal during warm and cold start, in urban canyons or in/under other obstructions like tunnels, roof of stations and vehicles occur on a regular basis (Schuessler and Axhausen, 2009; P. Stopher et al., 2008a). The warm and cold start problems are related to the fact that a GPS receiver whilst being turned on need to search for signals to compute its current position, called acquisition process. The time from turning it on to successfully estimating the valid position is called the acquisition time or Time To First Fix (TTFF). For warm start, the receiver first assumes that the current position is the last recorded. Then, the datum starts being recorded within 15-40s (P. Stopher et al., 2008a). If the duration between stopping and re-powering the receiver is large (e.g. over an hour), the receiver has to acquire good signals of (e.g. at least 4) satellites during a particular period to receive accurately description of satellites' orbit (i.e. ephemeris data) before its position is acquired. TTFF is 30-60s in case the receiver has been stationary but several minutes (e.g. 2-15 minutes) if it has been moved/moving (Chung and Shalaby, 2005; P. Stopher et al., 2008a). Cold start usually occurs at the beginning of a day whilst warm start coincides with switching from sleep mode (i.e. device stops for an hour or several ones) to operational mode. During TTFF, there is no GPS datum recorded, leading to losing either completely traces of short trips or partly traces of longer trips. Consequently, duration and trip length or vehicle kilometers traveled, the consumption of energy are possibly under-measured. Additionally, it causes difficulties in comparing trajectories with actual travel diaries. Urban canyon problems refer to typical issues taking place in urban canyons. One of them is multipath error that results from the refection of GPS signal mainly by high-rise buildings and walls, making respective GPS points jump around the actual position. Similar effects arise when an object is travelling under tree canopies. According to the report of Gong et al., (2012) conducted in the Wall Street (US), the average deviation between streets where participants walked and the GPS points was about 52m compared with 13m in the whole area of New York. In fact, issues interact each other rather than function separately, making themselves more serious. Also in the work of Gong et al., (2012), TTFF of warm start in areas under stronger effects of urban canyon is longer up to a few minutes. In many cases, a receiver fails to receive signals. Common situations of signal loss are movement in tunnels, buildings or under over-bridges. The use of transportation modes pertains to signal blocking. The GPS reception for traveling by public transport is worse than that for traveling by other modes. The vast majority of metro trips cannot be recorded by GPS receivers whilst parts of a trip by bus or tram may be omitted due to the obstructions of its roof (Biljecki et al., 2013). Better signals are always obtained in case the receiver is closer to windows of public transport means (Draijer et al., 2000). Notably, roofs of cars generally do not affect much the quality of GPS data (Chung and Shalaby, 2005). Every now and then, an object's behaviors cannot be captured in that the receiver's battery runs out or the receiver is not attached to the object.

Due to the technical problems and their adverse influences on obtaining insight into travel behavior from logs, some conclude that GPS surveys is far from even never altering completely conventional travel datum collection techniques (Marchal et al., 2011; Vij and Shankari, 2015).

**** Evolution of GPS-assisted surveys based on devices used***

The progression of GPS surveys in general and the analyses of GPS data particularly have depended greatly upon the types of receiver used to collect data. There have been three main devices, that is, (1) in-vehicle device, (2) wearable device and (3) smartphone. We divide the evolution of GPS-based surveys according to the dominant use of each among these. The temporal border of each period is not clear with significant overlaps because devices have been developed and tested simultaneously.

- *The period of in-vehicle device*

In very first studies, GPS data were collected by devices attached to cars (Bricka and Bhat, 2006; Forrest and Pearson, 2005; Murakami and Wagner, 1999; Wolf et al., 2004, 2003b, 2001). The greatest advantage of on-board device is to take power source of vehicle instead of having its own energy source. Additionally, its size and weight may not be concerned about not to mention its simple installation in cars. It can be set up to gather operational information of the vehicle like speed, acceleration and engine status, which is useful for identifying trip ends. For example, very low speed represents the stops whilst turning off engine is often involved in the leave from vehicle.

In-vehicle devices has made a breakthrough in auditing, estimating and correcting missed data in conventional travel surveys, frequently indicated as the under-reporting problem. Under-reporting of trip rate comes from survey length, participants' memory decay, lack of understanding of or failure to adherence to survey guidelines, carelessness and unwillingness to report full details (Wolf et al., 2003b, 2003a). The relationship of (1) household characteristics, (2) personal characteristics, (3) travel behavior characteristics with levels of and/or the likelihood of under-reporting are analyzed and reported in (Bricka and Bhat, 2006; Forrest and Pearson, 2005). Specifically, single-member households achieve the obvious higher reporting accuracy than that of two-member ones, that is, 63% versus 31%. Short trips in a trip chain or a very short round trip/tour are often omitted. The data of vehicles enable to develop simple algorithms to detect trip purposes by rules or probability (Wolf et al., 2004, 2001).

This period witnessed the great and partly exaggerated expectations about the use of GPS, reflected by the strong belief in the potential replacement of GPS-based data to self-reported data (Auld et al., 2009). Compared with great concentration on under-reporting issue, fewer trips recorded than those entered in travel diaries are analyzed and explained by technical problems such as the poor signal reception, the disconnection between devices and vehicle power (Schönfelder and Samaga, 2003; Wolf et al., 2003a, 2003b). This is very important to make researchers think about more investment of scientific efforts in dealing with GPS's limitations to extract better knowledge about travel pattern from positioning data.

- *The period of wearable device*

Car-mounted devices were first used because of its convenient use and its ability to track almost all trips by car that is the dominant mode in survey areas. Unfortunately, they are insufficient for drawing the comprehensive picture of travel behavior that encompasses other travel modes like walk, bike and transit, inducing the preference of wearable GPS dedicated device to on-board counterparts. Handheld loggers are capable of collecting person-based and multimodal GPS data. The mark of this period is by 1997 when the first survey using wearable GPS devices was carried out in the Netherlands (Draijer et al., 2000). From then on, great number of researches employing person-based logs in Australia (P. Stopher et al., 2008a), the Netherlands (Bohte and Maat, 2009; Feng and Timmermans, 2019), Denmark (Rasmussen et al., 2015), Japan (Shafique and Hato, 2015), Canada (Tsui and Shalaby, 2006), France (Marchal et al.,

2011), Switzerland (Schuessler and Axhausen, 2009), the US (Chen et al., 2010) and Belgium (Reumers et al., 2013) have been undertaken. To lessen burden on participants, GPS dedicated devices are made more and more portable and useful by the reduction in size and weight along with the addition of sensors. The very first devices mounted on bike were 2kg in weight (Draijer et al., 2000) but now around several hundred grams and possibly collect not only GPS data but also 3-dimensional acceleration information (Feng and Timmermans, 2013; Shafique and Hato, 2015).

Whilst explaining in-vehicle GPS data would be simple to some extent with the main tasks being to detect trip ends to estimate trip characteristics including length, duration and purpose, translating person-based data into sequence of trips and activities is more complex due to the inclusion of many travel modes. Splitting trajectories into segments corresponding to modes and detecting modes have become the major research emphases (Shen and Stopher, 2014b). Personal data are still feasible for assessing the accuracy of self-reported methods in household travel surveys (Stopher et al., 2007) and creating purpose inference models but mainly used for mode identification.

- *The period of smartphone*

Advantages of dedicated GPS logger are undeniable; however, its use is prone to both technical and economic problems. Money for purchasing GPS loggers makes up a significant rate in the total cost. As a result, the number of devices bought is smaller than the sample size. The survey is divided into phases. In each phase, a device is distributed to a participant before recollected to give another. This way has two main limitations. First, the survey cannot cover the entire sample simultaneously. Analyzing data of the whole sample may consider the change in behavior due to external factors like the distinct weather conditions. More importantly, the return of GPS loggers determines both the sample size of the following phases and the cost of projects. Whilst the rates of successfully recollecting devices in Australia and New Zealand are high at approximately 100%, the figure for others is apparently low 60-70% (Stopher et al., 2018). Almost all GPS device generations function by collecting and then storing all data on its memory before all data is retrieved at the end of the survey, possibly resulting in the entire loss of data in case loggers are not returned not to mention the increasing requirement for memory capacity. Newest devices thanks to the combination between GPS sensor and GSM chip are able to transmit data in real time, thus their memory is not large because data can be deleted after sent to the servers. Of course, the cost of using such devices is significantly higher. A GPS device has single function of recording positioning data solely as passively as possible; therefore, it is likely to be forgotten by participants. The inclusion of space for battery prevents devices from becoming more lightweight and small.

The rapid growth of smartphone-based GPS data collection is obvious by now. The ubiquitous pervasion of various brands and types of smartphone with in-built GPS sensor has brought about a number of advantages of smartphone-assisted data collection over handheld counterparts. First, the large number of, even the whole participants can be surveyed in parallel without provision of any devices at all. Second, smartphone is a multi-function device extremely useful for working and daily life, thus it is less likely to be left at home than a wearable device. Third, the collection by smartphone is more cost-effective because each user has a smartphone. The largest amount of cost would be for developing and maintaining an app but not purchasing devices. Fourth, smartphone enables users to review and check their trajectories visualized on their phone displays before giving ground truth information by validating and/or correcting recommendations of the app. By this way, participant burden is alleviated and the survey duration is extended. A smartphone app may be capable of implementing data transmission in real time, thus detect whether the app functions or not, to make request a user to turn on the app. Smartphones now can be packed

with a variety of sensor types (e.g. 3-dimensional accelerometer, gyroscope, WIFI, magnetometer) that play as useful sources to supplement GPS data.

Notwithstanding, smartphone-based collection may be very expensive in case participants do not either have or want to use their own smartphone, not to mention the compatibility between smartphone and the survey app. In the Singapore Household Travel Survey, some participants were borrowed phones whilst smartphones were provided for all participants in three Japanese cities in (Gong et al., 2018). As a part of the national household travel survey in New Zealand, an app, called ATLAS II was developed to be usable for the iOS platform only. Battery consumption of the app is a big problem. The more a person travels the more battery his/her smartphone consumes.

GPS data collected by smartphone generally is bigger in terms of time and sample than those of wearable loggers (Cottrill et al., 2013; Semanjski et al., 2017; Thomas et al., 2018; Zheng et al., 2010), enabling to apply and develop machine learning models to detect two main trip attributes that is modes and purposes. Detailed syntheses of studies of mode and activity inference are presented in the following parts.

2.2. GPS DATA PROCESSING

Processing GPS data here is the (offline) post-process following the completion of data collection in order to addressing the deficiency of trip characteristics in fixes.

Ideally, outcome of this process is a series of attributes of each trip, including travel mode, purpose, origin, destination, departing time, arrival time and route. Among them, origin, destination, starting and ending times can be estimated from data if a trip is detected; however, gaining knowledge of mode, purpose and route is complex. Mode and purpose are not apparently included whilst determining route requires the detailed geo-coded map of road network. Thus identifying mode, purpose and route is indicated as separate steps/fields pertaining to developing complex and quite independent algorithms. In this sense, GPS data processing is comprised of five steps. The first is data preparation related to download and assemble data recorded before removing bad points. The second is trip/segment identification. It aims to split each trajectory into trips to discriminate trips from activities. Also in this step, single-mode stages of each trip are found. The third is to determine mode of stages identified in the previous step. Based on stage mode information, the (main) trip mode is determined. The fourth is to detect trip purpose whilst the fifth targets at inferring routes to complete travel behavior profiles.

The last procedure that is frequently involved in map-matching field (Hashemi and Karimi, 2014; Quddus et al., 2007) is out of scope. Following sections focus on reviewing the first four steps.

2.3. DATA FILTERING

The purpose of this step is to check for the validity of points and make necessary corrections to provide adequate input for further analyses of mode and purpose inference.

As indicated above, NSAT and HDOP are the two major criteria to determine the quality of a fix. The threshold of NSAT depends on the number of targeted dimensions. To get a position specified by longitude and latitude only, a GPS antenna must receive signals from at least three satellites at the same time. If altitude is considered, four satellites in view are the minimum for an adequate point. The relationship between a point quality and NSAT is proportional. HDOP describes the arrangement of satellites in the sky when a fix is recorded. Lower HDOP value means that satellites more widely spread and thus fix is more accurate. The common threshold of HDOP value is 5 (P. Stopher et al., 2008a; Tsui and Shalaby, 2006).

Unrealistic values of indicators are used effectively to remove aberrant points; however, thresholds for the same indicator may vary between studies, possibly resulting from difference in geographic position of research areas, transport networks along with road regulations, the accuracy of data and researchers' experience. For example, bad points are those with speed over 250km/h in (P. Stopher et al., 2008a) and over 150km/h in (Wang et al., 2017). Generally, instantaneous speed is used the most. Information of NSAT, HDOP, altitude, changes in longitude, heading and latitude are not collected for many cases.

In order to provide better data for further analysis, some smoothing techniques (Kalman filter, Gauss kernel) are applied (Nitsche et al., 2014; Rasmussen et al., 2015; Schuessler and Axhausen, 2009).

2.4. SEGMENTATION

The objective of this step is to break trajectories into consecutive trips and activities. Algorithms to identify trip/segment depend heavily on the devices used to collect data and the further analysis goals.

To limit the confusion related to terminologies, we define them as follows. *An activity* is the main business undertaken in *a location*; hence, it is in harmony with *a purpose*. *A trip* is a one-way course of travel to do an activity. A trip is either single-mode or multimodal. *A segment* is a continuous movement by only one mode. Two consecutive trips are divided by a purpose (or an activity) whilst two segments in a trip are partitioned by a mode shift at *a transition point*. *A stop point* witnesses a short interruption of movement, to wait for the green light at intersection for example. A location witnesses both the end and the start of two consecutive trips, respectively. *A trajectory* is a series of GPS points of both activities and trips considered in a particular period and arranged in a chronological order.

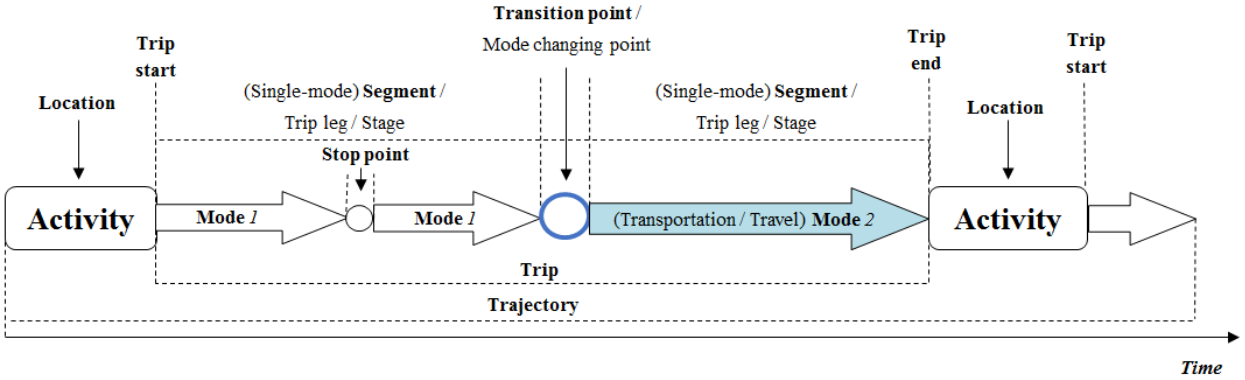


Figure 2 - 1. Graphically defining trip elements

Table 2 - 1. Summary of rules and techniques to filter and smooth data

Author	NSAT	HDOP	Speed	Heading	Altitude	Acceleration	Distance	Latitude	Longitude	Filter technique
Tsui & Shalaby, (2006)	< 3	> 5	0*	0	-	-	-	-	-	-
Stopher et al., (2008)	< 4	≥ 5	0* or > 250 km/h	0 or unchanged	-	-	-	< 15 m change	< 15 m change	-
Bohte & Maat, (2009)	-	-	> 200 km/h	-	-	-	< 10 m(1)	-	-	-
Schuessler & Axhausen, (2009)	-	-	> 50 m/s	-	< 200 m or > 4200 m	-	-	-	-	Gauss kernel smoothing
Nitsche et al. (2014)	-	-	-	-	-	-	-	-	-	Kalman filter
Rasmussen et al. (2015)	-	-	-	-	< -37 m or >201 m	-	-	-	-	Gauss kernel smoothing
Safi et al (2016)	-	-	not recorded or > 42 m/s	-	-	> 10 m/s^2	-	-	-	-
Wang et al (2017)	< 4	-	> 150 km/h	-	> 200 m	-	> 200 m(2)	-	-	-

Note:

Records meeting at least one rule will be eliminated, except for *

*: Zero speed is used with: (zero directional) heading (Tsui and Shalaby, 2006), and with heading, latitude, and longitude (Stopher et al., 2008);

(1) distance to the previous point;

(2) distances from the point to two centre points of 5 previous points and 5 following points, respectively.

2.4.1. Identifying trip and segment in studies of mode detection

Data for mode detection are person-based ones collected by dedicated devices or smartphones. To detect transportation mode, trips are found from trajectories before segments are identified from each trip. Detecting trips and segments is on the basis of the similar principles related to the loss of signal and the stability of device during a sufficiently long period.

Stationary status is defined by nearly zero instantaneous speed between consecutive points. The speed thresholds are exactly 0 m/s (Shen and Stopher, 2013a; P. Stopher et al., 2008a; Tsui and Shalaby, 2006) or less than 0.01 m/s (Rasmussen et al., 2015; Schuessler and Axhausen, 2009) or 2 km/h (Biljecki et al., 2013) or not documented (B. Wang et al., 2017). If there is a stable status whose duration is over a time boundary, a trip end is detected. The most common cut-off value of dwell time is 120s (Schuessler and Axhausen, 2009; P. Stopher et al., 2008a; Tsui and Shalaby, 2006). The higher values of 180s and 210s are reported in (Patterson and Fitzsimmons, 2016) and (Safi et al., 2016), respectively. Whereas, some studies use the considerably lower level of 60s (Rasmussen et al., 2015).

The time of signal loss is another criterion to determine the termination of a trip. Similar to the threshold of stable duration, the thresholds of missing data vary among studies. The most heavily used threshold is 120s (Gong et al., 2012; Rasmussen et al., 2015; Tsui and Shalaby, 2006) whilst the highest is 900s (Schuessler and Axhausen, 2009).

To detect segments, all points of each trip are considered. The thresholds of dwell time and the GPS interruption duration are lower than the corresponding levels to detect trip. For example, the dwell time of 12s and the signal loss duration of 30s were introduced in (Biljecki et al., 2013).

A notable advantage of trip/segment identification by time thresholds is simple and easily understandable; however, the result of segmentation is sensitive to the thresholds adopted, likely leading to both under-segmentation and over-segmentation issues. Over-segmentation is involved in dividing one real segment into several segments, whilst under-segmentation omits to detect transition points, inducing to merge segments with different modes into one segment. Over-segmentation is less serious than under-segmentation because if two consecutive segments have identical mode prediction results, they can be merged. Yet, under-segmentation causes the loss of information and there is no way to detect and fix it. A lower time threshold causes a higher magnitude of over-segmentation and a lower level of time under-segmentation. However, it is worth noting that both over- and under-segmentations occur in parallel at almost all thresholds.

An activity would be defined by the movement of a GPS device within a small area or a cluster. Therefore, segmentation can be implemented by setting a radius based on measurement accuracy (e.g. 30 meters (Stopher et al., 2008), 15 meters (Schuessler and Axhausen, 2009) and 50 meters (H. Gong et al, 2012)) to search for points belonging to the cluster and estimate the centroid during a particular period (e.g. 300 seconds (Schuessler and Axhausen, 2009), 200 seconds (Gong et al., 2012), and 60 seconds (Rasmussen et al., 2015)).

Uniqueness of a trip is frequently beginning and/or finishing by walking that is typical by very low speed and acceleration (e.g. 2.78 m/s and 0.1 m/s^2 for maximum speed and maximum acceleration, respectively (Schuessler and Axhausen, 2009)). In this sense, detecting walking segment is the key to finding mode transfer points and thus the remaining legs in a trip (Gong et al., 2012; Schuessler and Axhausen, 2009; Zheng et al., 2010). A common criterion for a walking segment is to last over 60s. This method has a shortcoming with respect to failing to detect short and immediate walking, for example between car and transit.

Table 2 - 2. Segmentation rules in the mode detection field

Author	Availability of signal				No signal
	Dwell time	*Speed*	*Radius of cluster/searching*	*Walking detection*	*Duration*
Tsui & Shalaby, (2006)	≥ 120 s	0	-	-	≥ 120 s
Stopher et al., (2008)	≥ 120 s	0	30 m(1)	-	-
Schuessler & Axhausen, (2009)M	≥ 120 s	≤ 0.01 m/s	15 m with ≥ 15 points in a sequence ≥ 300 s or ≥ 10 points	Speed, acceleration profiles and segment duration	≥ 900 s
Zheng et al., (2010)	-	-	-	Speed and acceleration with fusion rules	-
H. Gong et al., (2012)	-	< 1.6 km/h	50 m for ≥ 200 s	Speed profiles and segment duration	> 120 s
Biljecki et al., (2013)	≥ 12 s	≤ 2 km/h	-	-	≥ 30 s
Shen & Stopher, (2013b)	≥ 60 s	0	-	-	-
Rasmussen et al., (2015)	≥ 60 s	≤ 0.01 m/s	Within limited area for ≥ 60 s	-	≥ 120 s
Wang et al., (2017)	≥ 120 s	-	10 m	-	≥ 120 s

Note:

A segmentation is deduced if there is (1) no signal during a predefined period (see column "No signal"), or (2) a cluster of points (see column "Radius of cluster/searching"), or (3) very low speed (see column "Speed") between consecutive points during a predefined period (see column "Dwell time"), or (4) detection of walking segment (see column "Walking detection").

(1) gauged by 3 times standard deviation of accuracy measurement

2.4.2. Identifying trip in studies of purpose inference

Dissimilar to mode detection studies, purpose-specific researches use both in-vehicle data (Schönfelder and Samaga, 2003; Wolf et al., 2004, 2001) and person-based data (Feng and Timmermans, 2015; Gong et al., 2018; Xiao et al., 2016). They do not pay attention to identify segments.

Rule-based routine is the key to handle the trip detection challenge. The minimum threshold of duration witnessing the nearly stationary status of the observed object is heavily utilized. The 120-second threshold yields the best trip identification performance (Wolf et al., 2001) and agreed by (Deng and Ji, 2010; P. Stopher et al., 2008a), but smaller than 180 seconds (Bohte and Maat, 2009; Yazdizadeh et al., 2019) and 300 seconds (Wolf et al., 2004). The highest is at 10 minutes (Usyukov, 2017). Gong et al., (2018) also apply the principle of minimum time at a location to recognize trip ends but not report the cut-off value. Obviously, there is no boundary causing a satisfactory identification of all trip ends. The low and high thresholds coincide with over-segmentation and under-segmentation, respectively.

Clustering points is a solution to identify trip ends (Chen et al., 2010; Montini et al., 2014). Specifically, a cluster represents a trip end that is marked as points within 50 m of each other for more than

200 s. The trip end is assigned as the first stopped point of a cluster (Chen et al., 2010; Gong et al., 2012). Montini et al., (2014) decide a trip end position as the coordinate with the highest density in a cluster.

Table 2 - 3. Trip end detection in the purpose detection field

Method	Description
Rule-based	Zero speed over 120s (P. Stopher et al., 2008a; Wolf et al., 2001); At certain place over 180s (Bohte and Maat, 2009; Yazdizadeh et al., 2019), 360s (Wolf et al., 2004); under-5-km/h speed level over 120s (Deng and Ji, 2010); no movement during at least 10 mins (Usyukov, 2017)
Clustering	Points are within 50m of each other over 200 s (Chen et al., 2010; Gong et al., 2012), coordinate with the highest density (Montini et al., 2014)
Signal of vehicle status	Car engine switches from on to off (Lu et al., 2012); Whether passenger on taxi (Gong et al., 2016)
Comparison between ground truth and GPS data	Merge or divide GPS data into trips and activities based on examining travel diaries reported by participants (Cui et al., 2018)
Probability-based	Probability based on the distance to the road with the support of street map (Liao et al., 2007)

In case of using in-vehicle devices, a trip end is detected through signal showing the change in engine status from on to off (Lu et al., 2012). It is not reliable enough because the saving power mode can switch a car off when it is waiting for the red light for a fairly long time. In another situation, the engine-off event may not happen for short activities like dropping out someone. For taxi trips (Gong et al., 2016), because signals include information on whether passengers on board, a trip end is easily detected as a drop-off point.

Cui et al., (2018) examine both GPS data and travel diaries reported to merge or divide trajectories into trips and activities. Although being time-consuming and requiring great efforts, this method can eliminate spurious activities that pertain to signal gaps when a device passes through tunnels for example. Liao et al., (2007) detect trip ends by using street map to develop a complex probabilistic method.

Identifying trip ends does not exist in (McGowen and McNally, 2007) because they geo-coded origins and destinations of trips in California household travel survey. Geo-coded data are as surrogate of GPS data.

2.5. MODE DETECTION

The beginning of the mode detection field is around 2006 thanks to the introduction of wearable devices that collect person-based data. From then on, it has become the heart of studies using GPS data. Mode detection has been considered in three domains including Location-Based Services (LBS), Transportation Science (TS) and Human Geography (HG) (Prelipcean et al., 2017). The vast majority of mode detection studies are possibly arranged into LBS and TS and the differences in studies belonging two domains generally are not clear.

2.5.1. Method group

According to Gong et al. (2014), mode imputation methods can be divided into three main groups: deterministic, probabilistic, and machine-learning methods.

Deterministic methods are based on predefined, ad-hoc rules of speed, acceleration, and distance to locations of bus stops and train stations (Bohte and Maat, 2009; Gong et al., 2012). They are simple and

easy to interpret because they are based on transport practices. For example, a trip would be inferred to have been taken on foot if its nearly maximum (i.e., the 85th percentile of) speed does not exceed 10 km/h and its average speed does not exceed 6 km/h (Gong et al., 2012). Rules are useful for cases wherein the specific characteristics of travel modes are shown; however, these rules are insufficiently flexible to deal with the reality of travel, such as the slow movement of almost all modes on congested roads. The performance of deterministic methods depend largely on experts' experience and knowledge of the travel environment in the research area of interest. Furthermore, the use of a large number of variables is not practical, because this would result in an exponential increase in the number of rules, defined as the combinations of variables.

The second type of mode inference method (probabilistic) involves extensions of rules in fuzzy-logic-based models (Rasmussen et al., 2015; Schuessler and Axhausen, 2009; Tsui and Shalaby, 2006) and a probability matrix (Stopher et al., 2008). Instead of strictly making decisions, probabilistic methods consider the overlap of modes' behaviors to generate probabilities for each of the modes simultaneously. The mode that has the highest probability is attributed to the trip. Probabilistic approaches are flexible classifiers; however, they share the limitations of deterministic methods mentioned above owing to their reliance on devising rules. Most deterministic and probabilistic models use fewer variables than there are transportation modes, with accuracy levels of over 90% being attained in GPS-enabled tests (Rasmussen et al., 2015; Stopher et al., 2008; Tsui and Shalaby, 2006) but a level of only 70% being reached in a regional experiment (Bohte and Maat, 2009). Schuessler and Axhausen (2009) ignored the model accuracy owing to the absence of ground truth.

Machine-learning algorithms (e.g., support-vector machines, random forests, and artificial neural networks) are currently preferred owing to their ability to learn directly from big data to effectively classify modes (Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Gong et al., 2018; Semanjski et al., 2017; Shafique and Hato, 2015; Stenneth et al., 2011; Xiao et al., 2015). The advantage of machine-learning methods over deterministic and probabilistic methods is that they create powerful classifiers using various types of variables, the number of which exceeds the number of classes (Bzdok et al., 2018). For example, Feng and Timmermans (2016) used 17 variables related to movement, participants' information, and the quality of GPS points to detect 10 modes. The contribution and importance of variables in machine-learning models may be confusing to some extent, as their combination and interaction occur in black-box and mathematically complicated processes. Notably, mode use has not been balanced with the majority of trips belonging to some modes (e.g., walk, car) and the small minority of trips employing other modes, such as bus/tram and bicycle (Dabiri and Heaslip, 2018; Nour et al., 2016; Xiao et al., 2015). If big data are not collected, the imbalance in the mode shares would mean that relatively little data of minor modes would be used to train the model; thus, their detection would be significantly poorer than that of major modes. If adequate data are provided, the overall accuracy levels exceed 90% (Semanjski et al., 2017; Xiao et al., 2015) and can reach nearly 100% (Feng and Timmermans, 2016; Shafique and Hato, 2015).

2.5.2. Feature/variable selection

The most important variables are those related to speed and acceleration. By means of deploying instantaneous values of speed and acceleration at each point of a trip, transportation modes can be predicted satisfactorily. According to (Feng and Timmermans, 2013), the addition of acceleration data increases the accuracy from 78.4% to 91.7%. The most heavily used speed/acceleration indicators are nearly maximum, median, average values. Because a GPS point locates around rather than matches exactly the actual position

of a receiver; therefore, the maximum speed/acceleration may be exaggerated due to jumping points. Nearly maximum levels are defined as the 95th percentile or the 85th percentile of speed/acceleration (Feng and Timmermans, 2019; Schuessler and Axhausen, 2009). The 95th percentile of speed is the speed level that is higher than 95% of instantaneous speed values of a trip. Interestingly, the 85th percentile of speed is an important guideline in determining and evaluating the speed limit on roads (Johnson and Kuby, 2008). To gain a high success in detecting transit, GIS data are necessary (Bohte and Maat, 2009; Gong et al., 2012).

Some recent studies have reported interesting mode detection results based on using acceleration data or GIS data solely. Shafique & Hato (2015) use data collected from 46 participants in 3 Japanese cities (Niigata, Gifu and Matsuyama) between 2010 and 2011 by dedicated GPS devices to detect four modes including walk, bike, car, and train. The study takes advantage of the detailed raw acceleration data. Along with resultant and average resultant acceleration, 18 other acceleration-related variables enable to detect correctly over 99% of cases in every city with the application of Random Forest. This supports the great importance of acceleration in mode detection. However, apart from the difference in methods, the approximately 100% accuracy may come from the exclusion of bus easily misclassified as car (Gong et al., 2012; Xiao et al., 2015a). In addition, the modal share is obviously dominated by walk and car (e.g. 62.6% and 30.3%, respectively in the case of Matsuyama) that can be distinguished correctly at a high level by acceleration only (Feng and Timmermans, 2013).

Semanjski et al. (2017) deploy spatial data to discriminate between walk, bike, car, bus, and train. Notably, the high-resolution records are collected at a very large scale with over 30,000 segments of 8303 persons in Leuven (Belgium) in 2015. The authors employ the OpenStreetMap to delineate spatial contexts of a track point through calculating its distances to road types and transport points. The overall accuracy is high at 94% with all bike and train segments detected correctly. Walking that is the most easily identified in previous studies (Gong et al., 2012; Zheng et al., 2010) reaches the lowest level of 75.5%. Additionally, 18% of walking segments are labelled as bike cases. The reason for the confusion is the shared spatial contexts. The infusion of a speed variable to the Support Vector Machine-based classifier makes the overall accuracy rise by 1.2% and the misclassification of the modes of bike, bus, car as walking takes place simultaneously. Hence, new features can facilitate the correct detection but more or less foster the ambiguity between modes. The availability of superior GIS data is the great advantage of the study and also the impediment to transferability of the spatial-context-dependent method to other areas not being mapped digitally well.

The previously mentioned analyses show that the use of only acceleration or GIS data can give promising results but in limited areas. An addition of features would cause both positive and negative impacts on the classifier's performance in case of all modes being processed concurrently, which is prevailing in recent machine learning-based studies (Dabiri and Heaslip, 2018; Feng and Timmermans, 2016; Semanjski et al., 2017; Shafique and Hato, 2015; Xiao et al., 2015a). The hierarchical detection procedure, although requiring more steps, possibly alleviates the counter-effects. For example, walking and biking (non-motorized classes) are discriminated from together and from the rest by movement-related variables before the use of spatial data helps to classify well motorized modes (Biljecki et al., 2013; Nitsche et al., 2014).

Several new features have been introduced. Low-speed point rate, the proportion of points with speed under 1 m/s, is believed to promote bus segment identification through capturing the periodical bus stops (Xiao et al., 2015a). The rate of acceleration change (i.e. jerk) which is an important indicator in the domain of road safety is firstly deployed to carry out mode detection in (Dabiri and Heaslip, 2018). The

importance of the features could not be measured because they are in use with others to feed complex neural network algorithms.

2.5.3. Transportation mode list and validation

Although the lists of transportation modes and the methods used vary across studies, all researchers have paid close attention to basic modes in developed countries and modern cities of China, including walk, bicycle, car, bus/tram, and metro (Bohte and Maat, 2009; Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Gong et al., 2012, 2018; Marra et al., 2019; Semanjski et al., 2017; Shafique and Hato, 2015; Stopher et al., 2008; Xiao et al., 2015).

Mode detection is a classification problem and its results can be evaluated by two ways depending on either the availability of ground truth. In case ground truth information is collected, the prediction results are directly compared with modes reported to estimate precision and recall for each mode and accuracy for the whole model. Precision and recall is frequently combined and presented by a confusion matrix. Almost all mode detection studies use this way. If ground truth data are absent, there are two solutions. The first is using the results of large-scale conventional household travel surveys (Schuessler and Axhausen, 2009). This comparison mainly focuses on demonstrating the better ability of GPS in terms of detecting trips. The second way is manually giving ground truth to GPS data based on visualization and experience (Gong et al., 2018).

2.5.4. Existing questions in mode detection

As indicated above, mode detection studies infer how people travel from data collected in urban areas in developed countries or modern cities in developing countries like Shanghai, Beijing in China. It is interesting to test and develop mode prediction models for data that are collected in urban areas of developing countries and cover motorcycle.

Machine learning methods are trendy; however, rule-based, probability-based and machine learning approaches have their own advantages and disadvantages. The choice of method(s) depends largely on the volume and the type of data collected. Developing methods to identify mode in case of without ground truth has received a little attention. Another gap is how to combine effectively different method groups to predict modes, particularly in case the volumes of different modes are clearly disproportionate.

2.6. PURPOSE IMPUTATION

In order to make this review systematic, a search for papers published by mid-January 2020 on four databases (Web of Science, Scopus, TRID and Science Direct) was conducted by a function using Boolean operations, that is, *("activity inference" OR "activity type" OR "activity types" OR "trip purpose" OR "trip purposes") AND ("GPS" OR "social media").* To the best of our knowledge, *purpose(s)* and *activity/activities* are interchangeable terminologies in the literature because a purpose refers to an activity performed at a trip destination. To limit the number of candidates found but still ensure sufficient pinpointing of publications, *trip, type(s)* and *inference* were added. Because some authors only indicate *social media* but not GPS in titles, abstract and keywords (e.g., Cui et al., 2018), *social media* were coupled with *GPS*. Only journal articles, conference papers and book chapters/lecture notes published officially and in English were examined.

The numbers of papers found was: 1,098 (Scopus), 127 (TRID), 103 (Web of Science) and 38 (ScienceDirect). After removing duplicates, 1,162 were kept. Then, a screening of their titles, abstracts and

keywords was undertaken. For papers including *social media* but not *GPS* in titles, abstracts and keywords, research data descriptions in full texts were scrutinized. Finally, a corpus of 25 publications was selected based on the four criteria that follow: (1) available full text, (2) showing clearly method with variables/features used, (3) detecting at least three purposes from GPS data, (4) achieving ≥ 1 citation, ≥ 3 citations and ≥ 6 citations for those published in 2019[1], 2016–2018 and 2001–2015, respectively. The fourth criterion enables the removal of publications with low impacts on the literature. The number of citations was based on statistics from Google Scholar.

Compared with mode detection, imputing trip purpose although being conducted by the infancy of GPS datum collection has been understudied. The secondary role of purpose detection results from its complex nature. In contrast to visible transportation modes, inferring purposes involves seeking the most probable activity at a specific location type. In fact, different people can engage in dissimilar activities at the same place in cases of multiple functionality (e.g., a shopping mall with restaurants and a cinema inside). Additionally, while movement parameters (e.g., instantaneous speed, acceleration), which are usually provided by devices or easily calculated from coordinates and timestamps, are sufficient for labelling modes (Feng and Timmermans, 2013; Nguyen et al., 2019a; Schuessler and Axhausen, 2009), purpose prediction depends largely on external sources like land use data (Oliveira et al., 2014; Xiao et al., 2016). GPS traces and/or GIS data may not be accurate enough to recognize the places of activities (Meng et al., 2017). Shortage of data also prevents researchers from conducting more studies. A trip comprises some single-mode segments (i.e. stages/trip legs) but has only one purpose.

GPS data are useful for research in two domains, namely Transportation Science (TS) and Human Geography (HG).

Transportation Science (TS) studies persons' or objects' movement from place to place by an array of methods like physics, operation research, probability, control theory (Hall, 2003). Trip purposes coupled with their respective locations are useful since they are predictors of mode choice and travel forecast models (Burbidge and Goulias, 2009; Ho and Mulley, 2013).

Because of long-lasting and unresolved drawbacks of data quality collected in conventional travel surveys (Chen et al., 2016), TS researchers pay assiduous attention to the potential of GPS for supplementing and even replacing the traditional methods of data collection (Auld et al., 2009). That is to say, the focus of TS is on developing and validating methods of deriving trip purpose and other characteristics from GPS data.

The majority of purpose imputation studies are found within TS. In this domain, developing methods are undertaken after travel surveys are completed (i.e. post-collection analysis). Twelve out of 19 studies gather data using personal devices or smartphones ((Bohte and Maat, 2009; Chen et al., 2010; Cui et al., 2018; Deng and Ji, 2010; Feng and Timmermans, 2015; Gong et al., 2018; Montini et al., 2014; Oliveira et al., 2014; Shen and Stopher, 2013; Stopher et al., 2008; Xiao et al., 2016; Yazdizadeh et al., 2019). TS studies is to regularly collect ground truth, which refers to confirmed travel diaries corresponding to GPS streams, through prompted recall surveys following up GPS-enabled ones (Auld et al., 2009). A variety of strategies to collect ground truth have been carried out, including the use of paper diaries (Chen et al., 2010; McGowen and McNally, 2007; Oliveira et al., 2014; Reumers et al., 2013; Wolf et al., 2001), websites (Bohte and Maat, 2009; Cui et al., 2018; Deng and Ji, 2010; Feng and Timmermans, 2015; Krause and Zhang, 2019; Lu et al., 2012; Montini et al., 2014; Shen and Stopher, 2013; Stopher et al., 2008), telephone (Xiao et al., 2016), mail (Cui et al., 2018) and smartphone (Yazdizadeh et al., 2019). Another

[1] Papers are usually published online in *article in press*; therefore, they may have citations before published formally.

way is to manually label activities by means of visualizing trip ends in association with POI (i.e. Point Of Interest) data (Gong et al., 2018; Liao et al., 2007).

Human Geography (HG), a major discipline of the subject field of geography (Gibson, 2009), studies the interrelationships between people, place, and environment and how they vary spatially and temporally (Castree et al., 2013). It places attention on elements of human activity and organization, such as culture, urbanization, population and transport (Castree et al., 2013; Fang et al., 2017).

Similar to TS, GPS-based purpose imputation studies in HG undertake post-collection analyses in order to acquire general knowledge on mobility and whereabouts of activities. They do so by segmenting trajectories into parts of stopping and moving before semantically enriching them through inference algorithms, usually called the semantic enrichment process (Furletti et al., 2013; Gautama et al., 2017; Prelipcean et al., 2017).

Table 2 - 4. Overview of purpose imputation studies in Transportation Science and Human Geography

Indicator	Transportation Science	Human Geography
Focus	Methods of deriving purposes from GPS data	Semantic enrichment for GPS trajectories
Validation	Directly by ground truth	Indirectly by previously known knowledge(1) or by ground truth
Mainly assessed by	Quantitative measurement (e.g., accuracy, precision, recall)	Consistent with previously known knowledge
Main devices used to collect GPS data	Personal devices and smartphones	Devices attached to vehicles
Data source	Travel surveys	Travel surveys and taxi (probe) data
Studies for this review	19: (Bohte and Maat, 2009; Chen et al., 2010; Cui et al., 2018; Deng and Ji, 2010; Feng and Timmermans, 2015; Gong et al., 2018; Krause and Zhang, 2019; Liao et al., 2007; Lu et al., 2012; McGowen and McNally, 2007; Meng et al., 2017; Montini et al., 2014; Oliveira et al., 2014; Shen and Stopher, 2013; Stopher et al., 2008; Wolf et al., 2004, 2001; Xiao et al., 2016; Yazdizadeh et al., 2019)	6: (Chen et al., 2019; Furletti et al., 2013; Gong et al., 2016; Reumers et al., 2013; Usyukov, 2017; Wang et al., 2017)

(1) refers to common sense and information extracted from large-scale travel household surveys

Compared to studies in TS, GPS-based purpose imputation studies in HG are fewer and published within the last decade. Most HG studies use data from specific dweller groups (e.g., cyclists [Usyukov, 2017] or taxi passengers [Chen et al., 2019; Gong et al., 2016; Wang et al., 2017]) gathered by on-board devices (Chen et al., 2019; Furletti et al., 2013; Gong et al., 2016; Wang et al., 2017). Ground truth is optional because the common sense or the travel patterns from previous surveys are sufficient for validation (Chen et al., 2019; Gong et al., 2016; Usyukov, 2017; Wang et al., 2017). In Furletti et al. (2013) and Reumers et al (2013), ground truth is collected, enabling application of learning methods to semantically annotate activities to GPS traces. The poor prediction of non-home and non-work purposes in Reumers et al. (2013) and of "services, shopping" in Furletti et al. (2013) may not be a concern since the main travel patterns are obtained.

Among other technologies, including Mobile Phone Positioning, Smart Card, Social Media Network, WIFI, Radio Frequency Identification Devices and Bluetooth, only the three first are possibly

comparable to or complement GPS in imputing outdoor activities efficiently and effectively (Lee et al., 2016; Motlagh et al., 2009; Yue et al., 2014).

Mobile phone data are collected massively in passive solicitation and cover a considerable proportion of population (Chen et al., 2015). A number of studies (Ahas et al., 2010; Alexander et al., 2015; Chen et al., 2014; Jiang et al., 2017) infer the types of trip destination from mobile data. Compared with GPS-based purpose imputation studies, those on the basis of mobile phone data have much shorter history because this type of data has been emerged recently since 2010s. It is widely accepted that accuracy of GPS data, especially gathered by dedicated devices, is far better than that of mobile phone data whose location error vary between 100-600 m in urban areas and over 4,000 m in rural low-density areas (Abdulazim et al., 2013). Moreover, mobile data are numerous but not continuous. They are only recorded when there is a call or a message. Hence, they would be suitable for identifying basic places, where a phone frequently is in communication with others during specific time windows, like workplace and home. On contrary, GPS indeed continuously tracks the movement and stops. Along with this, GPS data tend to be enriched with personal information of participants who are anonymous in mobile positioning. Consequently, researchers are more likely to choose GPS data to predict trip purpose.

Smart card is a feature of public transport Automated Fare Collection (AFC) and sometimes available for hospitals, restaurants and shops (Yue et al., 2014). No information on trip purpose is recorded by AFC data; however, they are a rich, comprehensive, longitudinal data source useful for transit planners (Pelletier et al., 2011). Similar to mobile phone data, the smart card transaction data-based purpose inference problem has driven attention to only recently (Alsger et al., 2018; Devillaine et al., 2012; Lee and Hickman, 2014). The scope of GPS-based studies is broader than that of smart card ones targeting transit passengers solely. The purpose list of smart card studies is usually limited to mandatory purposes (home-related, work-related) whose trips are regular spatially and temporally (Devillaine et al., 2012; Lee and Hickman, 2014). Strength of smart cards over GPS is to be comprised of pieces of personal information (e.g. age group, student or adult), which aids to inference process. However, as for supplementing or altering conventional household travel survey, inferring purpose from GPS data is a better choice thanks to estimating at both aggregate and disaggregate level (Wolf et al., 2004; Xiao et al., 2016) whereas AFC data would be eligible for aggregate inference (i.e. work-related trips and education-related-trips) (Alsger et al., 2018).

Social media (e.g. Facebook, Twitter, Flickr, and Foursquare) is the environment for numerous people to create a huge amount of passive data. Beyond GPS data when it comes to quantity, temporal and spatial quality of social media data is much lower (Noulas et al., 2012), leading to its supportive role of GPS data rather than playing as the main source to deduce activities (Cui et al., 2018; Yazdizadeh et al., 2019).

Based on previously mentioned discussion in various contexts, it can be seen that the purpose imputation problem has grown along with the advancement of positioning technologies. GPS data that are either enriched by participants' social characteristics either followed by prompted recall surveys are favourite source of studies because they are eligible for inferring a series of purposes thanks to its high accuracy level and continuous observation. Researchers from TS have been the major motivators to its evolution; yet, counterparts from HG have proposed interesting directions to discover general knowledge on mobility of specific groups with limited amount of input. The rapid rise and prevalence of mobile phone data and smart card transaction data do not make inferring activities from GPS data behind, whilst social

media has offered hopes for implementing methodological improvement to predict activities at higher accuracy.

2.6.1. Feature/variable selection

All features are classified into four categories, including (1) geographic data, (2) activity-related and trip-related, (3) participant-related and (4) others.

**** Geographic data***

Geographic data are employed most frequently to infer activity types. Except for Montini et al. (2014) and Reumers et al. (2013), all other studies utilize GIS-related features. In studies developing random forest models, which compute and order variables by their importance (Gong et al., 2018; Montini et al., 2014; Yazdizadeh et al., 2019), variables based on the distances to POIs are the most significant. Gong et al. (2018) and Oliveira et al. (2014), by dropping out and employing again features, emphasize that adding spatial features considerably improves activity predictions, especially for non-mandatory purposes like eating out, banking and shopping (Oliveira et al., 2014). This argument is supported by Reumers et al. (2013), who reach only 20%, 26.9% and 7.1% success rates for social, shopping and leisure activities respectively due to lack of spatial features. In Lu et al. (2012), a land use-based model successfully labels 72.3% of total cases while those based on social demographics only and previous/next trip attributes solely produce accuracies of 51.9% and 60.6%, respectively.

The threshold of distance between a trip end and a place varies among studies, possibly resulting from the combination of four factors, namely (1) GIS data types, (2) device types, (3) place types and (4) quality of GPS data. An explanation of these factors follows:

- First, GIS data, which are available for regions or downloaded directly from Google Places (Cui et al., 2018; Furletti et al., 2013; Meng et al., 2017) or OpenStreetMap (Furletti et al., 2013; Krause and Zhang, 2019), exist in two forms, that is, a point representing a (small) place (i.e. POI) or a polygon representing a large place. Once polygon-shaped information is used, if a trip end is perfectly within a land use area (Deng and Ji, 2010; Xiao et al., 2016) or within a short distance of it (e.g., 25m of a commercial area in Oliveira et al. [2014]), the type of this area is assigned to the trip end. Once POIs are used, the higher distance thresholds are determined, shorter than 150m to a church in Oliveira et al. (2014), for example.

- Second, unlike personal devices, onboard devices fail to access within places. Consequently, the distance thresholds, which are indicated as the maximum walkable distance between an origin/destination and a parking lot (for car data) or a drop-off point (for taxi data), tend to be set at a (very) high value of 1000m (Furletti et al., 2013), 500m (Lu et al., 2012) or 300m (Krause and Zhang, 2019). Since taxi is a nearly door-to-door service, the threshold would be smaller (e.g., 200m in Gong et al. [2012]).

- Third, distance to known places and often visited could be higher than that to unknown ones. Bohte and Maat (2009) determine 50m for most POIs and 100m for homes and workplaces whose addresses are provided by participants.

- Fourth, for data with limited quality, the distance thresholds to POIs are high—for instance, at 250m in New York due to strong urban canyon effects (Chen et al., 2010) or 150m owing to the use of low-cost devices (Usyukov, 2017).

Original types of POIs and polygon-shaped locations vary across cities/regions and are much more numerous than activity types, which indicates a need for grouping them (Deng and Ji, 2010; Gong et al., 2016; Xiao et al., 2016; Yazdizadeh et al., 2019). Twenty polygon-level and 18 POI-level land use types in

Shanghai (China), for instance, are aggregated into 10 categories (Xiao et al., 2016). Street maps are used by Liao et al. (2007) to support trip end detection.

Schedules of land use types, which are not employed in all TS studies, are deployed in HG ones (see Table 2-5). A location will be disregarded if it closes when a person visits it (Furletti et al., 2013; Gong et al., 2016). Working time is unavailable in GIS data, leading Furletti et al. (2013) and Gong et al. (2016) to assign it to each type. To give an illustration, opening hours of shopping malls, government agencies and bars will be 9:00am–10:00pm, 8:00am–6:00pm and 2:00pm–03:00am, respectively (Gong et al., 2016).

**** Trip and activity-related***

An activity is the main business undertaken in a significant location; hence, it is in harmony with a purpose. A trip is a one-way course of travel to conduct an activity. Trip and activity have a close relationship (Ho and Mulley, 2013). Temporal attributes of trip and activity (i.e., time of day or day of week) are the same and listed as features of activity. Yet, mode, speed, distance, start time and end time characterize a trip. Except for Furletti et al. (2013), other authors deploy features related to either activity or trip or both to construct their models.

As for activity, the four characteristics, namely duration, time of day, day of week and start time are used most frequently (see Appendix 1); however, some of them rather than all are integrated in each study. If home type is neglected, duration and start time are sufficient for the models (Feng and Timmermans, 2015; Oliveira et al., 2014). The reasons would be that start time may be equivalent to time of day (e.g., morning, afternoon, night) and the confusion of duration between work and home does not exist. Activity history-related features occur in two studies (Chen et al., 2010; McGowen and McNally, 2007). Chen et al. (2010) conclude that the frequency of visiting locations, estimated from one-day data, carries little meaning. Based on the frequency of visiting a specific location per day, McGowen and McNally (2007) estimate the accumulated duration and give value to a dummy variable (1 for re-visiting case). Temporal profiles of activities are sufficient for detecting basic purposes (such as work and home) (Reumers et al., 2013). Shen and Stopher (2013) propose tour[2]-based corrections to enhance accuracy by approximately 8%. Notably, activity information is not available for HG studies using taxi data (Chen et al., 2019; Gong et al., 2016; Wang et al., 2017).

Trip attributes are employed in Deng and Ji (2010), Feng and Timmermans (2015), Gong et al. (2018), Lu et al. (2012), Montini et al. (2014), Oliveira et al. (2014), Xiao et al. (2016) and Yazdizadeh et al. (2019). Cui et al. (2018) use trip characteristics but not activity ones to impute purposes. Importantly, mode information is taken directly from ground truth in all existing studies while mode and purpose must be imputed in GPS data processing (Shen and Stopher, 2014). In this sense, mode detection and purpose imputation are now considered separately, although they should be combined in a continuous process. Travel mode is not included in HG studies since most of them use vehicle-based data (Chen et al., 2019; Furletti et al., 2013; Gong et al., 2016; Usyukov, 2017; Wang et al., 2017).

**** Participant-related***

Personal information is used in TS studies but not in HG ones, many of which employ anonymous data (i.e., taxi data) (Chen et al., 2019; Gong et al., 2016; Wang et al., 2017).

Addresses of home and work, which are seen in Bohte and Maat (2009), Gong et al. (2018), Montini et al. (2014), Shen and Stopher (2013), Stopher et al. (2008), Wolf et al. (2004) and Yazdizadeh et al. (2019), help in boosting considerably the overall classification performance since the number of work and

[2] Tour is defined as trips and activities occurring between a participant leaving home and returning home (Shen and Stopher, 2013)

home activities is by far greater than the number of other purposes. In case home address is not provided, home location can be estimated satisfactorily by examining the first trip's origin and the last trip's destination every day (Usyukov, 2017).

Among socio-demographics, age and occupation (Cui et al., 2018; Deng and Ji, 2010; Gong et al., 2018; Montini et al., 2014; Oliveira et al., 2014; Xiao et al., 2016; Yazdizadeh et al., 2019) are used much more frequently than the remainder (see Appendix).

In comparison with the two above feature categories, participant-related features are less informative for directly distinguishing activities and play a supportive role in increasing the classifier power (Lu et al., 2012; Shen and Stopher, 2013). Without participants' information, 96.8% of activities of 329 people in the Netherlands are correctly identified by Feng and Timmermans (2015).

**** Other features***

In TS, Gong et al. (2018) test the use of weather-based variables (e.g., temperature, precipitation and snow accumulation) in combination with the three above-mentioned feature categories by data collected in two distinct seasons in Hakodate, Japan. The authors report that weather features decrease the model accuracy without further explanation of reasons. This is a negative and surprising result given the significant relationship between weather and mobility in the literature (Böcker et al., 2013; Cools et al., 2010; Liu et al., 2017).

Social media (e.g., Twitter and Foursquare) information is useful for enhancing purpose inference in both TS (Cui et al., 2018; Meng et al., 2017; Yazdizadeh et al., 2019) and HG (Chen et al., 2019; Wang et al., 2017). Since social media data are noise, an information retrieval process is first conducted to find a place (i.e. POI) matching each post (e.g., tweet or check-in) (Cui et al., 2018; Meng et al., 2017). In this way, each post can be arranged into a category (e.g., food, education, leisure [Meng et al., 2017]). Then, the number of posts that belong to a specific category and are close to a trip end (e.g., with distance less than 250m [Yazdizadeh et al., 2019]) is used as a predictor variable. While studies in TS use social media data in tandem with geographical data achieved from other sources (e.g., Google Places and OpenStreetMap), city-wide POIs and their categories are achieved directly from Foursquare check-in data in HG (Chen et al., 2019; Wang et al., 2017). It should be noted that using social media data creates a lack of robustness owing to a variety of social media services and coverage. To give an illustration, in China, Weibo, a domestic social network, is far more common than Twitter or Facebook.

Table 2 - 5. Summary of input features

Feature category	Specific features	TS	HG
Geographic data	Polygon-based; POI; Street map	√	√
	Working time of land use types	-	√
Activity-related	Duration, Time of day, Day of week, Start time	√	√
	Activity history	√	√
Trip-related	Travel mode, Mode of next trip, Mode of previous trip	√	-
	Speed, Distance, Duration, Start time, End time	√	√
Participant-related	Home address, Work address, Occupation, Age, School address, Frequently visited places, Working hours, Gender, Driving license, Employment status, Income, Education degree, Race, Family structure, Household information, Marital status, Driving frequency	√	-
Others	Social networking: Foursquare, Twitter	√	√
	Weather: Temperature, Precipitation, Snow accumulation	√	-

2.6.2. Method group

This study follows the classification of Gong et al. (2014) to review three method categories in turn, namely (1) rule-based, (2) probability-based and (3) machine learning, before discussing how to validate prediction results.

**** Rule-based method***

Deterministic rules were the most common method for the initial era of purpose imputation because they are simple and easily interpretable. On the basis of polygon-based land use data and temporal characteristics of activities, Wolf et al. (2001) infer 10 purpose types to trips by 13 participants in Atlanta (USA) with a 79.5% accuracy. Using addresses for home, work and two frequently visited grocery stores, Stopher et al. (2008) report a set of rules to achieve an accuracy of over 60%. They improve their rules by deploying tour information to correctly classify 66.5% of 4,133 trips in the Greater Cincinnati region (USA) (Shen and Stopher, 2013). The lowest accuracy is 43%, reported by Bohte and Maat (2009), who use data of 1,104 respondents in the Netherlands to classify seven purposes by participants' home and work addresses together with a geographic database. They assign a location type to a trip end if the location has the shortest distance to the trip end. However, due to the error of GPS points, the closest would not be the right choice. A notable point from these four studies is that the accuracy is inversely proportional to the sample size, explained by an increase in the heterogeneity of travel patterns and activities in diverse samples. What's more, the quality of rules hinges on experts' knowledge of transport conditions in the research area. Rules are the least transferable since daily lives and travel patterns in different environments vary, requiring great efforts to re-calibrate and re-set up thresholds. Deterministic approaches have been used to construct the first steps of deriving purposes from GPS data. It is now integrated as a way to choose potential locations of a trip end or to estimate location of home for both TS and HG studies (Furletti et al., 2013; Usyukov, 2017; Xiao et al., 2016) (see Table 2-6).

**** Probability-based method***

Probabilistic approaches are more flexible than rules because they simultaneously take into consideration spatial and temporal confusion between purposes.

In TS, the first effort is estimating probability of each POI based on distance to cluster of trip end (Wolf et al. 2004) rather than choosing the closest like Bohte and Maat (2009). Nevertheless, socio-demographic variables are included in a deterministic way. Chen et al. (2010) introduce a multinomial logit model to analyze 49-person data in high-density areas of New York and classify four purposes into home-based group and non-home-based group with accuracies of 67% and 78%, respectively. Using data from 1,352 participants and more variables with the addition of trip-related variables, Oliveira et al. (2014) create a two-level nested logit model to distinguish 12 purposes with an accuracy of 60% at the expense of a time-intensive computation. Therefore, statistical methods would be better than rule-based ones in dense environments with high confusion between location types, yet they become much more complex and fail to adequately address a list with various purposes.

In HG, probability-based methods are the most common. Similar to Chen et al. (2010), Usyukov (2017) develops a multinomial model using start time and duration of activity to discriminate work activities from others. Furletti et al. (2013) propose temporal and spatial rules to choose a set of potential locations corresponding to each stop of a car. Only locations that are within a 1,000m-radius-buffer and open when a stop occurs are kept. Subsequently, the probability of each location is estimated based on the gravity model. The chosen location type has the largest probability. Gong et al. (2016) re-use the temporal and spatial compatibility criteria, which are integrated into Bayes's rules with distance decay effect, and

attractiveness of POI (i.e. service capability, size and fame) to create a probability function of visiting each POI. Wang et al. (2017) utilize probabilistic distributions (multinomial and Dirichlet) to first categorize POIs into 10 topics (e.g., shopping, university, office) based on the similarity of functionality, and then assign topics to both origin and destination of each trip. The addition of time helps in uncovering the semantic meanings of trip purposes.

Thresholds of (working) time, on the one hand, are impossibly set for every place, resulting in misclassification for specific cases. On the other hand, annotating activities with a degree of approximation is accepted for the HG authors, who aim at semantically enriching trajectories and discovering general activity patterns rather than focusing on precision at an individual level. Thanks to their flexibility, probabilistic methods are more powerful than rule-based ones. Rules are specific cases of probabilities. Compared with rules, statistical algorithms are less dependent on the subjective knowledge of researchers and are thus more transferable. They are capable of deducing purposes from far larger data sets (e.g., one-week data of 6,600 taxis [Gong et al., 2016]) (see Table 2-6).

**** Machine learning method***

Supervised machine learning (SML) algorithms are trendy in TS and considered a better choice than both deterministic and probabilistic methods. Their evolution can be divided into two main periods.

The period between 2001 and 2010 has witnessed decision tree models (C4.5 [McGowen and McNally, 2007] and C5.0 [Deng and Ji, 2010]). A more complex process, hierarchical conditional random field, is proposed mainly to detect stops (Liao et al., 2007). Most of these studies are small-scale tests. Deng and Ji (2010) use 226 trips by 36 participants while Liao et al. (2007) use data from three persons to train their model before validating it with the trips of a fourth person. With small sizes, samples would be homogeneous, resulting in high accuracy levels of 85.2%–90.6% (Liao et al., 2007) or 87.6% (Deng and Ji, 2010). With a larger sample of 17,000 households, McGowen and McNally (2007) report the precision at 73%–74% for five major activities and only 63%–64% for disaggregated purposes.

Since 2011, a boom in building up and enhancing a host of SML algorithms has been seen. Lu and Liu (2012) and Oliveira et al. (2014) continue employing decision tree and fail to generate an accuracy of over 80%. In contrast, Montini et al. (2014) pioneer a random forest model correctly classifying 84.4% of total trips by 156 persons in Zurich, Switzerland. Notably, random forest has replaced decision tree to become the most-utilized method (Feng and Timmermans, 2015; Gong et al., 2018; Montini et al., 2014; Yazdizadeh et al., 2019). It hits 96.8% accuracy with 329-person data in the Netherlands (Feng and Timmermans, 2015). However, it tends to encounter bias towards the categorical variables with many more levels (Deng et al., 2011) and the activity types occurring more frequently in training sets (Montini et al., 2014).

Some authors display impressive performances with improved neural networks. Xiao et al. (2016) use particle swarm optimization instead of back propagation in artificial neural network structure to obtain a 96.5% accuracy. By estimating parameters between nodes of different layers using probability distribution based on Bayesian inference rather than fixed values, Cui et al. (2018) successfully label 90.5% of activities.

Evaluating SML algorithms across studies would be questionable owing to the differences in purpose lists, features used, sizes and homogeneous levels of samples, not to mention the effects of the trip under-reporting problem in prompted recall surveys. However, in some studies reporting performances of multiple methods (Cui et al., 2018; Feng and Timmermans, 2015; Meng et al., 2017; Xiao et al., 2016), comparisons are feasible. Unsurprisingly, newly proposed models yield the best performances. Random forest has the second highest accuracy (Cui et al., 2018; Meng et al., 2017), while support vector machine

only generates reasonable accuracy levels, which are much lower than the highest (Cui et al., 2018; Meng et al., 2017; Xiao et al., 2016).

Contrasting opinions about the power of Bayesian networks are reported. Observing the accuracy of 46.2%, Feng and Timmermans (2015) conclude that it is not appropriate for inferring purpose, contradicting the high accuracy levels of 87.8% in Meng et al. (2017) and 86.6% in Xiao et al. (2016).

Recently, with the view that a purpose, which is decided at origin, will determine the trip end type, Krause and Zhang (2019) integrate a decision tree-based purpose inference into a model aimed at predicting vehicle destination. The purpose of a trip is inferred by features related to the trip origin before the derived purpose is used to estimate the trip destination.

Based on the comparison of Oliveira et al. (2014), probability-based models are outweighed by SML algorithms. SML is the most transferable because it learns from data to make predictions instead of being explicitly programmed by rules. In addition, it is suited to big data like GPS and GIS data as well as a host of input variables (Gong et al., 2018; Xiao et al., 2016; Yazdizadeh et al., 2019). For example, 29 features are deployed in Cui et al. (2018). As for SML's shortcomings, they need ground truth to train models and numerous data to avoid the overfitting problem. In that SML is a complex structure with many hyper-parameters (e.g., learning rate, number of trees), it is challenging to interpret magnitude of variables. A typical matter for SML is how to form training and test sets effectively (see Table 2-6). The common way is by dividing equal proportions of observations belonging to each purpose (e.g., 75% for training and 25% for testing [Montini et al., 2014]). However, Gong et al. (2018) highlight that data in training and test sets coming from different, distinct seasons weaken prediction results.

Authors in HG develop both SML and unsupervised machine learning (UML) models. Reumers et al. (2013) introduce a decision tree model, employing temporal features, to impute six purposes. Without ground truth, Chen et al. (2019) propose a novel combination of autoencoder and K-means. Autoencoder is an unsupervised neural network whose encoder maps a 22-dimension (i.e. original features) input vector to a more representative vector with 10 dimensions. Then, each trip that is represented by a 10-dimension vector is classified into one among five clusters (i.e. purposes) by K-means algorithm. One hundred thousand trips are used to train the autoencoder network. Because there is only one study, it is difficult to assess the transferable level and the importance of splitting data for UML. UML's shortcomings are complex structure with a need for big data coupled with low interpretation. Furthermore, its prediction is validated reasonably at aggregated level only (see Table 2-6).

Table 2 - 6. Synthesis of purpose inference methods

	Rule-based	**Probability-based**	**Supervised machine learning**	**Unsupervised machine learning**
Transferable level	Low	Medium	High	-
Suitable data size	Small	Medium and big	Big	Big
Role of data selection/division	Minor	Minor	Major	-
Number of variables	Small	Small	Large	Large
Ground truth	Optional	Optional	Mandatory	No
Mainly used in	Estimating home address; Choosing location candidate	Human Geography	Transportation Science	Human Geography
Power/performance	Low	Reasonable	High	Reasonable
Interpretation	High	Medium	Low	Low

2.6.3. Purpose list and validation

Prediction assessment is based on counting the matching times between the prediction and the ground truth for each category before summarizing them in a confusion matrix (Sokolova and Lapalme, 2009). The main measurement indexes are precision and recall for purposes and accuracy for the model. While the majority of studies report the confusion matrix, some (Chen et al., 2010; Lu and Liu, 2012) present the precision levels at aggregate levels only. The accuracy is not reported in Gong et al. (2018) or Krause and Zhang (2019).

Generally, the overall accuracy is inflated by identification of home trips, which are inferred at a high success rate and account for the largest percentage. When home activities are taken into consideration, the accuracy increases by 15% to 80% (Oliveira et al., 2014).

How to define purposes is scarcely presented. Most authors do not pay attention to mode transfer activity, although Montini et al. (2014) and Oliveira et al. (2014) do. Transferring from one mode to another is the purpose of a separate movement stage but not of a trip connecting between two significant locations. Because transition between stages during a trip is frequent, the inclusion of mode transfer exaggerates the model performance.

In a case of unavailable ground truth, Usyukov (2017) uses data from the 2006 Ontario Transportation Tomorrow Survey, which is believed to sufficiently represent the travel patterns of inhabitants. Because the shares of purposes extracted from GPS data are nearly in line with those obtained from the household survey, the model performs satisfactorily. The consistency between spatio-temporal patterns extracted from GPS data and the common distribution of activities according to time and space is deployed to validate models (Chen et al., 2019; Gong et al., 2016; Wang et al., 2017). To provide illustrations, the spatial distribution of home purpose is positively correlated with that of residential areas, and the direct distribution of all trips aligns with the geometry of central business districts in the research area (Gong et al., 2016). Since trips for the same purpose tend to have similar temporal distribution, the model of Wang et al. (2017)—with its smaller estimate of average time difference of trips annotated identically, as compared to its counterpart generated by a previous method—is validated.

2.6.4. Existing questions in purpose imputation

**** Feature/variable selection***

The above review emphasizes the great importance of geographic databases in purpose imputation, but relying on such databases would be the biggest barrier to carrying out purpose detection in areas where facilities and locations are not geo-coded well. GIS data, which themselves are numerous, are categorized differently in different regions, requiring great efforts to merge them with GPS data, especially for the simultaneous use of GIS data of several regions. In fact, to limit the complexity of problems and amount of computation, almost all existing studies involve experiments conducted in a specific area and neglect purpose imputation for long-distance trips. Krause and Zhang (2019) stress that including long-distance trips may exert adverse effects on purpose prediction results, but they do not show a detailed figure. Accordingly, it is worth digging and testing more features related to participants, activities and trips. Without geographic variables, a model tends to become more transferable and to be utilized not only to include several areas in a comparative analysis but also to identify purposes of long-distance trips.

As for participant-related features, more socio-demographics put a heavier burden on participants and thus result in a higher incomplete survey rate (Armoogum et al., 2014). Additionally, some locational information, such as school addresses, would be sensitive to parents (Feng and Timmermans, 2015).

Therefore, a model using the minimum number of these features would be welcomed. Future studies should carry out more tests and report the sensitivity of their models while adding or dropping person-specific variables.

Transportation mode and purpose are desired by researchers. In GPS data processing, mode detection is the step prior to purpose imputation (Shen and Stopher, 2014). However, up to now, they have been handled separately. As discussed above, trip mode is informative for inferring purposes, but all existing studies use mode information extracted from ground truth. Therefore, deducing purposes from GPS data with the support of the results of trip identification is promising, likely introducing the potential for incorporating both in an automated process.

Using social media information enables an improvement of detection but with lack of robustness. When driving, attending classes or working, people are less likely to tweet compared to when going shopping or eating out. Moreover, social media data would be biased toward young people and against the elderly (Chen et al., 2019). There are a number of social networking services (e.g., Facebook, Twitter, Foursquare, Weibo), and they offer different ways of accessing and exploiting their data. Accordingly, more efforts should be invested in testing different social media networking services to clarify how they complement participant, activity and geography-related variables to help in obtaining satisfactory purpose imputation results.

Reports of the negative or marginal roles of some features spur more tests. Historical trip frequency should be extracted from data collected during an adequate period, rather than from one day as in Chen et al. (2010). The negative effects of variables related to seasonal phenomena (e.g., snow, temperature) on purpose imputation results require more attention to draw a comprehensive conclusion and uncover reasons for their influences.

**** Data selection and enhancing algorithm***

Developing and enhancing machine learning-based methods require masses of data collected over several years (Feng and Timmermans, 2015; Gong et al., 2018), leading to a concern over data selection. First, as a result of the seasonal variability of travel patterns in different weather conditions (Liu et al., 2017), training and test sets should not be formed by data from distinct seasons (Gong et al., 2018). Second, different people display different numbers of each type of activity. Division based on the rate of participants may induce a lack of specific trip types in the training process. In addition, different people tend to carry out the same activity in dissimilar ways; therefore, training and test sets should have a similar distribution of socio-demographics. The most common way is now randomly using the same percentage of each purpose type. This method ignores the logic between consecutive trips in a chain. If many more trips come from several specific participant groups (e.g., the young group), the prediction may be weakened in that the model over-fits these groups. Accordingly, thinking about how to populate training and test sets to ensure similar distribution with respect to seasonal conditions, people and activity would be benefit future researchers.

Generally, basic machine learning algorithms (e.g., decision tree, random forest, artificial neural network, support vector machine) have been constructed. To bring about methodological improvement, combining them in hybrid models to limit the long-lasting problem of over-fitting should be considered. Bayesian network may need to be tested in different circumstances to clarify the contrasting reports in Feng and Timmermans (2015), Meng et al. (2017) and Xiao et al. (2016).

**** Validation and assessment***

The variability of purpose categories prevents cross-evaluation among studies, raising the question of what constitutes an optimal list of activity types. An ideal list may be determined at the minimum level based on the balance between the specific research targets and the requirements of resolution quality for the travel demand forecast (Oliveira et al., 2014). Furthermore, it is worth considering the nature of data instead of using predefined purpose categories. In this regard, UML methods may be applied to classify activities into groups sharing mutual characteristics with unknown labels. The results of these classes would be serious candidates for deciding purpose lists. In any case, the mode transfer purpose should not be included if trip is taken into consideration. To limit the inflation of the overall accuracy, a higher penalty may be imposed on misclassified home cases, or home activities may be disregarded.

The error propagation from trip end detection to purpose imputation result would be a useful area for research. Dividing a trajectory is subject to under- and/or over-segmentation problems. While there is no way to detect neglected trip ends in order to fix under-segmentation, over-segmentation is generally preferred (Montini et al., 2014; Shen and Stopher, 2014). The challenge is how to detect a spurious trip end, which is either a mode transfer point or simply a stop at a traffic light. A potential solution is to detect purposes of all potential trip ends with a certain level for each. Subsequently, using the most certain ones as well as logic in trip chain or tour could help in detecting and eliminating wrongly labelled trip ends. The rate of unidentified spurious ends reflects the error propagation level of the trip end detection step.

Duration, start time and end time of an activity are estimated from timestamps of GPS points. Estimation from GPS is subject to technical and random errors (Chen et al., 2010; Stopher et al., 2008) whilst prompted recall travel diary tends to be prone to the under-reporting problem. Therefore, it is better to compare start time, end time and duration of both to determine the accuracy rather than based on the number of matching between detected result and reported diary only. Another problem that has never been documented is that what is the matching because the difference in time nearly happens for all cases (Kelly et al., 2013) (see Figure 2-2). An idea is to propose an accepted range for the difference. For example if start time of activity predicted differ that of reported one more than five minutes, the prediction is considered a failure. Because duration is typical for purposes, the accepted range can be determined according to activity types, a rate of average duration extracted from national travel household surveys for example. This proposition seems to be more related to assess the trip end detection than purpose imputation; however, validating segmentation by results of purpose imputation is worth trying.

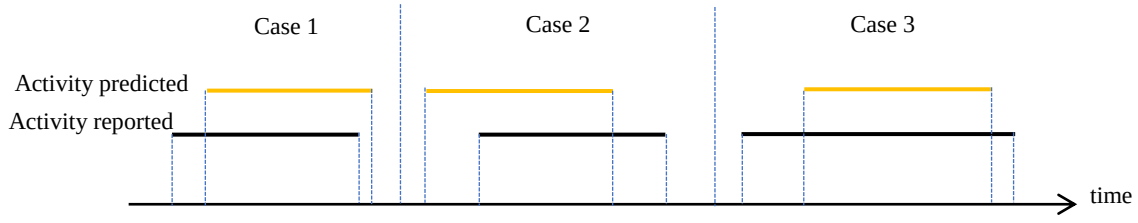


Figure 2 - 2. The difference in time of activities predicted and reported is common. In case 1 and case 2, duration of both is similar but the predicted starts later and sooner, respectively. In case 3, the predicted shows much shorter duration due to being imputed starting later and finishing sooner. And the question is that in which case(s) an activity is correctly inferred.

Inferring purpose from GPS surveys is seen predominantly in Belgium, the USA, the Netherlands, Japan, Canada, Switzerland and Australia (Bohte and Maat, 2009; Gong et al., 2018; Stopher et al., 2008; Wolf et al., 2004, 2001; Yazdizadeh et al., 2019), while among developing countries, China is the only

venue of such studies (Deng and Ji, 2010; Xiao et al., 2016). Because working, living and transport conditions in emerging countries are distinct from those in developed countries (Huynh, 2020; Nguyen et al., 2019b; Nguyen and Pojani, 2018; Pojani, 2020), it is worth carrying out more experiments in these nations.

Purpose imputation from GPS data is a complex problem that a wealth of studies have attempted to address. Here, we provide a critical synthesis of these studies based on the process of addressing a purpose inference challenge and according to two domains, TS and HG. In reality, researchers always confront one of two situations: analyzing a post-collected dataset or designing a survey to create an enhanced inference model. In both cases, positioning the domain of research is vital. For example, a post-gathered dataset without ground truth should be treated as a study in HG to semantically annotate GPS stream. On contrary, if aiming at a methodological contribution, researchers may need a dataset with corresponding travel diary collected in a prompted recall survey. Although applying the same process with trip end detection, feature selection and prediction result evaluation, studies from TS and HG draw up various schemes with both similarities and differences due to dissimilarity of research focus. This review, therefore, could help with efforts to obtain a clear understanding of purpose imputation in both domains. Furthermore, the unresolved issues mentioned in Section 5 could serve as suggestions for future research. Dealing with these issues is the key to (1) integrating purpose and mode imputation into an automated and continuous process, enabling GPS travel surveys to replace, in part or completely, conventional techniques and (2) gaining a better understanding of how human beings travel from GPS data.

2.7. SUMMARY

Based on the rigorous discussion above, it can be seen that GPS-based surveys have advanced considerably for the last two decades. This period has witnessed the rich literature of both mode and purpose imputation fields due to the lack of trip characteristics in positioning data, which has mainly prevented GPS from replacing completely conventional techniques (e.g. face-to-face interviews, CATI). Compared with mode detection, purpose identification has been understudied. Both of them have been primarily conducted on data collected in developed countries and modern cities of China, leading to the little understanding of (1) the potential of both GPS-assisted travel surveys and (2) the application of mode and purpose inference models in developing countries. The subsequent chapters concentrate on:

- Developing/extending algorithms to detect transportation modes in motorcycle-overwhelmed urban area of a developing country.

- Developing/extending algorithms to detect trip purpose without the use of GIS data in urban area of a developing country.

Besides, to challenge the ability of GPS-based surveys to be alternative to conventional household surveys, we compare findings of travel patterns estimated from GPS data and CATI data. The comparative results are the base for assessing algorithms created to take trip characteristics from GPS data.

Finally, recommendations for alleviating and/or addressing existing impediments to GPS surveys, particularly for developing countries are proposed.

Chapter 3: DESCRIBING SURVEYS IN RHONE-ALPES (FRANCE) AND HANOI (VIETNAM)

3.1. RHONE-ALPES SURVEY BY DEDICATED DEVICE

3.1.1. Regional household travel survey

The regional household travel survey in Rhone-Alpes was conducted by (CATI) technique and frequently indicated as EDR-RA, an acronym of *"L'Enquête Déplacements Régionale Rhône-Alpes'*. It originated from the need of observing continuously citizens' mobility in the long term to obtain the practice of travel pattern and evaluate transport policies towards sustainable development (Armoogum et al., 2018b). As for geographic survey scope, the Rhone-Alpes region was a large territory in the east of France and comprised of the Rhone region and the Alps mountainous region. Now, Rhone-Alpes is a part of Auvergne-Rhone-Alpes.

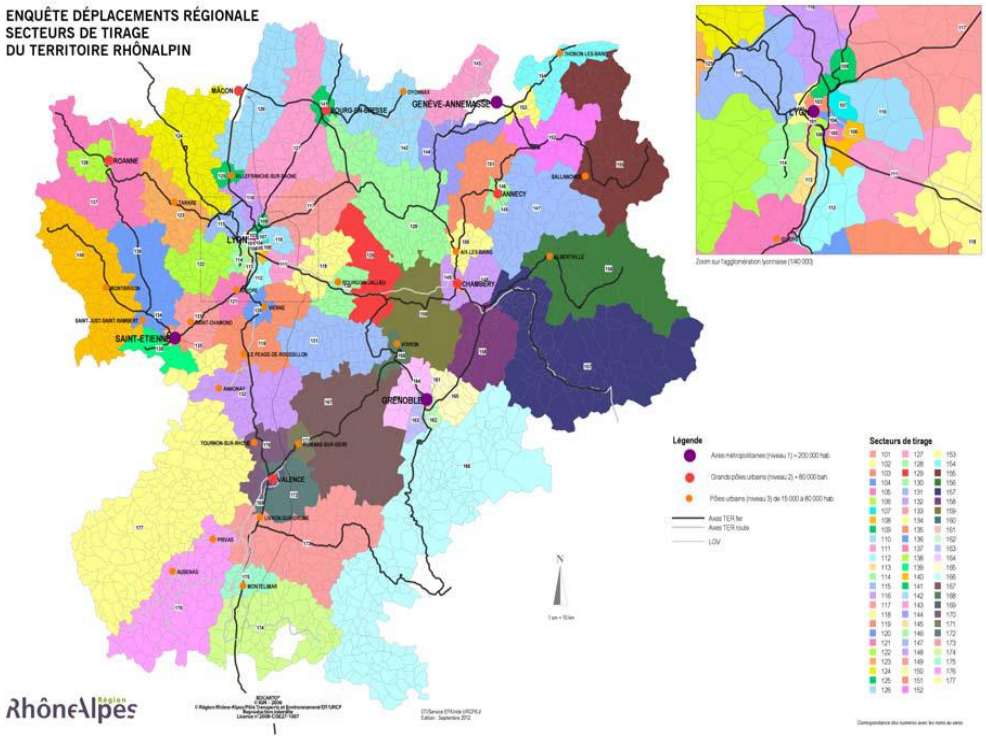


Figure 3 - 1. Scope of the Rhone-Alpes survey from 2012 through 2015
Source : (Armoogum et al., 2018b)

The three-phase survey from 2012 to 2015 was undertaken based on CERTU methodology. Specifically, randomly selected participants of a recruited household was asked to answer a series of questions categorized into four parts: (1) characteristics of household, (2) individual socio-economic characteristics, (3) trips during a (reference) weekday, and (4) opinions on transport problems (Bayart and Bonnel, 2015). The region was divided into 77 sections to make samples collected well represent the entire population. Those questioned were aged 11 and over. Each phase lasted during two years, from October of the previous to March of the subsequent. The survey time excluded the holidays of both the public and school and had fairly balanced distribution of day types ranging from Monday to Friday. After three phases, the total of 37,456 persons from 32,293 households reported 143.4 thousand trips.

3.1.2. GPS-based experiment

TOMOS, an acronym of *"Test de l'Observation de la Mobilité par Suivi GPS"*, is a sub-project of EDR-RA. It was carried out during two periods of 2015, from March 12 to April 11 and from May 05 to June 05, respectively. The TOMOS's sample encompassed the individuals who had participated in the second phase of EDR-RA (between 2013 and 2014) and were not under 18 years old. They lived in and around Lyon city that spreads over 47.9 km^2 and is both the second largest city in France and the capital of Rhone-Alpes region. During the course of observation, each of them was distributed a dedicated lightweight device designed by AlyceSofresco integrated a GSM chip to transfer records to the server on a regular basis, thus the issue of data storage was handled. With the hope of collecting travel traces in both weekend and weekdays, a participant was requested to take a GPS logger when leaving home during at least 7 consecutive days. Before the beginning of the mobility investigation, (s)he had to respond some questions similar to the first and second parts in the mentioned-above questionnaire of EDR-RA. Participants were not asked to provide travel diaries corresponding to days observed, which lessened the burden on them and thus fostering their involvement into the survey.

A GPS record was characterized by the identifier number of a device, longitude and latitude in decimal format together with timestamp. The interval of collection varied around 10 seconds to mitigate the battery drainage. The synthesis of TOMOS can be found in (Armoogum et al., 2018b).

3.2. HANOI SURVEY BY SMARTPHONE

3.2.1. TravelVU, the smartphone application of the Hanoi survey

TravelVU that is a software dedicated for smartphones was chosen for the study. It is developed by Trivector Traffic AB, a Swedish company and hereafter indicated as Trivector. The app is in a trial process. Similar to previously introduced apps like ATLAS II (Safi et al., 2015), Future Mobility Survey (Cottrill et al., 2013), PEACOX (Montini et al., 2015), TravelVU can collect GPS data and their corresponding ground truth. GPS data are analyzed by in-built algorithms that run to implement segmentation (i.e. detecting segments). The app classifies data into two segments types, that is, travel segment and activity segment. For the former, information of start time, stop time, speed, travel mode, length, speed, duration, and route are estimated. For the latter, start time, end time, type, and location are estimated. This information may be input for further researches on travel behavior, energy consumption, and health.

To conduct a person's survey, procedures are as follows:

**** Step 1: Finding the app and installing it***

A user first seeks TravelVU on Google Play or App Store depending on his/her smartphone's operating system before installing it on his/her smartphone. TravelVU is compatible with both Android (version 4.4 and higher) and iOS (version 9 and higher) platforms.

**** Step 2: Selecting survey and answering background questions***

Concurrently, the app allows a number of surveys. After the app is available on a smartphone, a user has to choose the survey (s)he wishes to take part in. To successfully access the survey, (s)he enters a password or using a QR code. Communications presented by the app are in several languages like Swedish, English, and Italian. The user is asked to grant permissions for enabling the location function and disabling the (battery) saver mode or battery optimization to ensure to collect well data.

**** Step 3: Answering background questions***

Next, (s)he is requested to provide his/her background information, that is, gender, year of birth, household size with age distribution, possession of driving license, number of cars in household, educational level, residence country, occupation, and income.

**** Step 4: Providing data and validating data***

Provision of data means the app starts collecting and transferring data in real time to the servers if the internet connection is available. Here, a series of algorithms run to separate trajectories into trip segments and activities. Indicators of speed are estimated to label segments. Coordinates of bus stops and stations collected prior to the survey are employed to detect transit trips. For activity, imputing types to activities is indeed complex. The app uses the first validations of the user to determine his/her main locations (i.e. home and work). Based on the distances between locations of activities to these locations, the app can infer home and work purposes.

Data analyzed then are sent back to the user's smartphone via internet connection and users are requested to check and validate them. The objective of initially analyzing and classifying data is to lessen the burden on users. If the initial interpretation of the traces by the app gains high accuracy, users' missions are alleviated significantly. For activities, home and work segments are frequently recommended by the app whilst the user has to give his/her answer to other activity segments. For modes, all travel segments are labelled by the app. The number of travel segments exceeds the figure for activity segments because a user can travel by many modes between two activities. In case a prediction recommended is true, the user makes no action. By contrast, based on the list of segments ordered chronologically and accompanied by visualization of GPS data on the map, a user needs to recall and correct segments. By means of tapping symbols on the monitor, a user can add an activity by splitting a trip into two, merge two (travel/activity) segments into one, and delete segments.

Each user after checking and correcting presses "day is correct" on the screen to show their agreement with the precision of data and upload them to the survey's server. Users have 7 days after the end of survey to validate data.

**** Step 5: Finishing data collection***

At the end of the survey, a user can either uninstall the app or keep it; however, the survey is closed.

3.2.2. Data collection in Hanoi

**** Introduction of Hanoi***

The Hanoi Capital Region is the capital of Vietnam and hereafter called Hanoi. It is located in the central area of the Red River Delta and is the political, educational, commercial, economic and cultural center of Vietnam in general and in the Northern region in particular. As for population, it is the second largest city with the 7.32-million inhabitants in the area of 3344.4 square kilometres (General Statistics Office, 2017). Hanoi has a burgeoning population with an annual growth rate of approximately 2.34%. The geographical distribution of population is greatly unbalanced. Urban citizens accounting for 42.5% of the whole population live in 10% of the Hanoi area (i.e. the 330-km^2 area).

The central area of Hanoi comprises eight districts (i.e. Caugiay, Badinh, Hoankiem, Dongda, Haibatrung, Thanhxuan, Hoangmai, Tayho) and surrounded by the suburban and rural districts and town. Hanoi is distinct from cities of developed countries and China in terms of mode use with the dominant role of motorcycle (Nguyen et al., 2019b).

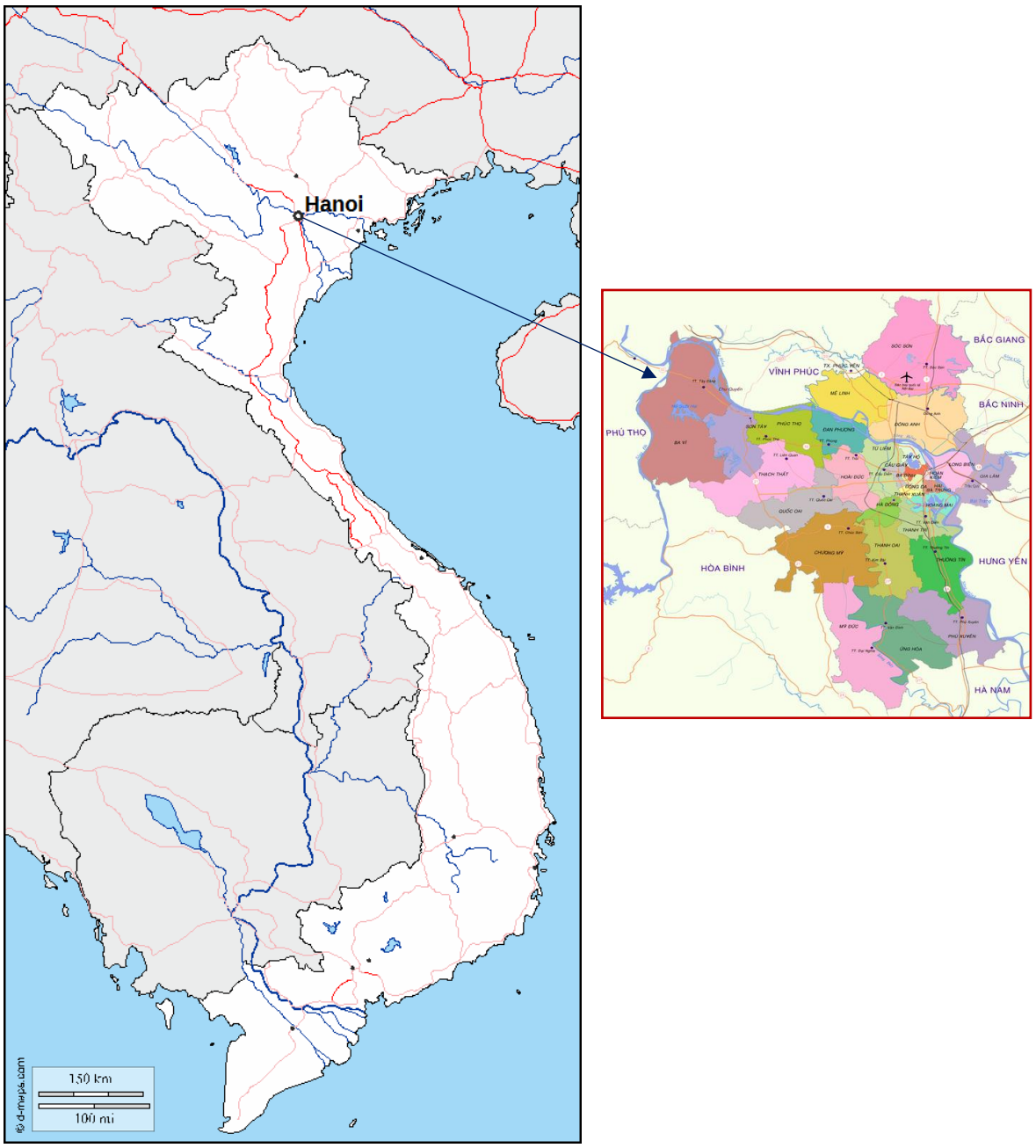


Figure 3 - 2. Hanoi map

Source: https://d-maps.com/carte.php?num_car=22366&lang=en (People's Committee of Hanoi, 2016)

**** GPS-based and smartphone-based survey in Hanoi***

In Hanoi, almost all surveys are conducted by face-to-face interviews and sometimes by telephone to collect data for making transport planning and public transport strategy during the 10-year period and vision to 20 or 30 years later (M. H. Nguyen et al., 2017; People's Committee of Hanoi, 2016). Interestingly, there was a GPS-based survey conducted in 2010-2011 by (T. T. Nguyen et al., 2017). The main objective of the survey is to compare trips recorded by wearable GPS devices with trips reported by participants. The results are evidence to clarify the potential of GPS technology and its existing drawbacks. 95 participants provided GPS data and their individual characteristics. After this survey, there have not been more GPS-based ones yet.

Until now, there have not been any smartphone-based surveys in which users actively participate in and give their data for research goals.

**** GPS-based survey using smartphone in Hanoi by 2019***

In a series of efforts to promote the continuous observation of mobility by GPS-enabled solutions, the laboratory DEST under IFSTTAR (France) signed a contract with Trivector to be eligible for running an application named TravelVU during a month on approximately 100 smartphones. The objectives of the survey are (1) assessing the potential of smartphone for observing travel patterns in urban areas and (2) developing methods to obtain trip information from GPS data. The survey time was determined from 2019-03-18 to 2019-04-15 to cover various travel modes. According to the official of the Ministry of Transport, the first metro line in Hanoi came into operation from April. Thus, we hoped that survey data would include metro trips; however, this kick-off was delayed.

- *Recruitment*

Firstly, ten motivated colleagues at University of Transport and Communications in Hanoi were invited and accepted to participate in the survey. Next, the authors made a Facebook post showing the scope and goals of the survey and tagging ten first respondents. The authors received many feedbacks from students, colleagues and relatives also. For those expressing interests in the survey, the authors kept in touch at individual level by Facebook Messenger to explain how to install and master the app to provide data effectively and efficiently. Once a user accepted to join the survey, the corresponding staff gave him/her the password to access the survey and a document of detailed instructions in Vietnamese with screen shots of the app. Based on the relationship between users, groups on Messenger and Zalo, a pervasive free message and call application in Vietnam, were created to save efforts of communications. For example, all users working at University of Transport and Communications were grouped. Answers to questions of a member can be read by others in the group. A participant could remove the app after at least a week with at least seven days validated. After reaching the target, validation was optional.

We got the updated numbers of persons installing the app by requesting them to take photos of your phone's screen. This way allowed us to know that users whether adhered to the instructions of turning off saver more and turning on location function or not.

- *Adapting collection of background information*

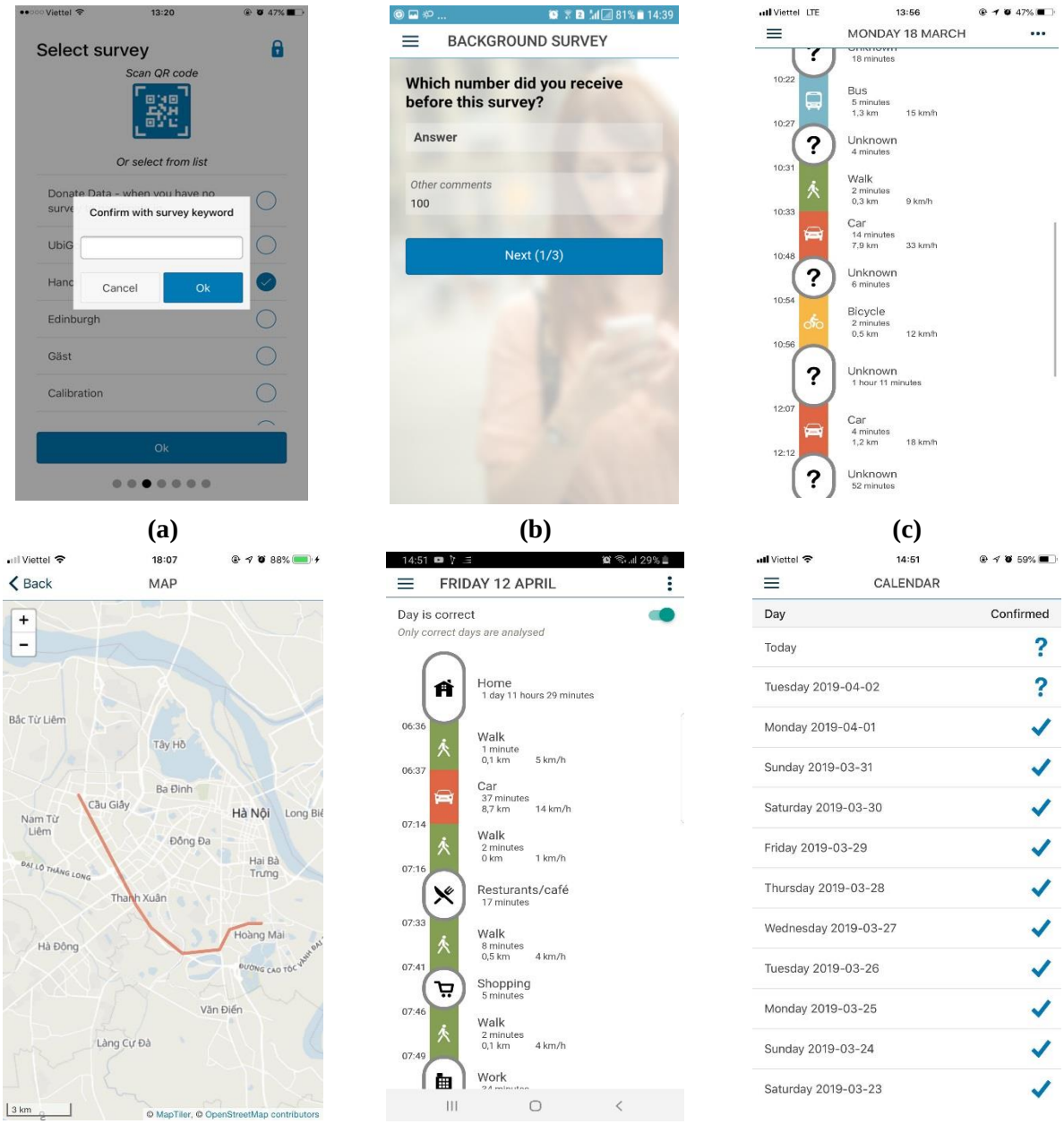
As indicated above, the use of smartphone apps to track travel behaviors is new in Hanoi whilst the most common technique to collect mobility data is face-to-face interviews. We realized four difficulties that participants would face.

The first is the language barrier. Generally, the English ability of the Vietnamese is not as good as citizens in European countries; therefore, answering background questions on smartphone in English can raise the burden and unpleasantness. In case a user does not understand exactly questions, information provided may be incorrect.

The second is the technological barrier. For many persons who are not familiarize with setting up apps, doing many tasks to complete installation would encourage them to ignore the survey or answer questions as quick as possible but without carefulness.

The third is users' understanding of concepts of the app. To be specific, persons are usually familiar with trip rather than travel segment, family rather than household. Although we made explanation for such concepts in instructions; however, there is a likelihood that participants would glance it to know how to find and access the survey in Hanoi rather than read carefully every detail.

The fourth is the lack of incentives. Due to financial limitations, all users were volunteers and we could not show our appreciation for them by a financial compensation. This reminded us of alleviating difficulties in installing the app as much as possible.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 3 - 3. (a) User was asked to input password to access the Hanoi survey; (b) User was asked to input the number assigned after the initial individual profile survey; (c) example of daily segments not validated; (d) Visualization a segment route created by GPS points, (e) example of daily segments confirmed, (f) confirmation status of survey days

We decided to remove background question part on the app, instead of this, participants were requested to answer them by various ways including on email, chat on messenger or phone call via messenger. These ways were useful because we have two-side communications, which enabled us to explain more about questions and limit bad answers. I took responsibility for a coordinator who received and gave prompted answers in Vietnamese to users during both the installation process and the survey process.

We made some modifications in the questionnaires. A question on whether frequently working outside the main workplace was added. Besides, the questionnaire aimed at collecting more details of possessing driving license types (i.e. for motorcycle, for car) and ownerships of vehicle types (i.e. bicycle,

motorcycle and car) and monthly bus pass. The question about the age distribution of household's members was removed.

To match background profiles with travel data of each person, after finishing answer questions a respondent was received a number to type in the process installation.

3.2.3. Results of data collection

Data including GPS points and their corresponding ground truth were first stored by Trivector. After the survey finished and was removed from the app, at the beginning of May, all data were delivered to DEST in JSON files and structured into segments. A segment assigned an ID number includes ID of phone collecting it, a series of GPS points, duration and length. A point encompasses latitude, longitude and timestamp. Points were collected at high frequency ranging between 1 and 3 seconds. Different sampling frequencies came from the collection by users' various smartphone models. Information of these phone models was not collected.

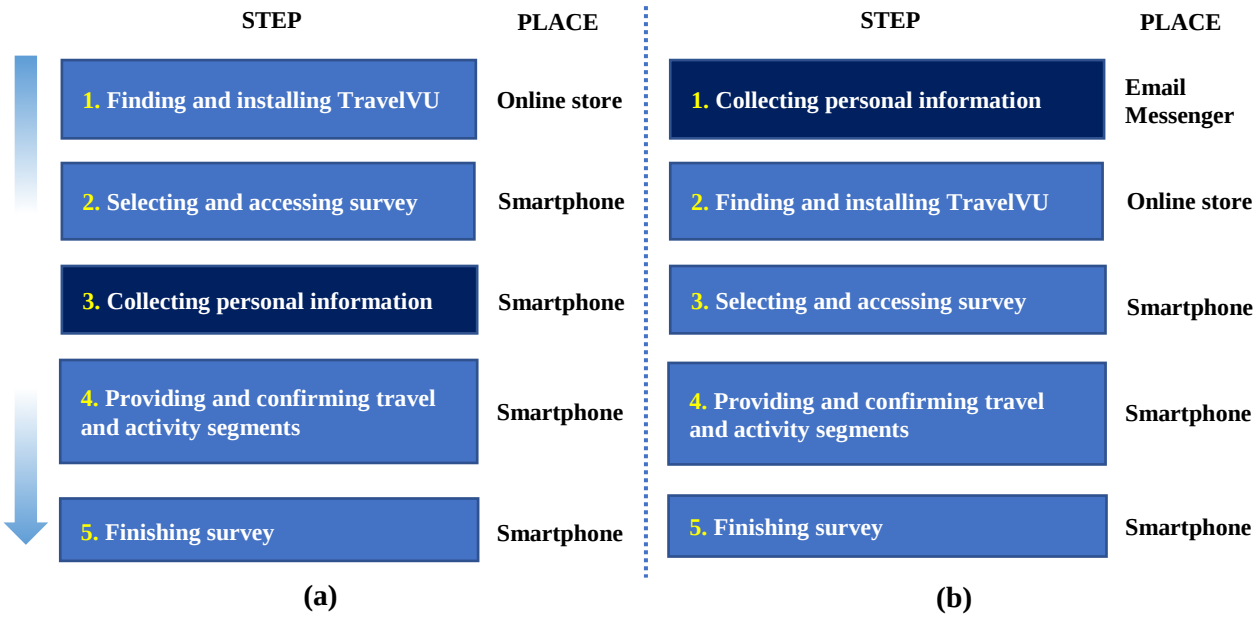


Figure 3 - 4. (a) Standard process of data collection by TravelVU; (b) Adapted process of data collection in the Hanoi survey

The list of participants who had accepted to install the app was comprised of 85; however, Trivector reported and delivered data of 72. So, the recruitment rate was of 84.7% that is much higher than the levels of 1.9% in (Stopher et al., 2018) and 39% in (Safi et al., 2017). The obvious difference in recruitment rates possibly come from the collection method. Whilst Hanoi survey focused on individuals by various communication techniques with the main use of social networking services, recruitment in the two previous studies were carried out mainly by CATI for households either participating in NHTS. Additionally, participants in the Hanoi survey had particular connection or relationship with surveyors more or less whereas random sampling was undertaken in (Safi et al., 2017; Stopher et al., 2018). Notably, 85 persons in our list and 72 users in list of Trivector were smaller than the number of registrations that achieved the maximum level of 100. This can be explained by several following reasons. First, some may register but they installed unsuccessfully it or simply did not install it. Second, the app counted the number of registered smartphones whilst we counted the number of potential participants. In fact, many in our list were using two phones simultaneously during the survey time. They could install the app for both smartphones but

validating data of one. The rest although registered did not record any data at all. Third, some may share the app's information with others who installed and removed it later. Fourth, some may uninstall the app because it failed to gather their travel data at all. Omission to collect data could result from the fact that a user (un-)intentionally forgot either turning on saver mode or turning off location function. Besides, technical problems of the app could be a culprit since it the app was a beta version.

Among 72 users, seven did not make any validation yet and two were staff conducting tests in Paris (France) and Lund (Sweden). As a result, we had 63 persons' validated data with 5652 travel segments and 4624 activity segments.

Data presented in Table 3-1 were comprised of behaviors in Hanoi, other provinces in Vietnam and abroad during the survey time. This is the reasons why ferry and metro that are not available in Hanoi were reported. We paid attention to the high share of bike compared with the very low use level in practice. Notably, based on data collected in Vienna (Austria) and Dublin (Ireland) in the project PEACOX, Montini et al., (2015) emphasized the over-representation of bike segments as an issue of smartphone-based survey.

Table 3 - 1. Mode share and purpose share in 63-person data collected

Travel mode	Frequency	Percent	Activity	Frequency	Percent
Walk	2065	36.54%	Home	1371	29.65%
Motorcycle	2034	35.99%	Work	1103	23.85%
Car	770	13.62%	Restaurant/café	329	7.12%
Bike	300	5.31%	Visit	261	5.64%
Bus	228	4.03%	Business	255	5.51%
Carpool	91	1.61%	Shopping	230	4.97%
Taxi	63	1.11%	Wait/Transfer	213	4.61%
Unknown	49	0.87%	Other	183	3.96%
E-bike	31	0.55%	Pick-up/drop-off	169	3.65%
Airplane	6	0.11%	Education	118	2.55%
Train	5	0.09%	Running	96	2.08%
Other	5	0.09%	Other errand	65	1.41%
Ferry	3	0.05%	Entertainment	63	1.36%
Metro	2	0.04%	Health	61	1.32%
Total	**5652**	**100.00%**	Unknown	48	1.04%
-	-	-	Hobby	42	0.91%
-	-	-	Temporary overnight	15	0.32%
-	-	-	Parking	1	0.02%
-	-	-	Prefer not to say	1	0.02%
-	-	-	**Total**	**4624**	**100.00%**

As can be seen in Table 3-1, the total number of travel segments and that of purpose segments are not equal. This is due to the multiple modes being used to travel between two meaningful places. For example, a user may walk to a parking lot to drive his/her car to the office. Consequently, the purpose is working whilst the travel includes a walk segment and a car segment. Although the list of purpose includes parking activity; however, the change between walking and driving would be so short that a user adds modes but not add a parking activity. Table 6-1 (see Chapter 6) gives an example where a participant reported two consecutive travel segments by walking and motorcycle but without an activity in the middle of them.

As for characteristics of users, lack of balance in terms of gender and age is apparent. Females made up 60.3% of total users. The elderly accounted for only 3.2%, overwhelmed by the figure for those aging between 23 and 60. The number of married users was double that of single counterparts. The vast majority of users (79.4%) were white-collar workers whilst the proportion of blue workers was much smaller at 9.5%. The sample also covered the retired and job seekers. 17 out of 63 users (27%) had dynamic jobs that required them to do usually business outside their main offices. The rate of users whose households included no child was insignificantly different from that of those living with children.

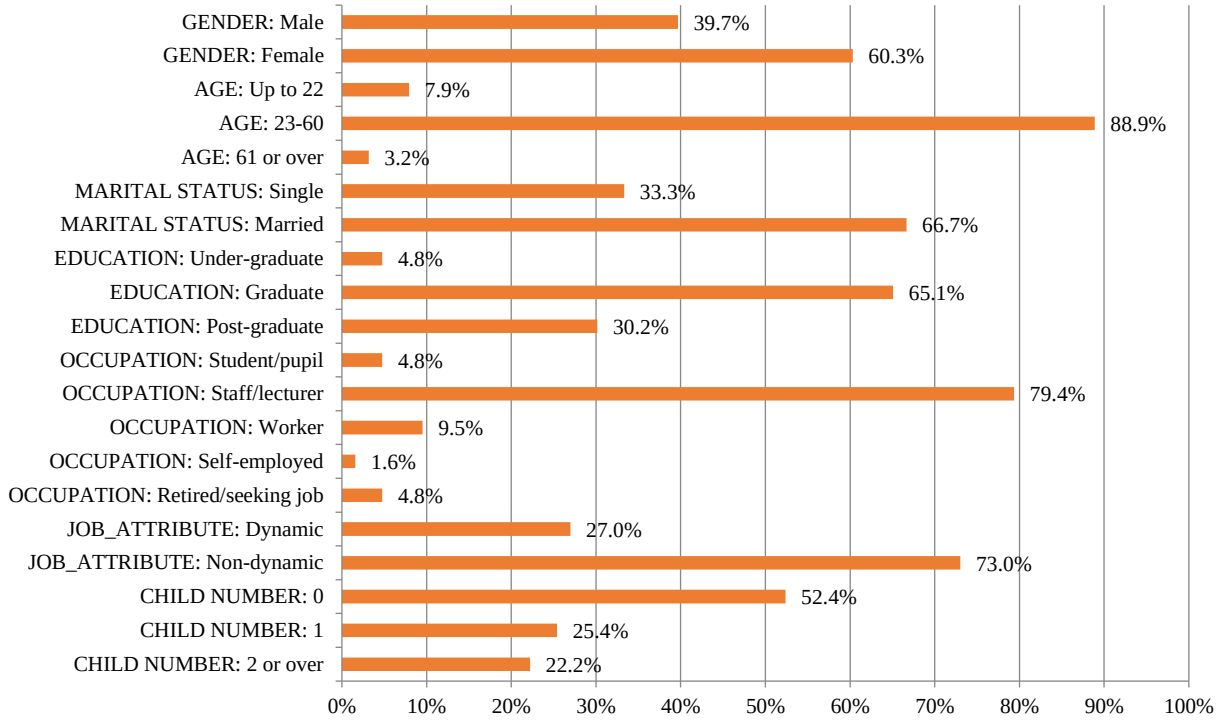


Figure 3 - 5. Profiles of 63 users

3.2.4. Respondents' views on TravelVU

In the end of May 2019, after doing some initial analyses on data, we conducted post-surveys to listen to feedbacks of participants about both advantages and disadvantages of the surveys in general and the app in particular.

**** Positive views***

Most of all showed their happiness and interest in the mobility survey using smartphone. The summary of travel by mode and time provided a clear picture of their daily travel patterns, especially how to make active travel, which they had never seen before. They appreciated the app's friendly display with pictures for choices, which allowed them to validate data with a limited English level. They also agreed with the great potential for using smartphone than self-reported investigations to collect mobility data.

**** Negative views***

- *High battery drainage*:

Over half users made complaints about the significant increase in energy consumption due to the operation of TRavelVU. Even, some ended up flat battery and had to take portable power banks. High drainage possibly came from the wish of collecting high-resolution data at the frequency of between 1-3s per point.

- *Wrong recommendation*:

As indicated before, data after transferred to the server were analyzed by a series of algorithms to label both travel and activity segments. Unfortunately, the vast majority of users showed their unhappiness because a large number of travel mode recommendations were wrong. The most serious misclassification was between motorcycle and bike segments. Even, several responded that TravelVU seemed to be designed to detect bike but not motorcycle segments. Every now and then, travelling by car was wrongly labeled as bike. In fact, we had anticipated this problem. The app has been developing based on data of Sweden and other countries in the Global North where car is dominant in mode use and bike is being promoted by dedicated infrastructure like bike lanes and bike-shared facilities. Bike can reach the maximum speed of up to 40 Km/h on cycle lanes (Bohte and Maat, 2009; P. Stopher et al., 2008a), which is impossible in Hanoi. Several persons were disappointed at the app's poor ability to detect home and work activities. The app remembered locations of home and work provided by users in first confirmations. Probably users validated wrong home and/or work segments, leading the system to save these locations and falsely identify home and work in the following process. On the other hand, in case a user validated a number of locations as home, the system was unable to do further analyses to give recommendations for home segments (see Figure 3-6).

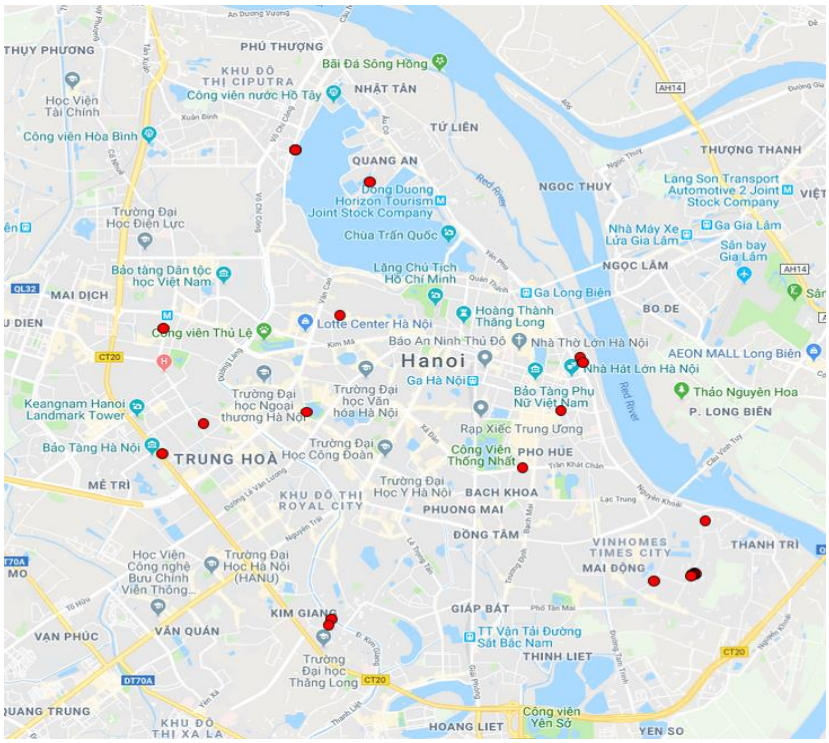


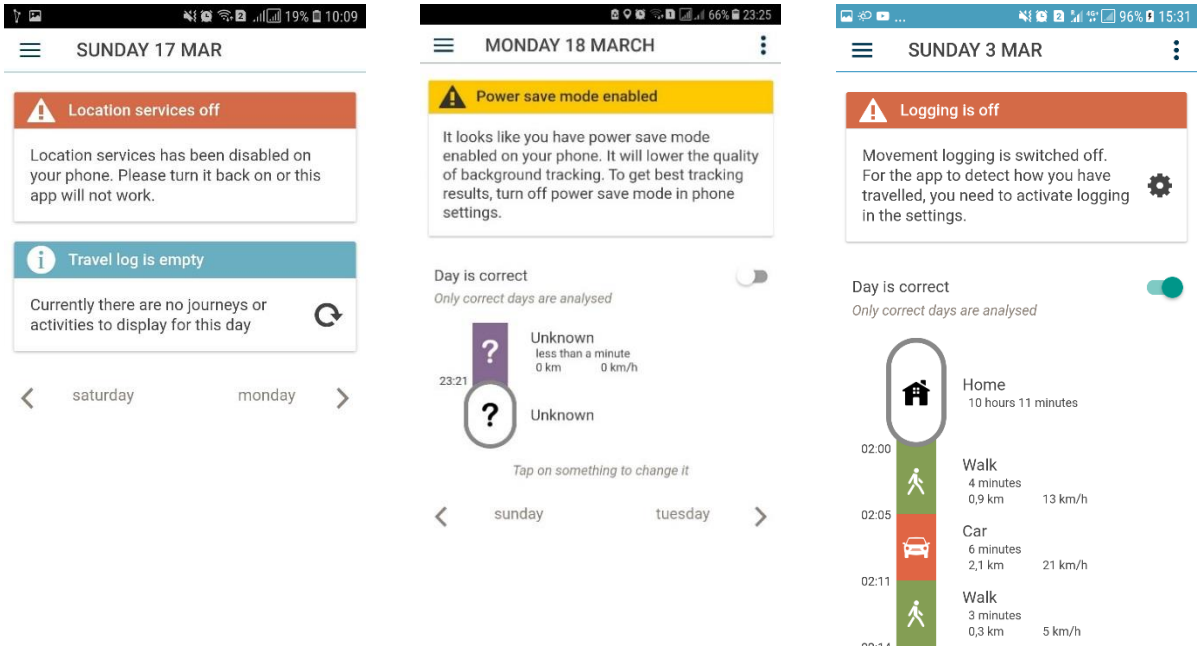
Figure 3 - 6. Typical example of inconsistent home location distribution of a user

- *Missing data*:

Omission to track data usually took place because of three reasons. Two first ones were either turning on power save mode or turning off location service on phone, resulting in failure to collect data at all. Some declared that they made a habit of activate power safe mode to extend the battery life. For a number of users, they actively deactivate location function and/or log of the app to save energy once the remaining battery was low (e.g. below 30%) or keep their positions in secret for example; however, forgetting turning on them or re-logging in the app.

There was an issue, which we could not found a persuasive explanation for it. Almost all users reported the loss of trips. This problem was more serious in the morning whilst almost all trips in the rest

of day were recorded sufficiently. For those aggressive with this situation, we proposed technical reasons in terms of save mode or location function; however, they confirmed that they had adhered to instructions. The evidence was the successful collection of data in the afternoon and in the evening. In (Montini et al., 2015), authors documented missing data in case of still turning on the app; however, there was not further explanation.



(a) (b) (c)

Figure 3 - 7. TravelVU cannot operate well because (a) location function of smartphone is off; (b) power save mode of smartphone is on; (c) the app is turned off

- *Long time to log in*:

The app usually spent very short time in starting whilst a user called it. However, after a week or two weeks it took long even very long time (several minutes) to operate and display information, causing considerable inconvenience for confirming data. The long lag may relate to uploading data to the server and receiving analyses from the server. Specifically, there were numerous data that had recorded but not uploaded due to the unavailability of internet connection. Whilst a user operated the app to check data and the phone had internet, the app immediately took time to upload and receive data. Information was displayed after the process finished.

- *Language barrier*:

A number of users did install the app by instructions displayed in Swedish because ignoring unintentionally choosing language in the first introduction of the app. In this case, users contacted the corresponding staff to receive step-by-step instructions in Vietnamese. There were two old persons asked their relatives to install and validate data due to limitations related to English and mobile use.

- *Privacy concern*:

Whilst many felt free to provide both background information and GPS data, some paid close attention to the privacy issue. They refused to reveal their addresses. They also emphasized that their data were used for research goals only and they should not be published or shared in any cases.

3.2.5. Correcting ground truth data

Data confirmed by users are used to train or calibrate models, thus its quality is very important. In fact, the reliability of ground truth can be checked for some segment types. Here, we took consideration into: bike segments, bus segments, and business segments.

**** Checking bike segment:***

Compared with other modes like motorcycle, car and bus, the rate of bike use is much smaller. 5.31% of segments were reported by bike, which made us suspect the reliability of confirmations for this mode. In the survey that aimed at collecting users' views on TravelVU, we asked them about whether travelling by bike during the survey time in Hanoi. Once a user said "no", we further examined their households' possession of bicycles. Interestingly, all those answering "no" came from non-bicycle households. Thus, we had strong belief that bike segments confirmed were false and we neglected them.

**** Checking bus segment:***

There is confusion in terms of using terminologies for travel modes in Hanoi. Whilst designing the survey, we defined that bus type refers to the urban or suburban services that strictly follow a predefined routes and predefined timetable. Specifically, a bus should stop at transit/bus stops to allow boarding and alighting. Besides bus mentioned-above, there is a service type that connects stations of two different provinces. This type is entitled coach. Because the platform of TravelVU does not offer this type and cannot add this, we recommended users to give car label to segments by coach. In this sense, the car covers all automobile types except from bus. Users generally easily discriminated coach from bus. However, there is a type that was frequently misclassified as bus. This is a specific-passenger-dedicated service that is for carrying staff and workers of particular companies only. It ran two or four trips per day between the center of Hanoi and the company located at suburb. In the center, it picked up passengers at several points close to bus stops before running directly to the destinations. We did not consider it as bus because it limits passenger types and passes almost all bus stops. Users were recommended to label these segments as car. By visualizing all bus segments on the map, we realized that many segments reported by bus of several persons were actually passenger-specific services. For those persons, we requested them to re-validate the use of bus. The results were as our expectations. They were not by bus and we changed their mode from bus to car. This systematic error occurring was due to fail to abide by instructions.

(a) **(b)**

Figure 3 - 8. (a) Bus services and (b) services dedicated for workers and staff
(Source: http://baobacgiang.com.vn/bg/kinh-te/211825/dua-dich-vu-don-tra-cong-nhan-vao-nen-nep.html; https://tintucvietnam.vn/lo-trinh-xe-buyt-ha-noi-nam-2018-41010)

**** Checking business segment:***

In the following table, all trips that had confirmed as work activities before are presented along with their dates, start times, durations and locations on the map. Please check and verify them.

Thank you for your kind support!

Dưới đây là địa điểm của các chuyến đi được Ông/Bà báo cáo là *chuyến đi làm*. Xin hãy kiểm tra lại giúp chúng tôi về tính xác thực của thông tin dựa trên bản đồ và thông tin thời gian của hành động.

Trân trọng cám ơn

TT Code	Ngày Date	Giờ Start time	Thời lượng – phút Duration (mins)	Mục đích Purpose		
				Đi làm Work	Đi làm bên ngoài Business	Không nhớ Not remember
1	19/03/2019	09:15	560	✓		
2	20/03/2019	09:35	8		✓	
3	20/03/2019	10:13	9		✓	
4	20/03/2019	10:23	107		✓	
5	20/03/2019	12:40	8		✓	
6	20/03/2019	13:01	494	✓		
7	21/03/2019	09:09	654	✓		
8	22/03/2019	11:36	87	✓		
9	22/03/2019	14:56	179	✓		
10	23/03/2019	09:08	172	✓		
11	23/03/2019	12:02	21		✓	
12	23/03/2019	12:27	311	✓		
13	25/03/2019	09:36	181	✓		
14	25/03/2019	12:40	18		✓	
15	25/03/2019	13:00	241	✓		
16	26/03/2019	12:50	73		✓	
17	26/03/2019	14:06	16		✓	

Figure 3 - 9. Example of re-confirming work and business activities

17 work activities had been reported during 7 days. After re-checking activities with their code, start time, duration profiles, the user relabeled 8 activities, from work to business. Finally, all work locations were spatially in line each other whereas business locations were fairly scattered. The user re-declared that his job were dynamic. Accordingly, we believed more in the accuracy of re-confirmations

In case a participant did a business outside his/her workplace, this should be labeled as a business segment. However, there was a possibility that users failed to follow this note. Instead, they confirmed business activities as work ones. To detect this problem, we considered persons who had dynamic jobs. All their work locations were visualized. For those whose spatial distribution of work were not consistent, there

would be likelihood that they either validated business purpose as work purpose or simply gave wrong answer about the dynamic status of his/her job (see Appendix 2, question 6). In the post-survey, those who were suspected were requested to do more two tasks. The first is checking a map with distribution of their work locations numbered along with their corresponding time profiles, which could help them to recall the activities. They next saw and chose one among three responses, that is, 1: work activity, 2: business activity, 3: not remember/prefer to reveal. The second question was their job types being whether dynamic or not. Before they answered two questions, we re-emphasized the difference between work and business activities. The results showed that all persons asked re-confirmed their dynamic jobs. Many work activities were changed into business activities.

3.3. SUMMARY

This chapter describes the datasets used in this thesis. The first is the sets collected in Rhone-Alpes, France. One of them was telephone-based data of the regional household travel survey (i.e. EDR-RA) whilst the rest was GPS points of a sub-sample of EDR-RA (i.e. TOMOS) gathered by dedicated devices. The EDR-RA data were at large scale and thus adequate for evaluating the mode-based travel patterns translated from TOMOS data that were without the corresponding ground truth.

The second data set was collected in Hanoi by smartphone and attached with travel diaries provided by 63 valid users. Generally, participants assessed positively the use of smartphones in mobility survey; however, they grumbled about the significant burden of validation process, relatively high battery consumption and technological errors. Due to privacy concern, the home addresses of some users were omitted. Visualization of data highlighted some issues related to the precision of ground truth information. They were corrected by post-surveys. Therefore, the Hanoi data would be eligible for developing algorithms of imputing both travel mode and trip purpose.

Chapter 4: MODE DETECTION FROM GPS DATA WITHOUT GROUND TRUTH

4.1. INTRODUCTION

To evaluate the potential of GPS in terms of interpreting better travel patterns than conventional surveys, numerous GPS-based surveys have been conducted at both national and regional scales (Shen and Stopher, 2014b). In case the major objective is developing inference methods, ground truth data are often collected; however, ground truth may not be necessary if the survey's data is for comparing with those of conventional household travel surveys. In this chapter, we attempt to translate TOMOS data that were collected in Rhone-Alpes, France without respective ground truth. There are two main challenges we have to address.

- The first is choosing and enhancing an appropriate method to detect modes.
- The second is evaluating the mode prediction results in case of unavailability of ground truth.

Finding out answers to the mentioned-above problems can help to clarify more remaining issues preventing GPS from being more useful for travel behavior analysis. As for the chapter structure, the following reviews the literature to seek suitable methods for detecting mode in case no ground truth data are gathered. Section 4.3 is for filtering datasets in both EDR-RA and TOMOS to make them comparable spatially and temporally. Detailed description of the adopted method is the content of Section 4.4. Subsequently, results and discussions are presented. The last section concludes the chapter and points out unresolved matters.

4.2. MODE DETECTION AND GROUND TRUTH

The existing literature witnesses the great attention to the travel mode detection field with the introduction of numerous methods that can be classified into three main groups, namely rule-based, probability-based and machine learning (Gong et al., 2014). The pioneers (Bohte and Maat, 2009; Stopher et al., 2005) proposed rules to classify modes. Simple and understandable criteria are determined from practice. However, their performances are poor once confronting the ambiguous behaviors of modes in real life such as movement in congested roads. Consequently, deterministic methods are not preferred, mainly used to discriminate between motorized and non-motorized segments (Safi et al., 2016). The current preference is donated to machine learning solutions because they are able to differentiate modes at very high accuracy by various variables (Feng and Timmermans, 2016; Semanjski et al., 2017; Shafique and Hato, 2015; Xiao et al., 2015c, 2015b). However, learning-based methods require intensive computation and great number of data to overcome the problem of overfitting, not to mention lack of interpretability. In addition, supervised machine learning needs ground truth to train models according to minimizing cost function. Mode detection results by machine learning are useful for reconstructing travel diaries only when ground truth is reliable. If ground truth is far from the actual travel history, a mode identification model may fail to give precise information on how a person traveled but still deliver a very good performance in terms of mathematics. Unfortunately, ground truth is collected by traditional techniques like website, paper and pen, telephone (Bohte and Maat, 2009; Semanjski et al., 2017; P. Stopher et al., 2008a; Xiao et al., 2015b) thus it is subject to the under-reporting problem. In this sense, machine learning and ground truth seem not to be a good answer to the biggest question in Transportation Science about how to make a move

toward the replacement of conventional travel household surveys with GPS (Auld et al., 2009; Wolf et al., 2014a).

Fuzzy logic-based algorithms function far better than criterion-based heuristics thanks to the inclusion of the overlaps between features (e.g. speed and acceleration) to produce simultaneously the probabilities that pertain to the fall of a segment into each class. The first application was introduced by Tsui and Shalaby, (2006) before deployed in (Das and Winter, 2018; Nitsche et al., 2014; Rasmussen et al., 2015; Schuessler and Axhausen, 2009). Interestingly, different from other authors, Schuessler and Axhausen, (2009) validate their mode prediction outcome by the Swiss National Travel Household Survey data.

Regarding features employed to detect modes, movement-related variables such as speed and acceleration profiles are the most common for any method types (Dabiri and Heaslip, 2018; Feng and Timmermans, 2016; Schuessler and Axhausen, 2009; P. Stopher et al., 2008a). Even, with acceleration solely, the accuracy rate of prediction can achieve reasonable levels (Feng and Timmermans, 2013; Shafique and Hato, 2015). However, due to similar speed and acceleration characteristics between modes, especially bus and car (Gong et al., 2012; Xiao et al., 2015b), GIS information should be in use to identify more correctly public transport segments (Biljecki et al., 2013; Gong et al., 2012; Nour et al., 2016; Rasmussen et al., 2015). Recently, authors adopted new features like heading change rate (Zheng et al., 2010), jerk (Dabiri and Heaslip, 2018) and the frequency of stopping in the close vicinity of bus stops (Nour et al., 2016).

The afore-said synthesis emphasizes that probabilistic and machine learning approaches are capable of detecting well travel modes from GPS data and external sources. The former is a better choice to test the potential of GPS-based survey thanks to being able to avoid the entire reliance on ground truth collection. Validating mode detection in GPS-based surveys by data of representative large-scale samples is a good choice in case of unavailability of ground truth.

4.3. DATA PREPARATION

GPS data in TOMOS and CATI data in EDR-RA can be compared because GPS carriers in TOMOS were interviewees in EDR-RA. However, due to the difference in recruitment criteria along with spatial and temporal distribution of sample, filtering data of both EDR-RA and TOMOS to make them comparable is essential.

- *Spatial difference in residence:*

Participants in TOMOS lived in Lyon city and its outskirts (see Figure 4-1) whilst those questioned in EDR-RA lived in the large area Rhone-Alpes. In EDR-RA data, we removed all persons whose residences were not in Lyon, Bron and Villeurbanne.

- *Difference in age:*

Participants in TOMOS were over 17 years old whilst EDR-RA sample included participants over 10 years old. Thus, those in EDR-RA who were younger than 18 years old were out of scope.

- *Difference in time:*

EDR-RA spanned weekdays only whilst TOMOS covered all days of week. Therefore, weekend data in TOMOS were eliminated.

GPS data of the first and the last days of each person were disregarded due to failing to cover all trips of each among these days. Since most of first and last days were at weekends that are convenient time

for participants to travel and receive/return devices, this rule reduced more the marginal number of surveyed days.

- *Signal loss and immobility*

There were days of both mobility and immobility in EDR-RA whilst TOMOS included days with and without positioning data. It should be noted that unavailability of GPS data on the one hand was the fact that GPS holders did not move. On the other hand, it possibly came from technical problems or forgetfulness of taking devices or flat battery. Distinguishing between immobility and mentioned-above issues was impossible. Hence, all days without mobility in EDR-RA and those without GPS data in TOMOS were neglected.

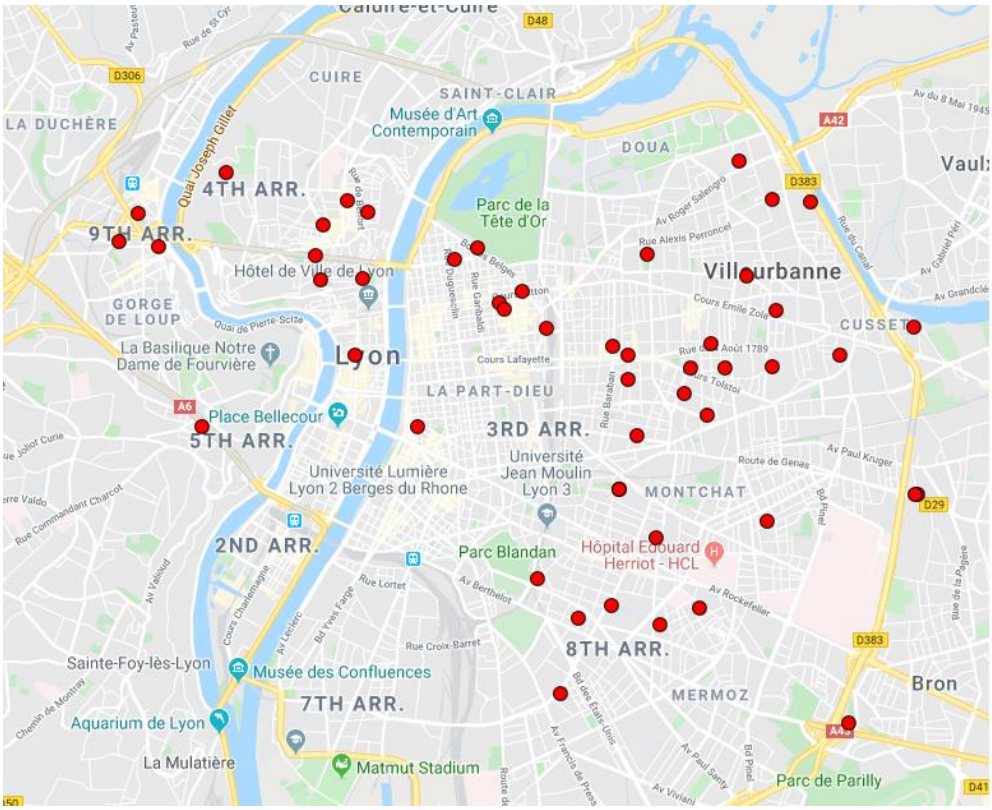


Figure 4 - 1. Residence distribution (red points) of TOMOS's sample

Finally, the sample of TOMOS consisted of 311-day database of 80 persons whilst EDR-RA was comprised of 2859 persons and 2859 days also. Courses of both surveys had balanced distributions of weekday types in week (i.e. Monday, Tuesday and so on).

Figure 4-2 describes a breakdown of participants in the EDR-RA and TOMOS by gender, working/studying status, age and number of car in household. As regards the two first ones, the shares of groups by each variable in both surveys were similar. There was a difference in age groups. TOMOS had a larger percentage of the oldest at 36.3% compared with 31.5% in EDR-RA. The opposite pattern was seen in two younger groups but with smaller difference levels. As for car ownership, those whose households did not possess car in EDR-RA made up at 22.5%. By contrast, TOMOS's figure was only 10%. When it comes to the group of persons coming from households with at least two cars, the proportion in TOMOS (40%) was much higher than that in EDR-RA (24.5%). The group of persons with one car accounted for the largest percentage in both surveys with the similar rates of around 50%.

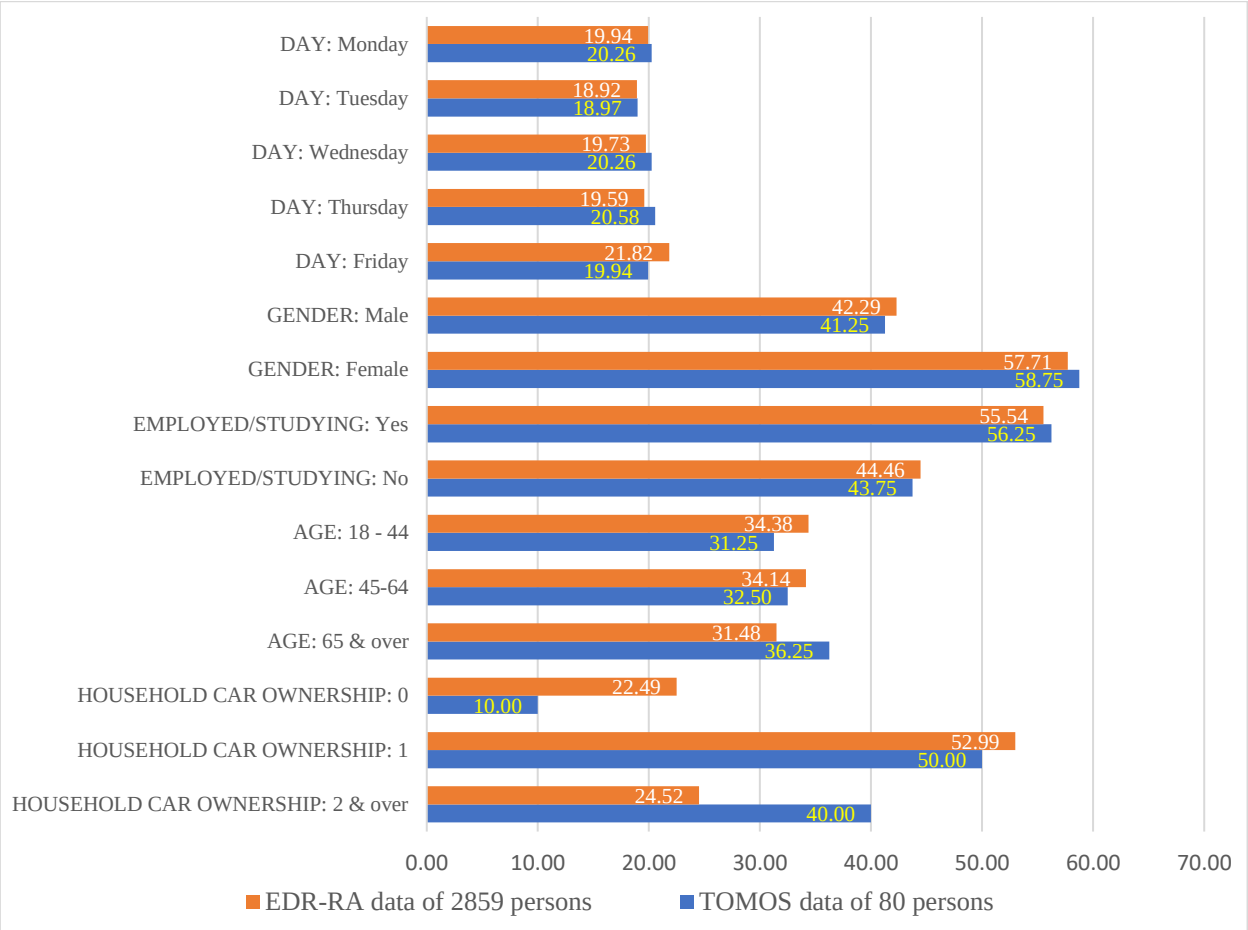


Figure 4 - 2. Distribution of demographics variables in EDR-RA and TOMOS data (unit: %)

Assuming that EDR-RA is sufficiently representative to travel pattern of TOMOS sample, we then tested the associations of mode-based trip rates with age, car ownership by analysis of variance (see Table 4-1 for results).

With the 95% level of confidence, the oldest group walked significantly more but traveled by bike significantly less than younger groups. For car, three groups have significant difference in trip rate. The 45-64-year old group travelled by car the most at 1.68 on average whilst the oldest did the least at only 1.11 trips per day. The use of public transport by different groups shows significant difference. An older group utilized public transport less than the younger group. The trip rates by public transport of the groups whose ages are 18-24, 45-64 and over 64 are 1.04, 0.67 and 0.50, respectively. The total trip rates per day of the youngest and the middle-age group are close together, that is, 4.33 and 4.25. The figure for the oldest group at 3.93 is significantly smaller than those of the two younger groups.

Regarding the car ownership, persons from households possessing more cars walked significantly less than those from households having fewer cars. Whilst non-car households' persons traveled 2.2 walk trips per day, over two-car households' participants walked only 1.47 trips. The difference in using bike is not significant. For both car use and public transport use, trips rates of car ownership-based groups are significantly different. Persons with no car traveled by this mode only 0.21 trips whilst persons from over one-car household made 2.35 trips. By contrast, the rates of public transport use for these two groups are 1.28 and 0.42, respectively. The daily trip rate of non-car ownership group and groups with available car in households are different significantly. However, the difference between the groups having one and more than one car is insignificant.

Table 4 - 1. Testing differences in trip rate of modes according to age and car ownership in EDR-RA data

Variables/Groups	Walk	Bike	Car	PT	Total
	TRIP RATE				
AGE					
18-44	1.69	0.19	1.41	1.04	4.33
45-64	1.76	0.14	1.68	0.67	4.25
over 64	2.27	0.05	1.11	0.50	3.93
CAR IN HOUSEHOLD					
0	2.20	0.14	0.21	1.28	3.82
1	1.97	0.13	1.48	0.66	4.25
2 and over	1.47	0.11	2.35	0.42	4.36
	DIFFERENCE IN MEANS AND ITS SIGNIFICANT LEVEL				
AGE					
18-44 vs 45-64	-0.07258	0.04782	-0.26731*	0.36962*	0.07755
18-44 vs over 64	-0.58037*	0.14033*	0.30515*	0.53116*	0.39627*
45-64 vs over 64	-0.50779*	0.0925*	0.57246*	0.16154*	0.31872*
CAR IN HOUSEHOLD					
0_car vs 1_car	0.23363*	0.00862	-1.27765*	0.61191*	-0.4235*
0_car vs 2+_car	0.72844*	0.02727	-2.14266*	0.85159*	-0.53535*
1_car vs 2+_car	0.49481*	0.01866	-0.86501*	0.23969*	-0.11186

**: Significant at the 0.05 level*
Method: Analysis of variance and compare means by Tukey's test

Subsequently, based on the combination between age and household car ownership, expected trip rates by walking, biking, car and public transport in TOMOS were estimated. Table 4-2 describes how to compute expected car trip rate. For each row, an expected rate (column 8) is the multiplication between *means* calculated from EDR-RA (column 4) and *the share* of the group characterized by age (column 1) and household car ownership (column 2). Thus, an intersection between column 8 and each row presents the contribution of each responsive group to the total car trip rate (1.69) that is reported in bold in the last row. By the same way, we estimated the expected trip rates for walk (1.83), bike (0.12) and public transport (0.61). The total expected trip rate was 4.24.

Based on the comparison between trip rates of EDR-RA and expected trip rates of TOMOS, TOMOS sample would be more mobile than EDR-RA sample. Participants in TOMOS were expected to walk slightly less, utilized public transport at lower level but travelled by car considerably more than those in EDR-RA.

Table 4 - 2. Estimating expected trip rate by car in TOMOS

Age	Household car number	EDR-RA data		Share in sample		Trip rate in EDR-RA	EXPECTED trip rate in TOMOS
		Frequency	*Means*	*EDR-RA*	*TOMOS*		
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
18 - 44	0	246	0.20	8.60%	5.00%	0.02	0.01
18 - 44	1	477	1.39	16.68%	8.75%	0.23	0.12
18 - 44	2 & over	260	2.60	9.09%	17.50%	0.24	0.45
45 - 64	0	141	0.14	4.93%	2.50%	0.01	0.00
45 - 64	1	541	1.67	18.92%	18.75%	0.32	0.31
45 - 64	2 & over	294	2.44	10.28%	11.25%	0.25	0.27
65&over	0	256	0.25	8.95%	2.50%	0.02	0.01
65&over	1	497	1.37	17.38%	22.50%	0.24	0.31
65&over	2 & over	147	1.73	5.14%	11.25%	0.09	0.19
Total trip rate						**1.41**	**1.69**

*Column 7 = Column 4 * Column 5*
*Column 8 = Column 4 * Column 6*

4.4. MODE DETECTION METHOD

4.4.1. Terminology and principles to determine trip mode

A trip is defined as a one-way course of travel to do an activity; therefore, a trip is either single-mode or multimodal. Trip mode is the main mode that pertains to mode of a segment.

In EDR-RA survey, although requested to provide detailed itineraries with modes of all stages, almost all participants declared the main mode of each trip only. For example, a train trip was reported without stating mode(s) to access to the station and mode(s) to arrive at the destination. In the existing literature, the main mode of a trip has been used to validate the mode detection results (Patterson and Fitzsimmons, 2016; Schuessler and Axhausen, 2009). Here, we applied following principles to assign a (main) mode to a trip from the mode classification result(s) of its segment(s).

- First, if a trip has at least one segment labelled as transit (i.e. metro or bus/tram), it is classified as public transport.

- Second, if a trip omits to the first rule and includes at least one car stage, it is classified as car.

- Third, if a trip omits to both the first and second rules and includes at least one bicycle stage, it is classified as bicycle.

- Last, if a trip includes walk stages only, it is classified as walk.

Simply, the main mode is done using the heaviest vehicle.

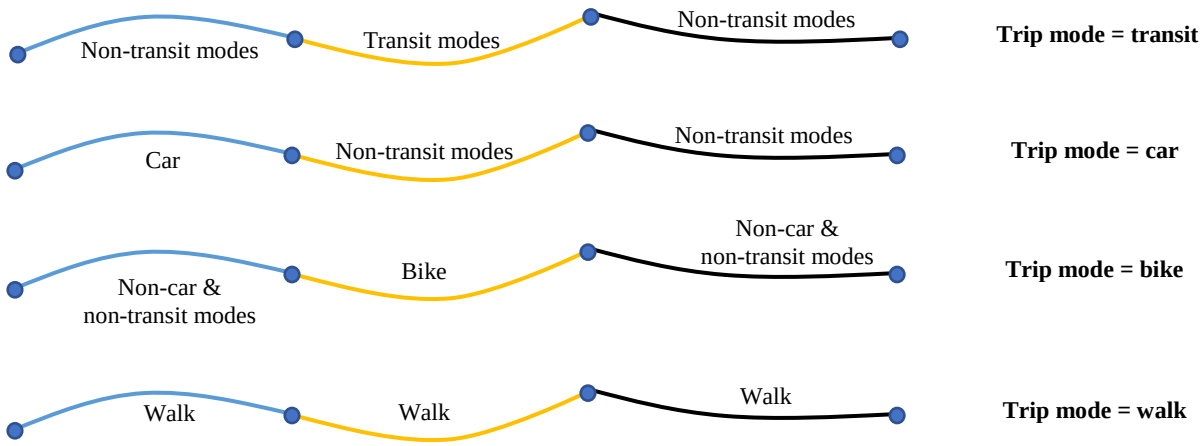


Figure 4 - 3. Graphical descriptions of trip mode rules

It is notable that the four mentioned-above principles were applied for both EDR-RA and TOMOS data to detect modes of trips.

4.4.2. Data filtering

Instantaneous speed between two consecutive points is one of main indicators to remove jumping points that usually having very high/unrealistic speed levels. A TOMOS datum consists of longitude, latitude and acquisition time only. To estimate speed, first we computed the distance between two points. Based on the assumption that the figure of the Earth is oblate spheroid, Vincenty (1975) proposed a method widely applied. Here, a distance between two points was gauged according to Vincenty's formula by Python with the support of the *GeoPy* library. After distance was estimated, an instantaneous speed between two consecutive points was computed by the division of distance by the corresponding time lag.

Raw data were filtered by examining three criteria.

- First, all points without coordinates and with same timestamps or coordinates as the previous' ones were not eligible for being accepted.

- Second, the maximum instantaneous speed should be less than 350Km/h that is possibly generated by high-speed train.

- Last, all points outside France were abandoned.

In TOMOS, one participant visited Italian cities during the survey period and days of her tourism were excluded. After being filtered, fixes were smoothed by an implementation of Gauss Kernel technique like (Schuessler and Axhausen, 2009).

4.4.3. Segmentation

Segmentation is associated to two steps. The first is splitting trajectories into trips by detecting trip ends. Then each trip continues to be divided into segments. Stops and gaps are the key to identifying trip ends and transfer points. We applied trip/segment-breaking rules, as follows:

Table 4 - 3. Criteria to filter and segment data for mode inference

Indicator	Minimum	Maximum	Explanation
Data filtering and smoothing			
Available both longitude and latitude	-	-	-
Speed (km/h)	> 0	350	-
Latitude (decimal format)	42.500	51.100	French area
Longitude (decimal format)	-4.900	8.500	
Gauss kernel smoothing	-	-	For all filtered points
Segmentation			
Speed of non-movement status (km/h)	0	1.4	Nearly zero speed
Dwell time or time gap (s)	180	-	Trip end detection
Dwell time or time gap (s)	30	180	Stop detection
Kalman smoothing	-	-	For points in each segment lasting over a minute and with more than 10 points

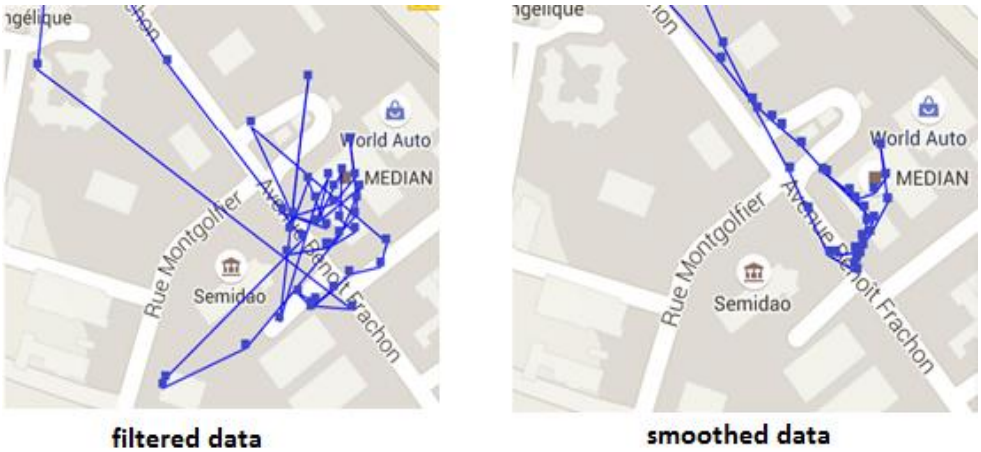


Figure 4 - 4. Example of filtered and smoothed data

A trip will terminate if (1) there is a signal loss gap of at least 180 seconds or (2) there is a stationary status during at least 180 seconds (i.e. dwell time). The non-movement is defined as the speed between two consecutive points below 1.4km/h (approximately 0.4m/s). The 180-second threshold was used by (Bohte and Maat, 2009; Patterson and Fitzsimmons, 2016) before.

In each trip, its segments were detected by identifying stops that were possibly transition points. Similar to trip end detection, two rules related to time gap and dwell time were employed albeit with the threshold of 30 seconds.

Both trips and segments whose durations were smaller than 60 seconds were neglected. Each segment lasting over a minute and including over 10 points was smoothed by the Kalman filter. Trips with high-speed rail segment(s) that has 95th percentile of speed of over 40m/s were disregarded.

4.4.4. Metro segment detection

Generally, most metro segments are underground, thus coincide with signal loss. Performing spatial analysis by the GIS data of stations has a high chance of success (Biljecki et al., 2013; Gong et al., 2012; Rasmussen et al., 2015). There are some challenges of examining the proximity of the first and the last points of signal loss period to metro stations. One of them is involved in choosing the distance threshold. Because a station usually has more than one entrance, so it is better to use the coordinate of every entrance like (Gong et al., 2012) rather than a center point of the station. However, a station can be integrated into large places like department store, commercial center or square. The obstructions of these places prevent devices from receiving signals from satellites. The signals occur again whilst a person leaves from them and at that moment, he/she is fairly far from the station. In this situation, the use of entrance positions and the center point of a station generate similar performances.

We considered the distance to the center point representative of a station location. All stations were classified into two types. If a station is integrated into large areas like commercial centres and rail stations, the threshold of 250m was determined. The threshold of 150m was applied for others. These boundaries were 50m higher than the case of using location of metro entrances (Gong et al., 2012); however, smaller than 300m in (Patterson and Fitzsimmons, 2016).

Ideally, all parts of each metro segment are underground, leading all movement time to correspond to signal interruption (*case a in Figure 4-5*). In this case, the first and the last point of signal gap should be close to stations. However, some stations are on the ground, making only some parts of a metro segment be underground. Therefore, it is acceptable that a gap with either its start point or its end point adjacent to a station belongs to a metro segment (*case b and c in Figure 4-5*).

Another challenge is seen in *case d in Figure 4-5* where there is no signal interruption in a buffer of any stations. To deal with this, the rule is that if the start point of the previous segment or the end point of the next segment in a buffer of a station, the signal loss stage together with the previous and the next are merged and are assigned the metro label.

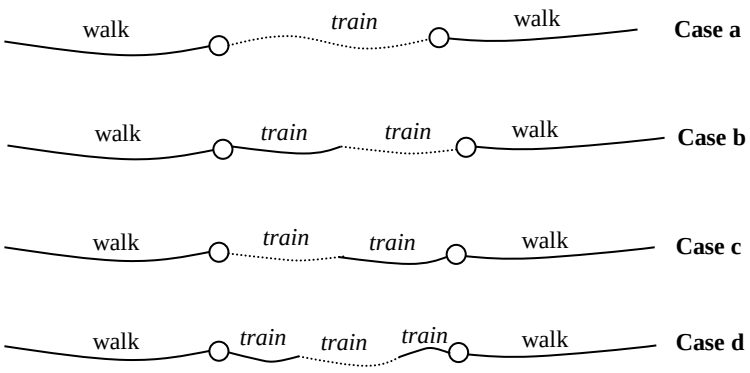


Figure 4 - 5. Descriptions of metro trip cases

There is a four-metro-route network connects stations in Lyon city and its suburbs. The majority of stations are underground; hence, the rate of *case d* compared with those of *case a, b and c* was expected to be insignificant. In the metro network, the time to travel between two consecutive stations whose distance between them is around 500m is roughly one min. Hence, a signal gap belongs to a metro segment if it lasts over 60 seconds and corresponds to a distance at more than 300m.

4.4.5. Fuzzy logic-based algorithm

A fuzzy logic-based model was designed to detect transportation modes whose movement was recorded by GPS. Previously, fuzzy inference has managed to differentiate the ambiguous characteristics of modes with and without prompted recall surveys (Das and Winter, 2018; Rasmussen et al., 2015; Schuessler and Axhausen, 2009; Tsui and Shalaby, 2006). In this part, after discriminating metro segments, we aimed at classifying walk, bike, car and bus/tram segments by a model based on the fuzzy logic theory. Bus and tram had similar operational characteristics and thus were grouped. Whilst buses run on roads, trams operate on tracks that are in parallel with and close to roads. Tram stations and bus stops are adjacent each other to provide convenient access and transfer.

Fuzzy logic theory, frequently termed *fuzzy logic*, is a type of reasoning theory making decisions on the basis of ambiguous instead of crisp (precise) data. It was introduced by Prof. Zadeh of the University of California in 1965 (Zadeh, 1965). The theory began receiving a great scientific attention since Mamdani applied it to control an automatic steam engine (Mamdani and Assilian, 1975). From then on, it has been extended and applied widely in many fields like automatic control, measuring service quality, academic education, hospitals (Sugeno, 1985; Zadeh, 1973; Zadeh and Aliev, 2018; Zedeh, 1989).

The fuzzy logic-based mode prediction model proposed is comprised of five steps.

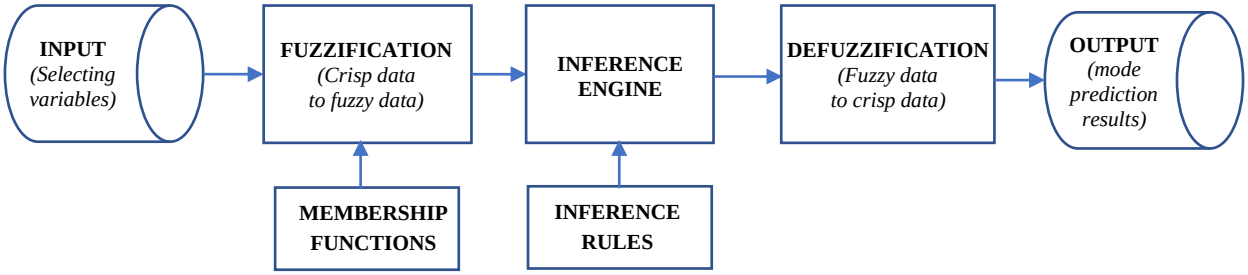


Figure 4 - 6. Flow of fuzzy logic-based mode inference model

4.4.5.1. Variable selection

The first and the vital procedure of any mode detection model is choosing appropriate variables that represent uniqueness of modes' behaviors. 95% maximum speed and 95% maximum acceleration are often the most useful indicators to discriminate motorized modes from non-motorized ones. Whilst car and bus can run at (very) high speed and need short time to achieve this speed with high acceleration, walk cannot do it.

However, the confusion related to speed and acceleration profiles raises in practical life. Bike in different cases can show similar speed levels either to car or to walk. Once running on cycle lanes, bicyclists can move at 30-40 km/h, raising the ambiguity between bike and motorized modes in cities giving prioritized infrastructure to biking. If biking is for entertainment or going sightseeing, bike's speed profiles may be close to those of walking. The ambiguity between car and bike, even walk can grow up if a participant drives slowly to find parking places.

Movement-specific indicators of car and bus generally are similar in urban areas. Buses are offered dedicated lanes, thus possibly run at 40-50 km/h. In many cities, to reduce bus trip duration, the bus's acceleration ability is fostered.

In case nearly maximum speed and acceleration indicators cannot help much to classify modes, the use of average speed can be a handy complement. For example, a person can walk considerably faster to reach his/her friends/colleagues or crossing an intersection before the green light switch to red. However, (s)he cannot maintain high speed levels on a long distance. Similarly, although a bus can run at speed akin to that of car but it has to decelerate to be stable at bus stops, thus the average speed of bus is expected lower than that of car.

The most challenging situation may be on congested roads where movement of motorized and bike is comparable with movement of walk.

Mentioned-above analyses demonstrate that a single variable is not enough to detect modes. The combination of 95% maximum speed, 95% maximum acceleration and average speed were expected to be useful but may be insufficient. If data are collected densely, every second for example, indicators can be more typical for modes. Yet, TOMOS data were collected at low sampling rate of 10 seconds. During the 10-second interval, speed of vehicles may change 10 times compared with speed observed every second. For example, a car achieves a very high-speed level at the first second but then slow down and keeps moving slowly until the 10th second. Consequently, we will fail to keep the distinctly high speed of the segment. To enhance the power of the mode classifier, we decided to use heading change rate that is regarded as the frequency of changing heading direction within a unit distance. Heading (or bearing also) is the angle between the ray connecting two consecutive points and a reference ray (e.g. true north). The heading change rate is useful for mode detection because it is relatively independent of traffic conditions compared with acceleration and speed and typical for specific modes. Cars and bus move alongside roads, consequently their levels of direction change are small. Trams are limited by tracks, thus they alter their directions at low levels also. On the contrary, bicycle and walking are so flexible that their heading directions fluctuate, even in case of intending moving straight (Zheng et al., 2010). The heading rate is computed by using coordinates of three consecutive points (Figure 4-7) by the following equations:

$$y = sine\ [(p_2(lon) - p_1(lon)] \times cosine\ [p_2(lat)] \quad (4\text{-}1)$$

$$x = cosine\ [p_1(lat)] \times sine\ [p_2(lat)] - sine\ [p_1(lat)] \times cosine\ [p_2(lat)] \times cosine\ [p_2(lon) - p_1(lon)] \quad (4\text{-}2)$$

$$Heading_{(P_1)} = artangent\ (y, x) \quad (4\text{-}3)$$

$$Heading\ rate = \left|Heading_{P_2} - Heading_{P_1}\right| \quad (4\text{-}4)$$

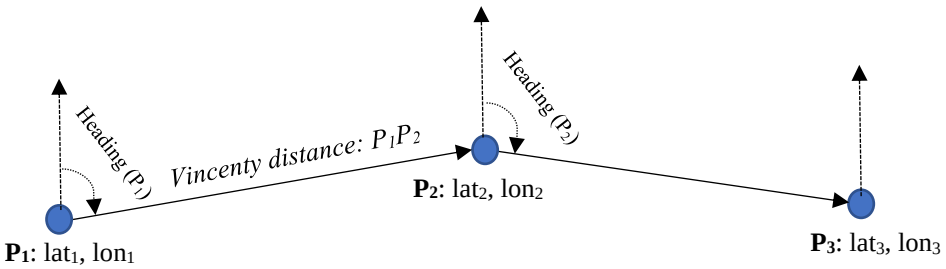
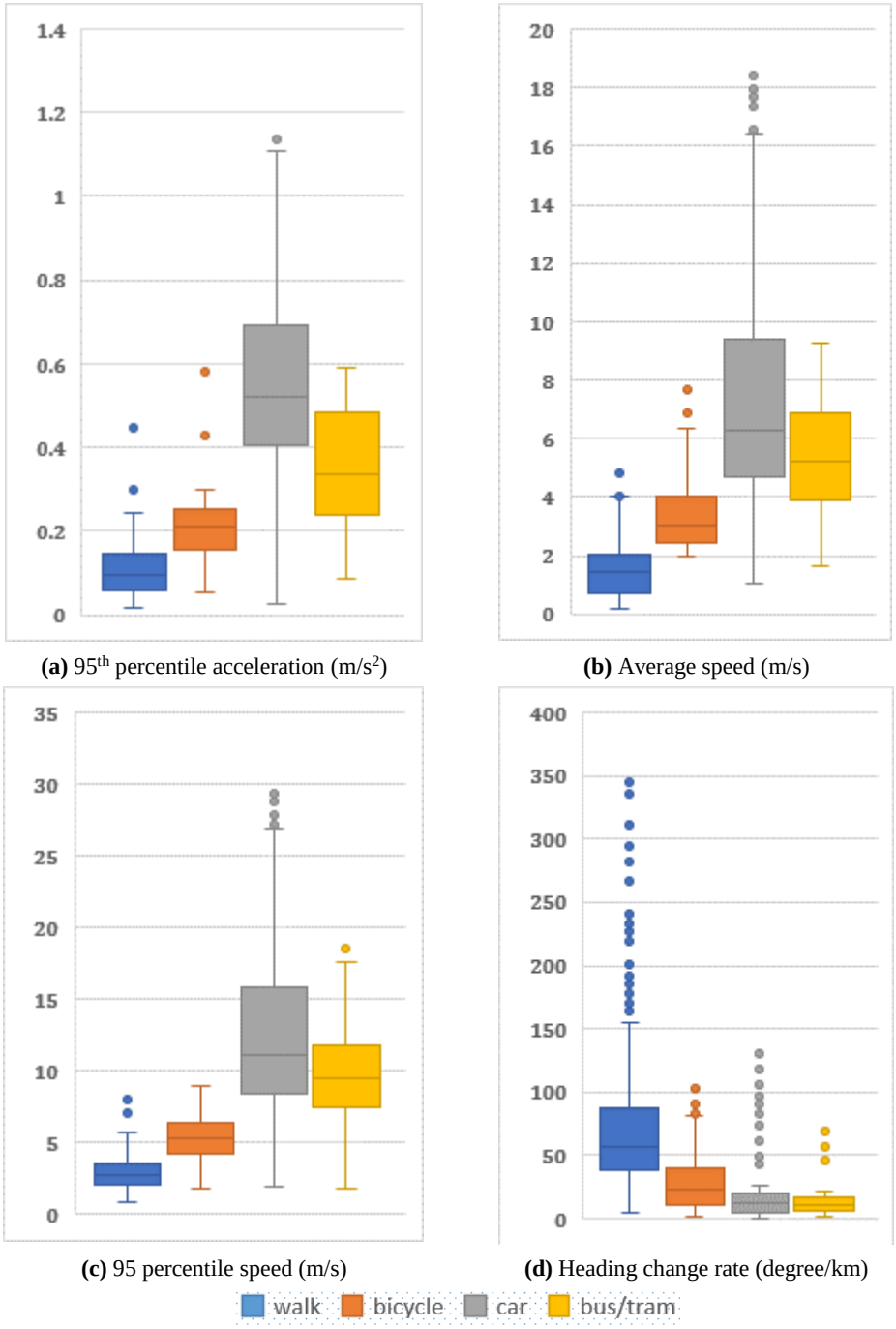


Figure 4 - 7. Description of heading between consecutive points

Where *sine*, *cosine*, *arctangent* are trigonometric functions. *Lon* and *lat* stand for the point's longitude and latitude, respectively. They should be in radians. *Heading rate* in equation 4-4 is converted to degree.



(a) 95th percentile acceleration (m/s^2) **(b)** Average speed (m/s)

(c) 95 percentile speed (m/s) **(d)** Heading change rate (degree/km)

Figure 4 - 8. Distribution of variables based on data of volunteers in pilot tests

Before TOMOS, pilot phases were implemented in 2014 to test devices by our 24 colleagues. Each provided data of one or several days. They were used for building and calibrating the mode detection model. Filtering and splitting data were conducted according to approaches presented above. Next, we considered distributions of four variables.

As can be seen in Figure 4-8a, 95% max acceleration of car has the broadest range. Approximately 75% of walk segments have the 95% max acceleration levels akin to those of bike. Bike's distribution of this variable were covered completely by car's one and almost completely by bus. All 95th percentile acceleration values of bus were achievable by car.

Figure 4-8b and 4-8c reveal that mode-based distributions of both average speed and 95% max speed are similar. Speed values of bus and car are obviously ambiguous. Car is the most confused mode because it can show similar speed levels to those of other modes. Walk differs from other by very low speed

levels, under 2 m/s for average speed and less than 3.5 m/s for 95th percentile speed for example. Levels at over 9 m/s for average speed or/and over 19 m/s for 95% max speed are unique for car segments.

Figure 4-7c emphasizes the difference in heading change rates of modes. In line with above analysis, walk has the widest range and generally has higher rates than bus and car. Heading change rate would be informative for discriminating walk from bike.

For better seeing the potential of combining 95th percentile speed, 95th percentile acceleration, average speed and heading change rate, we visualize four variables of some walk, bike, car, bus/tram segments on figure 4-9. It can be seen that if considering four variables, the opportunity for correctly identifying modes is relatively high.

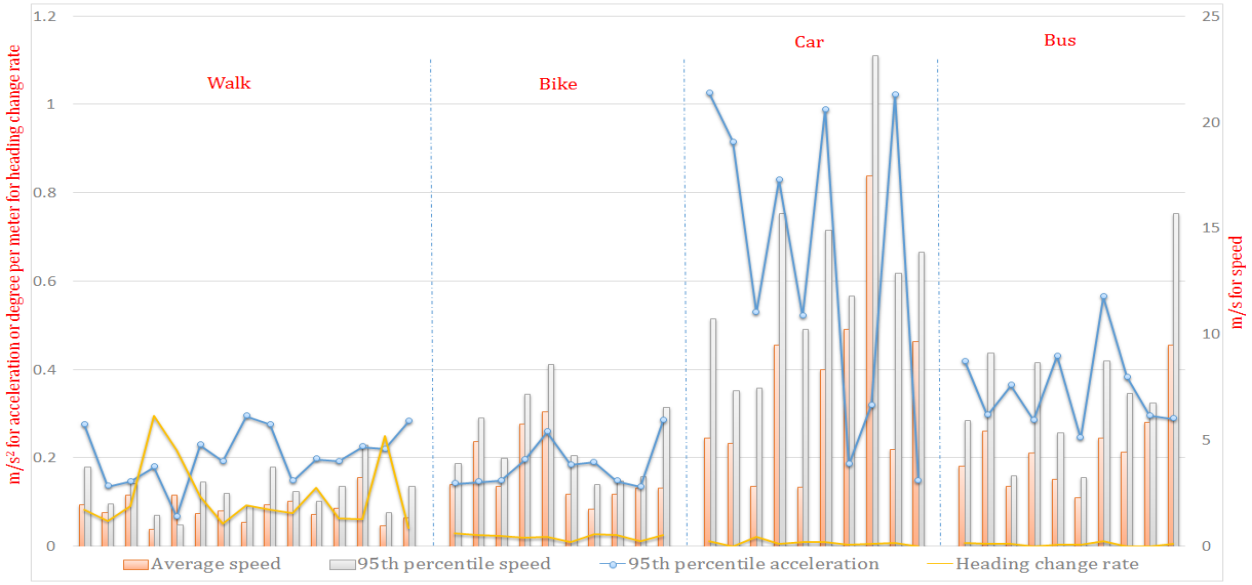


Figure 4 - 9. Example of distributions of variables by modes

4.4.5.2. Fuzzification

Fuzification is the process of converting crisp input variables' data into fuzzy values by Membership Functions (MF). A fuzzy value is involved in a membership degree of an observation to a linguistic variable that is defined vaguely by an expression like "high" or "medium" or "low". To specific, a linguistic variable takes words as values that can be considered as labels of the linguistic variable. For each label, it has a number of values (e.g. *low*, *medium*, *high*) that are called its term set.

Suppose that X is a vector of input crispy variables, $X = \{x_1, x_2, x_3, x_4\}$ with x_1, x_2, x_3, x_4 being 95% max acceleration, average speed, 95% max speed and heading change rate, respectively.

A variable x_i has its own value set. In the fuzzification step, this value set is re-defined by a fuzzy set characterized by a term set $T(x_i) = \{T_{x_i^1}, \dots, T_{x_i^j}, \dots, T_{x_i^k}\}$ and a MF set $M(x_i) = \{\mu_{x_i^1}, \dots, \mu_{x_i^j}, \dots, \mu_{x_i^k}\}$. The term set $T(x_i)$ include k terms. Each term $T_{x_i^j}$ is defined by a MF $\mu_{x_i^j}$. Where, $\mu_{x_i^j}$ is for estimating the membership degree that quantifies the grade of membership of an observation to the $T_{x_i^j}$. There are a number of MF types, that is, triangular, trapezoidal, gaussian, bell-shaped, s-shaped, z-shaped functions. Trapezoidal MF (see Figure 4-10) is the widely used in previous mode detection studies (Rasmussen et al., 2015; Schuessler and Axhausen, 2009; Tsui and Shalaby, 2006).

To give an example, the fuzzified variable of 95th percentile speed (i.e. x_3) can take on 3 linguistic values, namely *low* ($T_{x_3^1}$), *medium* ($T_{x_3^2}$), *high* ($T_{x_3^3}$). Each term is defined by a trapezoidal MF. In Figure

4-10, the horizontal axis represents the input variable x_3, and the vertical axis defines the corresponding membership value *μ(x)* of the input variable x_3. A MF estimates the membership degree of the observation to the corresponding term. If the degree is 0, the observation does not belong to the fuzzy set. If it is 1, the observation is fully a member of the fuzzy set. If it is between 0 and 1, the observation partially belongs to the fuzzy set.

It is worth noting that ranges of membership functions overlap, which enables the model to examine and address flexibly ambiguous profiles of input variables between modes rather than in a deterministic way.

We created MFs for four variables presented in Figure 4-11. To illustrate how to estimate membership degrees, let's see a segment with 95th percentile of acceleration: 0.17 m/s^2, average speed: 2.3 m/s, 95th percentile speed: 4 m/s, heading change rate: 25 degree/km.

For the 95th percentile acceleration of 0.17 m/s^2, the vertical line from the point of 0.17 intersects MF of term *low* at the point with coordinate (0.17, 0.5), thus the degree of the fact that 0.17 m/s^2 belongs to *low* percentile acceleration is 0.5. By the same way, degree of the fact 0.17 m/s^2 belongs to *medium* percentile acceleration is approximately 0.25 whilst 0.17 m/s^2 is not a member of the *high* acceleration set (see Figure 4-12).

$$\mu(x) = \begin{cases} 0 & (x < a)\ or\ (x > d) \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \dfrac{d-x}{d-c} & c \le x \le d \end{cases}$$

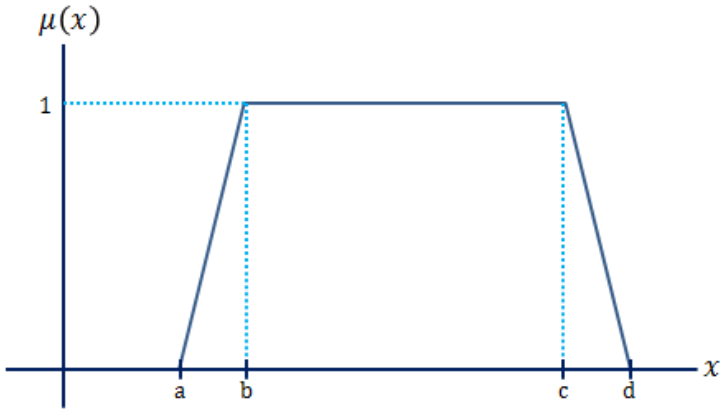


Figure 4 - 10. Trapezoidal membership function

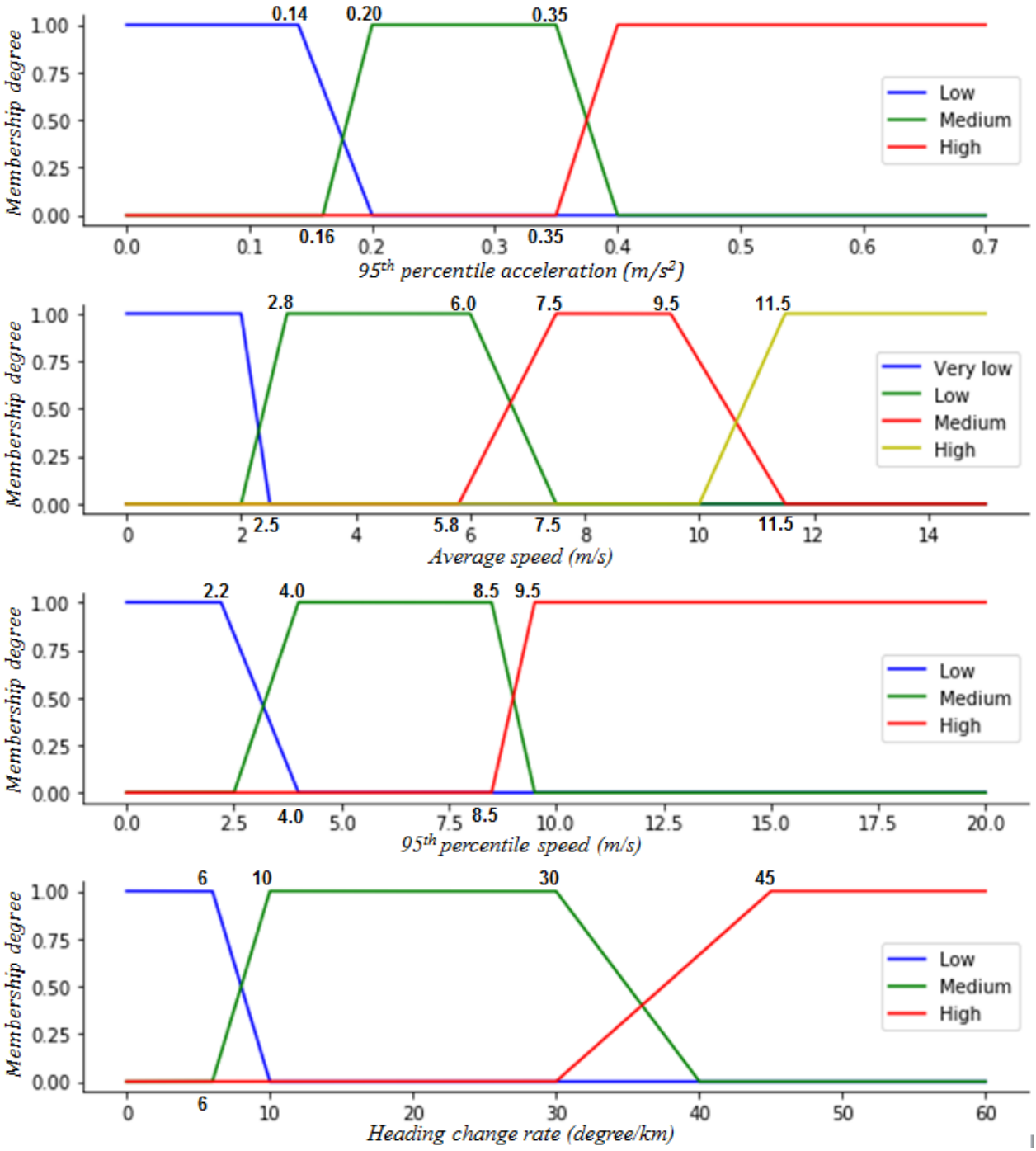


Figure 4 - 11. Membership function of variables

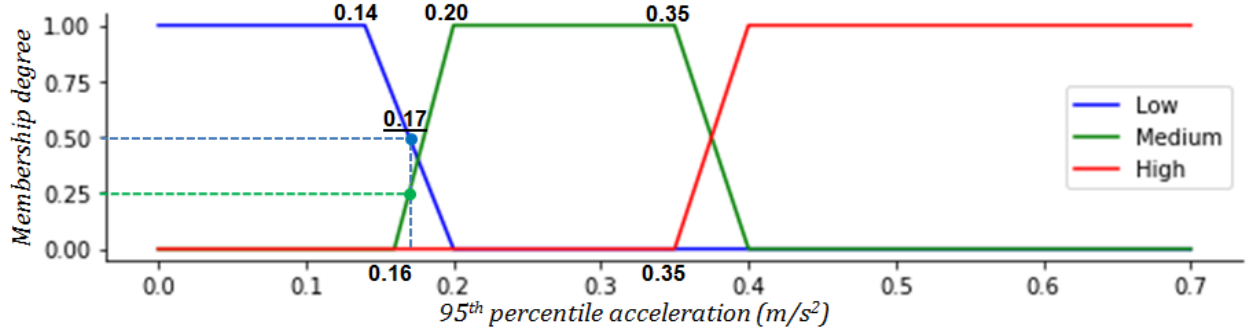


Figure 4 - 12. Estimating membership degrees of the 0.17 m/s² 95th percentile acceleration

The 2.3-m/s average speed belongs to *very low* and *low* with degrees of 0.4 and 0.375, respectively. It is not member of *medium* set and *high*. The 95th percentile speed of 4 m/s completely and only belongs to *low*. Similar, the heading change rate of 25 degree/km is a member of the *medium* set.

Table 4 - 4. Membership degrees of the exemplary segment

Variables	**Membership degree**			
	Very low	*Low*	*Medium*	*High*
95th percentile acceleration	-	0.500	0.250	0
Average speed	0.400	0.375	0	0
95th percentile speed	-	1	0	0
Heading change rate	-	0	1	0

4.4.5.3. Inference

Inference is deriving a conclusion from logic. Rules of inference specify conclusions that are drawn from assertions assumed or known previously to be true. In a fuzzy model, inference is the process where relationships between input and output fuzzy variables are defined to estimate the membership degrees of an observation to output linguistic terms. For the mode classification problem, the output variable is mode (x_{mode}). In fuzzy model, it corresponds to four (output) linguistic terms that is, walk, bike, car and bus/tram, which is presented by $T(x_{mode}) = \{T_{x_{mode}^{walk}}, T_{x_{mode}^{bike}}, T_{x_{mode}^{car}}, T_{x_{mode}^{bus/tram}}\}$

A rule is involved in a situation where travelling by a specific mode show values of four variables (i.e. 95th percentile acceleration, 95th percentile speed, average speed, heading change rate) simultaneously. Rules are set up based on knowledge about practical operations of modes.

The basic version is the single fuzzy *If-Then* rule that follows the form "If x is T_x, than y is T_y". It can be interpreted as if the input fuzzy x has the linguistic value of T_x, the output fuzzy y should have the linguistic value of T_y. For example, if the 95th percentile acceleration of a segment is *high*, the mode of this segment should be *car*. The rule version is "If $x_1$is *high*, x_{mode} is *car*". The condition part "If $x_1$is *high*" is antecedent whilst "x_{mode} is *car*" is the consequent part.

Nevertheless, the practice requires the combination of more than one variable in parallel to draw a conclusion. For example, a *very low* average speed may be *walk* but may be *bike.* However, a *very low* average speed together with a *low* 95% max speed would be indicators of *walk* with higher certainty than those of *bike*. To combine input fuzzy linguistic terms in order to figure out membership degree of the output fuzzy term, the two most common operators, that is, AND operation and OR operation, are used. AND operation is involved in the use of function *min* whilst OR operation is involved in the use of function *max*. We used AND operation to combine input fuzzy sets. Specifically, the membership degree of the output fuzzy set (i.e. mode) is the minimum degree of membership among degrees of input fuzzy sets considered.

Table 4 - 5. Rules of fuzzy-based model

Rule	Input				Output
	95th percentile acceleration	***Average speed***	***95th percentile speed***	***Heading change rate***	***Mode***
1	Low	Very low	Low	-	Walk
2	Low	Very low	Med	-	Walk
3	Low	Very low	High	-	Bike
4	Low	Low	Low	-	Walk
5	Low	Low	Medium	Low	Bike
6	Low	Low	Medium	Medium	Bike
7	Low	Low	Medium	High	Walk
8	Low	Low	High	-	Bike
9	Low	Medium	Low	-	Bike
10	Low	Medium	Medium	Low	Car
11	Low	Medium	Medium	Med	Bike
12	Low	Medium	High	-	Car
13	-	High	-	-	Car
14	Medium	Very low	Low	-	Walk
15	Medium	Very low	Medium	Low	Bike
16	Medium	Very low	Medium	Med	Bike
17	Medium	Very low	Medium	High	Walk
18	Medium	Very low	High	-	Bike
19	Medium	Low	Low	-	Bike
20	Medium	Low	Medium	Low	Bus/tram
21	Medium	Low	Medium	Medium	Bike
22	Medium	Low	Medium	High	Walk
23	Medium	Low	High	-	Bus/tram
24	Medium	Medium	Low	-	Bike
25	Medium	Medium	Medium	Low	Bus/tram
26	Medium	Medium	Medium	Medium	Bike
27	Medium	Medium	Medium	High	Bike
28	Medium	Medium	High	-	Car
29	High	Very low	-	-	Car
30	High	Low	Low	-	Bus/tram
31	High	Low	Medium	-	Car
32	High	Medium	Low	-	Car
33	High	Medium	Medium	-	Car
34	High	Medium	High	-	Bus/tram

Based on the practical relationships between four input variables' values and four modes, we created 32 rules presented in Table 4-5. Theoretically, with 3 terms of 95th percentile acceleration, 4 terms of average speed, 3 terms of 95th percentile speed and 3 terms of heading change rate, there should be 108

combinations and thus 108 rules. However, some cases are impossible in practice. For example, values of all variables are high. It is unreal because an object cannot run at high speed continuously and changes direction along with speed widely in parallel. Many cases can be combined by only one rule. For example, travelling with high average speed is enough for imputing car, no matter what other variables are. As indicated above, heading change rate is expected to support speed and acceleration variables, thus we considered terms of this variable in case other variables may be insufficient for addressing the ambiguity between modes.

Let returning our example to estimate the membership degree of the output term by rule 1. The rule 1 (see Table 4-5) is "*If* 95th percentile speed is *low* and average speed is *very low* and 95th percentile speed is *low* then the mode is *walk*". As can be seen in Table 4-4, the degrees of *low 95th percentile speed*, *very low average speed*, *low 95th percentile speed* are 0.5, 0.4, 1, respectively. Then the degree of segment belonging to *walk* term is 0.4, the minimum value of three input terms. By the same way, we estimated the membership degrees of modes according to rules (see Table 4-6).

4.4.5.4. Defuzzification

Defuzzification, as the name implies, is the opposite process of fuzzification. The output variable that is the predicted mode of segment is the most probable mode among modes rather than the exact mode. In this sense, the mode that has the larger membership degree than other mode's degrees should be chosen.

In the previous step, owing to the overlapping between MFs, a mode may fall into the effects of multiple rules, thus degrees of a mode generated by different rules should be combined. Operator OR with function *max* is used. For each mode, degrees of all rules are considered. The final membership degree of this mode is the highest degree of rules' degrees. For example, in Table 4-6, there are 7 rules pertaining to *walk*, including rule 1, rule 2, rule 4, rule 7, rule 14, rule 17 and rule 22. Among them, rule 1, rule 4 and rule 14 have degrees over 0. The largest of 0.4 is contributed by rule 1. As a result, the membership degree of *walk* is 0.4. By the same way, the degrees of bike, car, bus/tram are computed at 0.25, 0, 0, respectively. So the segment belongs to *walk* term at higher degree more than three other modes. And thus the segment is the most probable on foot. However, it should be noted that, the segment may be by bike also.

4.4.5.5. Output

In the last step of fuzzy logic-based travel mode inference, segments need to be labeled. The outcome of the previous step is membership degrees of modes. We are able to determine the mode of segment as which has the largest degree. However, we wanted to see how different degrees are, which reflects the certain level of travel mode decision. To do it, probability of each mode was estimated by the equation 4-5.

$$P_i = \frac{MD_i}{MD_{walk}+MD_{bike}+MD_{car}+MD_{bus/tram}} \quad (4\text{-}5)$$

Where:

MD_i : membership degree corresponding to mode i (i.e. walk, bike, bus/tram or car)

P_i : probability of mode i.

Applying this equation for the example above, the probabilities of walk, bike, car and bus/tram are 0.62, 0.38, 0 and 0, respectively. Hence, the mode given to the segment is walk.

Table 4 - 6. Results of estimating membership degrees based on rules for the example: a segment with 95th percentile of acceleration: 0.17 m/s^2, average speed: 2.3 m/s, 95th percentile speed: 4 m/s, heading change rate: 25 degree/km

Rule	Input				Output	
	95th percentile acceleration	***Average speed***	***95th percentile speed***	***Heading change rate***	***Mode***	***Membership degree***
1	Low	*Very low*	Low	-	Walk	0.4
2	Low	Very low	*Med*	-	Walk	0
3	Low	Very low	*High*	-	Bike	0
4	Low	*Low*	Low	-	Walk	0.375
5	Low	Low	*Medium*	Low	Bike	0
6	Low	Low	*Medium*	Medium	Bike	0
7	Low	Low	*Medium*	High	Walk	0
8	Low	Low	*High*	-	Bike	0
9	Low	*Medium*	Low	-	Bike	0
10	Low	*Medium*	Medium	Low	Car	0
11	Low	*Medium*	Medium	Med	Bike	0
12	Low	*Medium*	High	-	Car	0
13	-	*High*	-	-	Car	0
14	*Medium*	Very low	Low	-	Walk	0.25
15	Medium	Very low	*Medium*	Low	Bike	0
16	Medium	Very low	*Medium*	Med	Bike	0
17	Medium	Very low	*Medium*	High	Walk	0
18	Medium	Very low	*High*	-	Bike	0
19	*Medium*	Low	Low	-	Bike	0.25
20	Medium	Low	*Medium*	Low	Bus/tram	0
21	Medium	Low	*Medium*	Medium	Bike	0
22	Medium	Low	*Medium*	High	Walk	0
23	Medium	Low	*High*	-	Bus/tram	0
24	Medium	*Medium*	Low	-	Bike	0
25	Medium	*Medium*	*Medium*	Low	Bus/tram	0
26	Medium	*Medium*	*Medium*	Medium	Bike	0
27	Medium	*Medium*	*Medium*	High	Bike	0
28	Medium	*Medium*	*High*	-	Car	0
29	*High*	Very low	-	-	Car	0
30	*High*	Low	Low	-	Bus/tram	0
31	*High*	Low	*Medium*	-	Car	0
32	*High*	*Medium*	Low	-	Car	0
33	*High*	*Medium*	*Medium*	-	Car	0
34	*High*	*Medium*	*High*	-	Bus/tram	0

Note: In each rule, italic term(s) has the smaller membership degree than those of others; therefore, its degree is the membership degree of the output mode term

4.4.6. Post-processing by GIS data

A long-lasting challenge in mode detection is the big confusion between bus and car (Gong et al., 2012; Wu et al., 2016; Xiao et al., 2015c). As can be seen in Figure 4-8 and Figure 4-9, the overlapping between distribution of four variables for bus and car trips is significant. To deal with this, the support of GIS data is vital. Bus/tram is different from car by stopping at predefined stops. Each bus/tram segment should begin and finish at bus stops. So a solution to discriminate bus/tram from car is to examine the distances from the first and the last points of each potential bus/tram segments to the nearest bus/tram stops. Besides, to achieve higher certain level, it is necessary to check all stable points of each segment is either close to bus/tram stops.

The outcome of fuzzy logic-based step was the probabilities estimated to modes. Walk and bike were generally detected well by acceleration and speed variables along with the support of heading change rate. So in case a segment was labeled either walk or bike, this label was the final decision. Whilst the label was either car or bus/tram, we thought about whether making a post-processing check by GIS data. The result of performing fuzzy logic model on data of pilot tests revealed that in case the probability is 1, the label given was reliable. In case the probability of car was equal to or over 2 times as high as that of bus/tram, the segment was certainly by car. Similarly, in case the probability of bus/tram was equal to or over 2 times as high as that of car, the segment was certainly by bus/tram.

A suspected segment was one labeled as car or bus/tram and the division of its car probability by its bus probability ranged between 0.5 and 2. So it was subject to two checks, as follows:

- If it has both the first and the last points close to bus/tram stops, it will be classified as a bus segment.

- Or if it has either the first point or the last point together with at least one stationary point adjacent to bus/tram stops, it will be classified by bus.

A stationary point is the point whose instantaneous speed is under 1.4km/h, which is in line with the speed threshold used in the filtering step. Closeness to a bus/tram stop means the Vincenty distance from the considered point to this stop is not over 100m.

If a suspected segment fails to both criteria, it will be by car.

GIS data of public transport POIs were achieved from an official website of Grand Lyon Data project at: https://data.grandlyon.com/equipements/points-darrft-du-rfseau-transports-en-commun-lyonnais/.

4.5. RESULTS AND DISCUSSIONS

We implemented numerical and statistical analyses by SAS whilst spatial analyses and the fuzzy logic algorithm were carried out by Python with the support of the *skfuzzy* library. The accuracy for pilot tests was approximately 85%. This made us confident in applying for TOMOS data. As indicated above, the detection results were compared with EDR-RA data according to modes. In EDR-RA, respondents were asked about the use of public transport network rather than metro and bus/tram separately. Thus, in this section, all metro and bus/tram trips were grouped into public transport trips.

The comparison was conducted according to three aspects:

- Comparison by trip rate to conclude rule-based trip detection whether causes seriously over-reporting of trips.

- Comparison of trip rates by modes to assess the performance of mode detection model adopted.

- Visualization of trips by modes to assess the appropriateness of mode detection results.

4.5.1. Trip rate and rule-based trip detection

This Sub-section is focused on the outcome of identifying trips based on rules related to stops and gaps. As discussed before, applying minimum time of signal gap and stationary status has been the main method to detect trip ends. However, it is a controversial and context-sensitive technique due to causing both under-reporting and over-reporting simultaneously. Missing trips results from the fact (that) an algorithm neglects short activities and thus considers several trips as only one trip. In some cases, stops at certain locations to wait for public transport, the green light at intersections or vehicle queue stuck at public works are identified as trip ends, leading a trip to be broken into two/several trips. Therefore, determining the time rules is a trade-off based on specific contexts. The best trade-off generates the nearly balance between under- and over-reporting of trips.

By applying a smaller threshold of stationary status and signal gap, more trips will possibly be generated. To put it another way, the high trip rate in a GPS-assisted survey would result from intentionally determining a low time threshold rather than the ability of GPS to record sufficiently travel pattern. In this chapter, the adopted threshold of 3 minutes allowed to avoid significantly misclassifying stops at intersections to wait for traffic lights or at public works as activities. This level is one minute more than the most heavily employed threshold in the literature as discussed in Chapter 2. Moreover, approximately 4% of total activities were reported to last under 3 minutes in EDR-RA. Therefore, the 180-second time rule caused a level of under-reporting.

Notwithstanding, it would inflate the trip rate owing to considering waiting for public transport at stations/stops over 3 minutes as trip ends. A degree of over-reporting was unavoidable.

The 3-minute threshold was the best trade-off for data of pilot tests compared with 120s, 150s, 210s and 240s. In the literature, it was deployed in (Bohte and Maat, 2009; Patterson and Fitzsimmons, 2016).

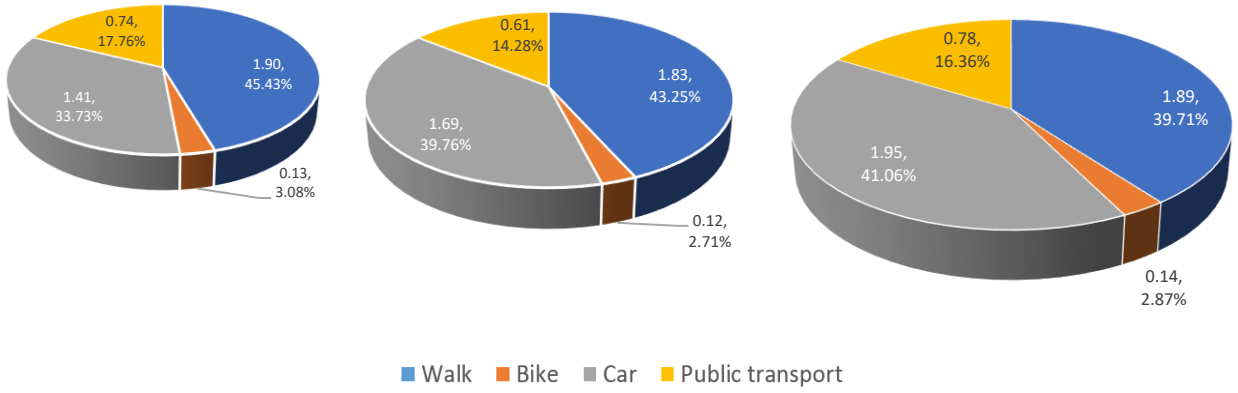
We thought about how to eliminate mentioned-above wrong identification of waiting for public transport as trip end by applying a higher threshold (e.g. 5 minutes) for continuous immobility within 100 m from the nearest stations/stops. Consequently, the total trips along with the numbers of both public transport and car trips decreased, which demonstrated some misclassifications were addressed. However, some car trips to pick up persons who get off public transport seemed to be omitted, not to mention waiting over 5 minutes still being as a trip termination. Although believing in the potential of modifying the time threshold based on contexts, with limited sample size of both pilot tests and TOMOS, finally the 180-second threshold was applied for all cases to generate the total of 1481 trips.

The TOMOS's trip rate of 4.76 is 114% times as high as the counterpart of 4.18 in EDR-RA. Although 4.76 is higher than the rates of GPS surveys reported previously, including 4.63 (Safi et al., 2015), 4.55 (Bohte and Maat, 2009), 3.74 (Patterson and Fitzsimmons, 2016), the 114% is still smaller than the level of 120% in (Bohte and Maat, 2009) and 139% in (Patterson and Fitzsimmons, 2016)[3]. Therefore, the time rule adopted in this study did not exaggerate seriously trip numbers and the high level of 4.76 trips per day mainly resulted from the nature of inhabitants' travel pattern in the research area.

[3] In these publications, the proportions between the rate in a GPS survey and that in a conventional survey were not presented, but we estimated as follows: for (Bohte and Maat, 2009), it is the division of 4.55 (GPS-based method) by 3.80 (DTS recall survey) whilst for (Patterson and Fitzsimmons, 2016), it is the division of 3.74 (DataMobile survey) by 2.70 (O-D survey 2013) in case of trips to Concordia.

4.5.2. Mode detection validation

As can be seen in Figure 4-13b and 4-13c, slightly more walking trips (1.89) were identified by the model than the expected number of 1.83. Because the walking rate in case 4-13b was estimated from telephone data in EDR-RA, it possibly included underreporting, especially in case walking was the dominant mode in the travel pattern (i.e. case 4-13a). The more a person makes a trip type, the higher its likelihood of being neglected (Forrest and Pearson, 2005; Patterson and Fitzsimmons, 2016).



(a) trip rate = 4.18 *(b) trip rate = 4.24* *(c) trip rate = 4.76*

Figure 4 - 13. Comparison of trip rates between EDR-RA and TOMOS. Case 4-12a describe shares and trip rates of modes in case of EDR-RA data. Case 4-12b describes TOMOS's expectation estimated from EDR-RA data as presented in Section 4.3 and Table 4-2. Case 4-12c describes the results of the mode detection model proposed

We were pretty surprised by the higher public transport rate (0.78) provided by the model in comparison with the expected value of 0.61. The difference may result from a particular level of over-estimating public transport trips, probably explained by the confusion in GIS data. Specifically, some public transport POIs are near-side stops that are located right before intersections to allow passengers to get in and out during the red light phase, thus limiting double-stopping. Waiting for green lights at these positions possibly caused misclassification of a non-transit segment as a public transport one. For segment detection, this issue may not be serious; however, for imputing main mode to a trip like this study, only one wrong public transport segment will lead to neglect other segments in the same trip. Nour et al., (2016) proposed to eliminate all the near-side stops before estimating the rate of stopping close to bus/tram stops. The threshold of the rate was determined based on empirical tests on ground truth data. The method was not appropriate for this study because eliminating near-side stops means partially blowing the chance of detecting public transport trips, especially short ones. And ground truth to calibrate the rate threshold was another barrier. Despite of the GIS confusion, the result of public transport trip identification was reasonable. Similar to the estimated level (14.3%), the public transport share of the mode detection (16.4%) is smaller than that in EDR-RA (17.8%).

As expected, the TOMOS car rate of the model showed a large difference at 0.54 compared with the car rate of EDR-RA. The model also delivered an obvious higher rate (1.95) compared with the expected level of 1.69, which made car become the dominant mode (Figure 4-13c). Interestingly, the car share in case 4-12c is slightly larger at only 1.3% than that in case 4-12b. This is in line with (Bohte and Maat, 2009) where a significant higher average trip number but a marginally larger trip share of car generated from GPS data than those from telephone data.

Regarding bicycle, its number and share in both surveys were so small that its changes were marginal.

Mentioned-above findings supported the conclusion that although the TOMOS sample seemed to be more mobile, GPS still detected more comprehensively and sufficiently trips. The mode shares and the trip rates of modes were reasonably logical with both travel pattern of inhabitants in research area and findings of existing studies. Accordingly, the mode detection method adopted performed satisfactorily.

4.5.3. Visualizing trips with modes detected

To validate mode detection results produced by the fuzzy-based model, two days of each person were randomly chosen to visualize on the map. Challenges were rising for trips that encompassed signal loss. However, most of them can be explained logically.

Figure 4.14 describes a typical example. The highly possible itinerary was that a person drove from suburb (Point A) to a metro station (Point B) to park and ride by metro in the morning. He/she did an activity at a place (point C) adjacent to a metro station in the city center. The device failed to receive signals during the period of moving from B to C. The signal re-occurred at noon at C once (s)he made a short walking trip to D. After about 45 minutes, (s)he returned to C on foot. Two trips would be a tour to have lunch. He/she returned B in the afternoon before walking and driving to A.

Trip	Segment	Depart at	Arrive at	Mode of segment	Trip mode	Color in figure
1	A-B	08:19:32	08:30:19	Car & Walk	Public transport	Red
	B-C	08:30:19	-	Metro		
2	C-D	11:31:59	11:34:56	Walk	Walk	Green
3	D-C	12:12:17	12:14:06	Walk	Walk	Blue
4	C-B	-	17:35:05	Metro	Public transport	Yellow
	B-A	17:35:05	17:47:06	Car & Walk		

During the segment B-C and C-B, there were no signals. B and C are connected together by dots. B and C are points close to metro stations Gratte-Ciel and Hotel de Ville Louis Pradel, respectively.

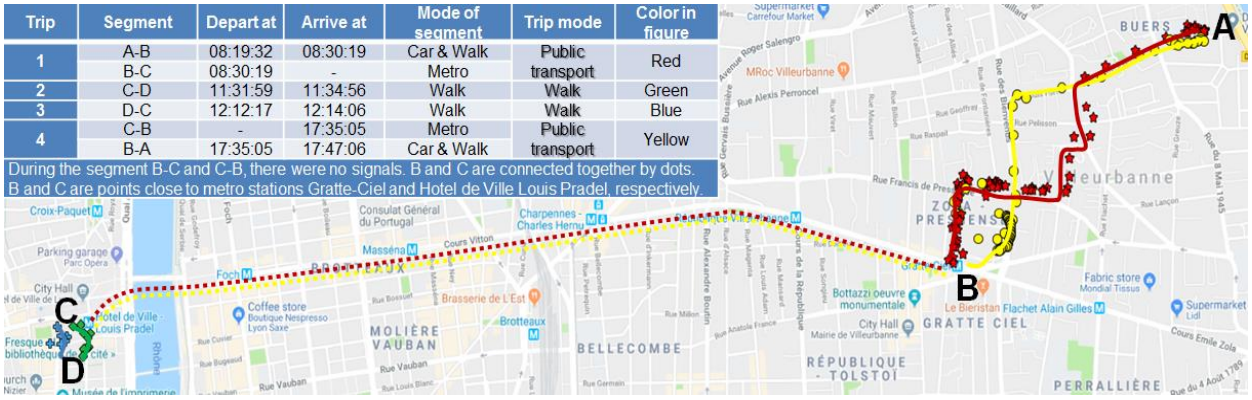


Figure 4 - 14. A person's trips during a day; GPS points of trips A-B, C-D, D-C, B-A are symbolized by red stars, green diamonds, blue pluses and yellow circles, respectively

Based on the itinerary, point A seemed to be his/her home and the segment B-C would be by metro. We were strongly confident in a four-trip tour thanks to the symmetrical travel patterns in the morning and afternoon together with the combination between public transport and private vehicle in the park-and-ride model of persons living in the outskirts of cities. Therefore, trip modes identified seemed satisfactory and reliable.

Notwithstanding, the arrival time to C of segment B-C and the starting time from C of segment C-B were questionable. In most other tested cases, we saw the similar logic between trips, modes and public transport facilities.

4.6. SUMMARY

This chapter developed a model using data of pilot tests before applying it for data of a GPS-based experiment. The results were compared with data of the CATI household travel survey in Rhone-Alpes, France. The method to translate GPS data into trips and detect their modes began with filtering data prior to using gap- and stop-related rules to break trajectories into trips and then segments. Travel modes were

recognized by a fuzzy theory-based model. To enhance the chance of correctly detecting transit segments, GIS data were used.

Besides the accuracy of 85% on pilot tests' data, the results of comparing travel patterns extracted from GPS data and CATI data, which were interpretable and compatible with previous findings, demonstrated the reasonable performance of the proposed model.

Similar to previous studies, the algorithm's drawback was misclassification between public transport and car due to the distribution of transit stops. Another limitation is that validating the mode detection model for TOMOS data were conducted based on assuming EDR-RA sample being able to represent TOMOS sample. The 80-person size of TOMOS would be small to some extent.

Findings confirm the advantage of GPS in capturing comprehensively and objectively citizens' travel patterns. With small amount of data of pilot tests, if well being developed, a probabilistic model would be sufficient for exploiting knowledge from a bigger GPS data set without ground truth collection. Avoiding requiring participants to validate travel diaries during the survey time is the key to alleviating burden on participants and thus promoting GPS-assisted travel survey in terms of both recruitment and duration. However, signal loss and limitations of inference model have prevented GPS from acting as a complete substitution of CATI.

Chapter 5: HIERARCHICAL PROCESS TO DETECT TRAVEL MODES FROM DATA OF MOTORCYCLE-DEPENDENT CITY (HANOI)

5.1. INTRODUCTION

Chapter 5 analyzes data collected by smartphones in Hanoi where the motorcycle is the most common travel mode. The inclusion of motorcycle is a new challenge to travel mode prediction because previous studies have concentrated on the modes of walk, bike, car, bus/tram and rail/train. Motorcycle makes the classification problem more complex because it can show similar values of indicators (e.g. acceleration, speed) to those of other modes.

A hierarchical process to separate each mode from others by its unique characteristics associated with speed, acceleration and spatial relationship with bus stops is developed. The ambiguities between modes are expected to be limited in each step; as a result, the accuracy can be enhanced. Fuzzy logic method introduced in the previous chapter is a part of the process. The prediction results help us to assess difficulties and challenges to impute travel modes from GPS data in a motorcycle-overwhelmed city like Hanoi.

The remaining content of this chapter is structured into six Sections. The next is reviewing the literature to highlight how previous authors combine mode detection methods (i.e. rule-based, probability-based and machine learning approaches) into hierarchical processes to detect modes. Section 5.3 presents transport conditions in Hanoi whilst data preparation is content of Section 5.4. Then, the description of method constitutes Section 5.5. Results and discussions are documented in Section 5.6. In the last Section, a summary is made.

5.2. LITERATURE REVIEW

As discussed in Sub-section 2.5.1, there are three main mode detection method groups, that is, (1) deterministic, (2) probabilistic and (3) machine learning methods. Each group has its own merits and shortcomings. In existing studies, travel mode lists vary; however, all authors have paid attention to basic modes in developed countries, including walk, bike, car, bus/tram and metro (Bohte and Maat, 2009; Burkhard et al., 2020; Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Gong et al., 2012, 2012, 2018; Semanjski et al., 2017; Shafique and Hato, 2015; P. Stopher et al., 2008a). Xiao et al., (2015) surveyed e-bike that was used as much as car in their sample collected in Shanghai, China. Motorcycles, which are common in developing countries, have not been considered.

Mode detection processes can be divided into two types according to the number of steps taken to generate an outcome. The first type includes all-in-one processes, in which the modes of all trips are detected by only one model. The second uses a hierarchical process to build a multi-step procedure to infer modes at aggregate levels (e.g., non-motorized and motorized modes) prior to disaggregated levels (i.e., each mode). Machine-learning models are typically used for all-in-one processes, and so they have both the advantages and disadvantages of the learning methods discussed above.

Hierarchical processes are based on separating modes that are sufficiently different from one another; thus, the division can attain a very high rate of success. The simplest versions of hierarchical classification are rule-based methods. In Bohte and Maat (2009), for example, walking trips were detected first using maximum and average speed, as walking is the slowest mode of transport. With a rule-based

hierarchical process, Gong et al. (2012) achieved an accuracy level of 82.6% for the case of New York. A complex hierarchical classification can be developed by combining methods. Rasmussen et al. (2015) successfully imputed 92.4% of trip segments. First, they distinguished rail segments from others by examining the proximity of points on each segment to the rail network. They then developed a fuzzy-logic algorithm based on speed and acceleration to distinguish walking and cycling segments from car and bus segments. Finally, they resolved the confusion between car and bus segments by using map-matching algorithms. Nour et al. (2016) applied the k-nearest neighbor algorithm to categorize all data into aggregate levels (i.e., motorized and non-motorized modes) and then disaggregated levels (walk, bicycle, car, bus). To enhance the detection of bus and car, all segments of motorized modes were analyzed to determine whether they involved bus segments by estimating the average rate of stopping close to transit stations. The researchers found increases of 65% and 10% in recall and precision of bus, respectively, compared with those of k-nearest neighbor. Marra et al. (2019) introduced a process that involved first segmenting trips into walk and non-walk segments on the basis of speed- and time-based rules. Next, train and bus/tram segments were identified by probabilistic functions using actual operational data of public transport. Historically visited places and routes extracted from multi-day GPS data improved the detection of transfer points. Finally, a random forest model was developed to infer bicycle and car trips. The proposed system was tested in Zurich and Basel (Switzerland); an accuracy of 86.1% was attained regarding classifying the four modes, namely, walk, bus/tram, car or bicycle, and train, and an accuracy of 87% was attained regarding classifying the modes bicycle and car.

Table 5 - 1. Comparison of hierarchical and all-in-one processes

	Hierarchical process	**All-in-one process**
Definition	Detect modes from aggregate levels to disaggregated levels	Detect all modes simultaneously
Main algorithms used	Rule-based and probability-based	Learning-based
Frequently used in	Infancy of GPS-based travel surveys	Recent times
Appropriate data size	Both data in tests and big data	Big data
Interpretability of results	High	Low
No. of variables used	Usually several; fewer than modes	Many; usually more than modes
Main modes classified	Walk, bike, bus/tram, car, metro and train	
Geographical research scope	Mainly in cities, metropolitan areas of developed countries and well-structured cities in China	

The advantage of the hierarchical process over the all-in-one process is noticeable in the variable selection. In an all-in-one process, each variable affects all trips unnecessarily, leading to an incorrect classification. Semanjski et al. (2017) reported that adding a speed variable to a support vector machine model, that had previously used only spatial variables, resulted in a small improvement of overall accuracy at the expense of increasing the misclassification of walk and bicycle segments. In each step in a hierarchical process, some unique variables for several mode groups are deployed to limit any confusion. Moreover, a hierarchy does not use numerous variables simultaneously but focuses on mode groups; therefore, it can fit small and imbalanced data with a low risk of overfitting. However, error propagation is a problem. As researchers consider more disaggregated levels, the accuracy will decrease. The accuracy of steps lower in the hierarchy is equal to or lower than that of the step above. If an observation is wrongly classified in the previous step, there is no way to correct it in the subsequent stage.

Table 5 - 2. Synthesis of mode detection studies based on methods and processes adopted

Authors and studies	Modes	Methods	Variables	Accuracy	Process types
Tsui and Shalaby, (2006)[T]	Walk, cycle, bus, auto, streetcar, subway, off-road	Fuzzy-logic and map-matching	Speed, acceleration, data quality, spatial information	94%	Complex hierarchical
Stopher et al., (2008)[T]	Walk, bicycle, car, bus, tram	Probability matrix	Speed, spatial information	95%	Simple hierarchical
Bohte and Maat, (2009)[E]	Car, train, bicycle, foot, other	Rule-based	Speed, spatial information	70%	Simple hierarchical
Gong et al., (2012)[T]	Walk, subway, rail, car, bus	Rule-based	Speed, acceleration, spatial information	82.6%	Simple hierarchical
Rasmussen et al., (2015)[E]	Walk, bicycle, bus, car, rail, other	Fuzzy-logic and map-matching	Acceleration, speed, spatial information	92.4%	Complex hierarchical
Nour et al., (2016)[T]	Walk, bicycle, transit, auto	KNN and rule-based	Speed, acceleration, jerk, spatial information (i.e. transit stop rate)	92.5% (1)	Complex hierarchical
Marra et al., (2019)[E]	Walk, bus/tram, train, car, bicycle	Rule-based, probability-based and RF	Speed, acceleration, heading, actual operational data of public transport, historical travel data	86.1%(2) and 87%(3)	Complex hierarchical
Schuessler and Axhausen, (2009)[E]	Walk, cycle, car, urban public transport, train	Fuzzy-logic	Acceleration, speed	Not available	All-in-one
Stenneth et al., (2011)[T]	Train, bus, walk, car, bicycle, stationary	**RF**, NB, BN, DT, MLP	Acceleration, speed, spatial information and real-time data	92.8% and 92.9% (4)	All-in-one
Shafique and Hato, (2015)[E]	Walk, bicycle, car, train	**RF**, SVM, AdaBoost, DT	Acceleration values for 3 directions	99.8%	All-in-one
Xiao et al., (2015)[E]	Walk, bicycle, E-bike, bus, car	**BN**, SVM, MNL, ANN	Speed, acceleration, average heading change, distance	92.7%	All-in-one
Feng and Timmermans, (2016)[E]	Walk, bicycle, bus, car, motorbike, running, tram, metro, train, activity	**BN**, NB, LR, MP, DT, SVM, C4.5	Speed, acceleration, distance, data quality, spatial information	99.8%	All-in-one
Semanjski et al., (2017)[E]	Walk, bus, car, foot, train	SVM	Spatial information	94%	All-in-one
Dabiri and Heaslip, (2018)[E]	Walk, bicycle, bus, driving, train	**CNN**, KNN, SVM, DT, RF, MLP	Speed, acceleration, heading change rate, jerk	84.8%	All-in-one
Gong et al., (2018)[E]	Walk, bus, tram, auto	RF	Travel time, length, speed, participant's information, spatial information	Not available	All-in-one
Feng and Timmermans, (2019)[E]	Car, bus, bicycle, walk, train	BN	Speed, acceleration, distance, data quality, spatial information	88.1%	All-in-one

(1) Estimated based on the confusion matrix in (Nour et al., 2016).
(2) For classifying walk, bus/tram, train and private modes (i.e. car and bicycle) in Basel data;
(3) For classifying bicycle and car in Basel data;
(4) Refer to overall precision and overall recall, respectively.
RF: Random Forests; DT: Decision Tree; SVM: Support Vector Machine; KNN: K-Nearest Neighbors; MLP: Multilayer Perceptron, CNN: Convolutional Neural Network, LR: Linear Regression, NB: Naïve Bayes, BN: Bayesian Network, ANN: Artificial Neural Network.
In a study reporting a number of methods, the main method is in bold and the accuracy presented in the table is its.
[E] Refers to an experiment whose valid data are comprised of either at least 50 persons or at least 350 days (equivalent to 50 persons * 7 days) or at least 1400 trips (equivalent to 350 days * 4 trips/day).
[T] Refers to a test at small scale and thus fails to meet the experiment-specific criteria.

Thus, although all-in-one processes with machine-learning methods are currently preferred, hierarchical processes would be a better choice to overcome the problem of imbalanced data and to improve the interpretability (see Table 5-1, Table 5-2). The inclusion of new travel modes in the classification list is interesting and contributes to the diversity of the mode detection field.

5.3. HANOI URBAN TRANSPORT

Urban mobility in Hanoi is mainly by motorcycle, walk, bike, bus and car. The striking characteristic is the dominance of motorcycles and the frequent occurrence of traffic jams at rush hours.

**** The dominance of motorcycles***

Motorcycle is attractive to inhabitants in developing countries like Taiwan (Hsu et al., 2003). In large Vietnamese cities, each household has at least one, typically two or three motorcycles (Huynh and Gomez-Ibañez, 2017). According to the statistics of the Hanoi police department, registered motorcycles made up about 90% of motorized vehicles in 2015. On average, a household possessed 2.84 vehicles with 2.56 motorcycles in 2014 (Huynh and Gomez-Ibañez, 2017). As can be seen Figure 5-1, in case of disregarding walking, about 80% of daily trips were by motorcycle in 2008. The shares of motorcycle during the period between 1995 and 2008 kept rising dramatically. There have not been any recent reports of prestigious sources on the updated trip rate by motorcycle; however, transport practitioners and researchers believe that its percent is around 80% (Nguyen, 2016; Vu and Ha, 2016).

Table 5 - 3. Breakdown of registered motorized vehicles in Hanoi from 2010 to 2015

Year	Car	Bus & Coach	Truck	Motorcycle	Other	E-bike	Total
2010	186 662	25 533	92 130	3 577 041	6 560		3 887 926
	4.80%	*0.66%*	*2.37%*	*92.00%*	*0.17%*		*100.00%*
2011	235 349	28 186	103 326	3 980 070	7 034		4 353 965
	5.41%	*0.65%*	*2.37%*	*91.41%*	*0.16%*		*100.00%*
2012	298 365	24 117	108 904	4 444 127	8 509	Not	4 884 022
	6.11%	*0.49%*	*2.23%*	*90.99%*	*0.17%*	available	*100.00%*
2013	311 061	24 889	113 827	4 660 761	9 010		5 119 548
	6.08%	*0.49%*	*2.22%*	*91.04%*	*0.18%*		*100.00%*
2014	333 252	26 028	122 797	4 852 380	9 736		5 344 193
	6.24%	*0.49%*	*2.30%*	*90.80%*	*0.18%*		*100.00%*
2015	368 665	26 927	137 466	5 045 672	12 999	10 686	5 602 415
	6.58%	***0.48%***	***2.45%***	***90.06%***	***0.23%***	***0.19%***	***100.00%***

Source: Data were provided by the Hanoi police department

There are four reasons explaining the uncontrolled proliferation of motorcycle. The first is the affordable price and cheap operational cost. With a budget of over 700 euros, a resident can buy a motorized two-wheel vehicle. These prices are possibly paid by almost all citizens because the GDP per capita per year in Hanoi is about 5000 US$. The use of motorcycle is mainly related to fuel costs whilst other costs like the parking fee are very cheap. The standard parking fees for motorcycle are 0.13 and 0.22 US$ in daylight and in the evening, respectively.

Travel by motorcycle obviously fits narrow alleys overwhelming in the road network of central business districts. It provides door-to-door access to destinations. Besides, in case increasingly serious

traffic jams, motorcycles can move and schedule flexibly to overcome and/or avoid congested road segments. Keeping and parking motorcycles are easier than other private vehicles. Hanoians usually have long trip chains. For example, in the morning a person may take his/her children to a restaurant to let them have breakfast before leaving them at school and then going to his/her workplace. It is difficult, even unfeasible to make this chain by bike and public transport whilst cars are unaffordable for numerous persons.

Another reason would be the poorly competitive capacity of alternatives, especially the public transport system, which will be presented in the following.

The last but not least is the shortage of effective and timely actions of the national/local government. Motorcycle use restriction has been issued; however, the authority has not pursued it with great determination. Discrete solutions such as fixing the maximum number of motorcycle registered per person or putting much high registration fees were soon out-of-date with endless debates on whether invading citizen's rights (Nguyen et al., 2019b).

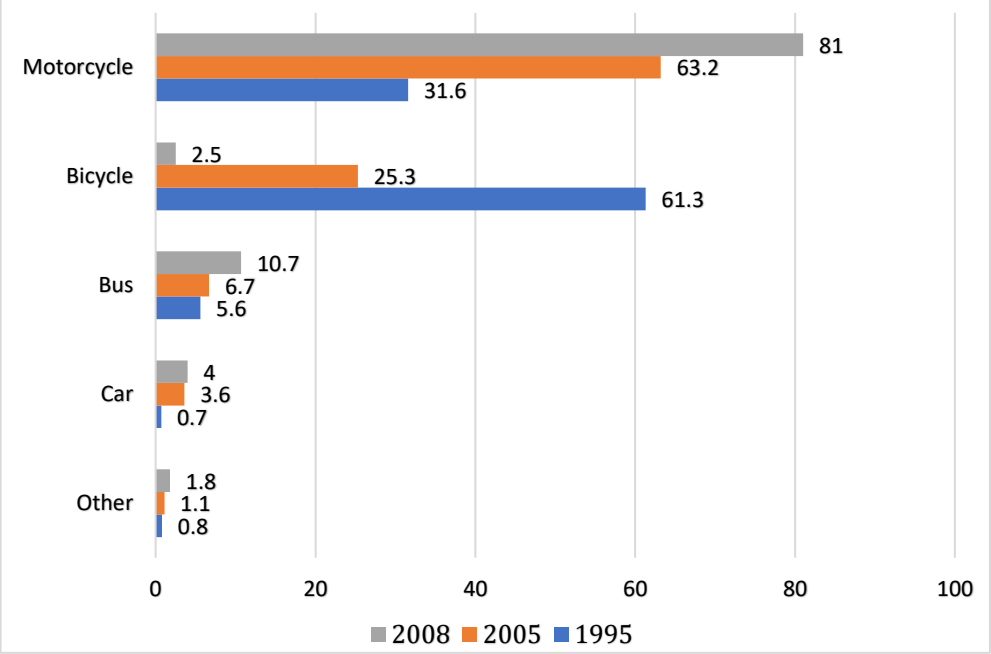


Figure 5 - 1. Mode share in Hanoi mobility (unit = %)

Source: Drawn by figures presented in (Huynh and Gomez-Ibañez, 2017)

**** The minor and decreasing use of bicycle***

Bicycle was the most popular means in Hanoi in the 1990s; however, its use has been decreasing rapidly. In 2008, only 2.5% of trips in a day were by bicycle. Now, bikes are primarily utilized by restricted-budget groups like the poor and pupils/students.

The demise of bike usage possibly comes from the boom in motorcycles. In fact, bikes cannot be comparable with the motorized mode in terms of speed and comfort.

The weather conditions in Hanoi are an impediment to travel by bike. In summer, the temperature is high at 40°C in the air and 50°C on the roads. The hot and humid weather occasionally goes with abundant rainfalls. By contrast, the winter is characterized a substantial drop in temperature at around 15°C, even under 10°C with drizzle.

The local government has not taken any actions to make Hanoi become a bike-friendly city. There are not bike-sharing facilities adjacent to public transport points. No cycle-lanes are offered.

**** The downward trend of bus and frequent delays of mass rapid transit***

The 10-year period from 2003 to 2013 was the golden age of bus in Hanoi. Thanks to the subsidy and new bus development-oriented policies, informal transits with the share of less than 5% was reformed to impressively become the Vietnamese urban bus model that offers high quality with a cheap ticket price. The share rose from 6.7% in 2005 to 10.7% in 2008. Although urbanization has processed at a very high speed, the government seemed to have misconceptions about the competence of bus, thus has not paid adequate attention to develop mass rapid transit. Although a number of bus routes have been introduced to widen the catchment, five recent years witnessed the decrease in bus ridership, leading the bus share to be around 9%. In 2019, over 110 bus routes are running.

Figure 5 - 2. Bus network in Hanoi

Source: https://xe-buyt.com/ban-do-xe-bus-ha-noi

2017 saw the kick-off of the first BRT corridor after the 10-year construction process; however, its performance is much lower than expected (Nguyen et al., 2019b). As a result of being disappointed with the poor operation of the first corridor, the BRT planning has been stopped.

The first metro line between Hadong and Catlinh has not been inaugurated although its infrastructure was completed at the beginning of 2019. The exact kick-off time has not been revealed. The second line connecting Nhon bus hub and Hangco rail station with the financial and consultation of French agencies are under construction.

**** Rapid increase in car possession***

The prospect of a quick rise in the number of cars is a serious issue because cars use far larger road space compared with other modes, 8 times more than motorcycle for instance (Hsu et al., 2003). Cars made up 6.58% among motorized vehicles and the rate of car trips is roughly 4%, half the rate of bus trips. The shift from motorcycles to cars is increasing considerably, which was warned by (OECD, 2018).

**** Serious traffic congestion***

The unavoidable repercussion of the boom in private vehicle usage coupled with the lack of effective public transport alternatives is traffic congestions that goes serious in terms of both temporal and spatial scales. Hanoi has more and more typical congested points/areas and the congested duration tends to extend. In a recent speech of the leader of the Institute for Strategy and Development in Transport under the Ministry of Transport, traffic congestion cost of the capital is between 1 and 1.2 billion US$ per year (Boltze and Tuan, 2016).

Figure 5 - 3. Unrelenting and serious traffic congestions in Hanoi due to the rapid growth of private vehicles and poor public transport alternatives

Source: https://vietnamnet.vn/vn/thoi-su/an-toan-giao-thong/ha-noi-quyet-khong-de-tac-duong-qua-30-phut-434525.html

5.4. DATA PREPARATION

The survey time between mid-March and mid-April was expected to cover trips by the first metro line planned to come into operation on the first day of April. Unfortunately, its introduction was delayed; consequently, we had data of five modes, including walk, bike, motorcycle, bus and car. Car here refers to taxi, private car, technology-based service (e.g. Uber and Grab) and non-bus auto (e.g. coach running between fixed places and tourism coach). All data of other modes like train, air were disregarded.

The validated segments of 63 users were used to develop a hierarchical mode-detection process. As the records included only timestamps and coordinates in the World Geodetic System 1984 format, the distance between two consecutive points was calculated using Vincenty's equations (Vincenty, 1975). With distances and timestamps, speed and acceleration profiles were easily gauged. The criteria used to filter the data were as follows:

- The speed limit for roads in Vietnam is 120 km/h; therefore, points with speeds over this threshold were removed[4].
- Any point that had the same coordinate or timestamp as its previous point was ignored.
- Any segment with a duration of under 60 seconds was ignored.
- Any segment whose points were all outside Hanoi was excluded because this research did not use GIS data of the public transport systems in other provinces or cities. Travel connecting Hanoi with other provinces, however, was still within the scope of this study.

.

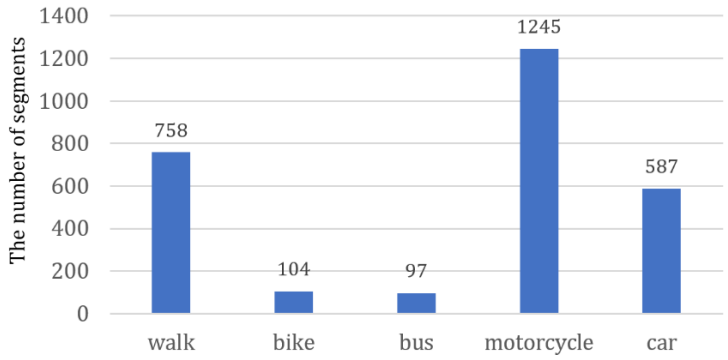


Figure 5 - 4. The numbers of valid segments by modes in the Hanoi survey

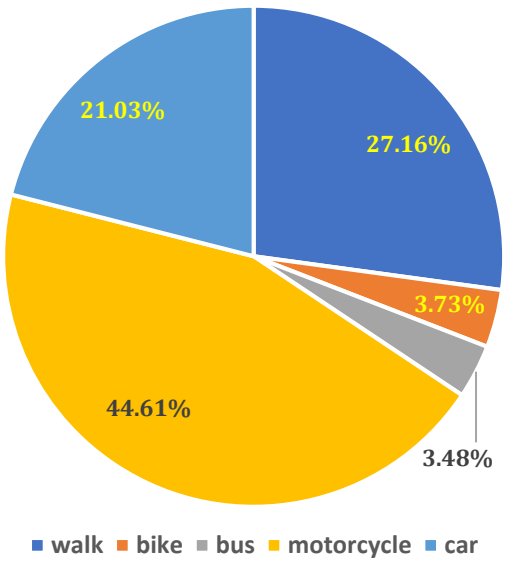


Figure 5 - 5. Mode shares in the valid data of the Hanoi survey

[4] It is important to note that a point with a speed over 120 km/h may not constitute noise or a bad record. Such points may represent an over-speeding situation of a car on an expressway. In Hanoi, the highest speeds of motorcycles and buses are around 70 km/h (the allowable level) because they are banned from running and do not operate on expressways, respectively. For this reason, disregarding points with speeds of over 120 km/h, enabling less computation, did not decrease the detection performance of segments by car, motorcycle and bus. However, if the comprehensive distribution of speed is desired in case of the speed limit at 120 km/h, the threshold to eliminate erroneous records may be 150 km/h, according to Wang et al. (2017).

Finally, we had 2791 adequate segments for further analysis with the numbers and rates of modes presented in Figure 5-4 and 5-5. It can be seen that the motorcycle segment number exceeds other figures, which is in line with the practical mode use in Hanoi. Excluding walk, cars are the second most heavily used with the number of 587 equivalent to the share of 21.03%. Bike and bus show moderate percentages at approximately 3.5%. Therefore, mode data of the Hanoi survey is unbalanced with major classes being motorcycle, car, walk and minor classes being bike, bus.

Compared with the initial outcome of the app (see Table 3-1), the number of modes reduced significantly, which can be explained by two major reasons:

- In Vietnam, 30/4 is the National Reunification Day or Day of Southern Liberation for National Reunification that marks the end of Vietnamese war in 1975 to begin unifying the North and the South. Staff and workers do not work on 30/4 and May Day (i.e. the first day of May). The Vietnamese Prime Minister approved the national long 5-day holiday ranging from 27/03 (Saturday) to 01/05 (Wednesday). During the holiday, many users travelled and were not in Hanoi, leading their data during this period to be eliminated.

- Among users, there were participants having international business days in the Philippines. During the most survey time a user who was a lecturer at University of Transport and Communications left Hanoi to give lectures and do research in Hochiminh City. A number of persons had jobs requiring them to travel to provinces around Hanoi like Ninhbinh, Thaibinh, Haiduong, Haiphong to do business tasks. Trips in these provinces were eliminated.

5.5. HIERARCHICAL PROCESS OF TRAVEL MODE IMPUTATION

The hierarchical mode inference process encompassed three steps. The process began by distinguishing walk and bicycle segments from motorized ones using a fuzzy logic algorithm. In the second step, bus segments were separated from other motorized segments using stop-related rules. In the last step, car segments were distinguished from motorcycle segments using a random forest model.

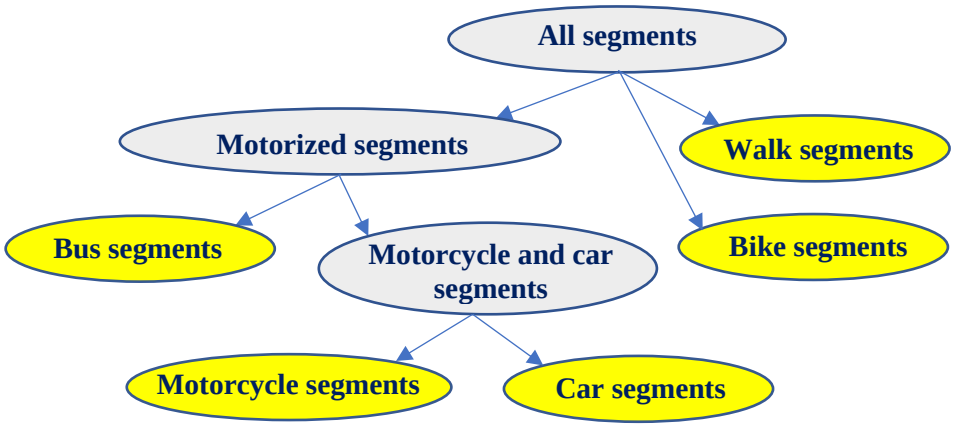


Figure 5 - 6. Three-level hierarchical mode detection process

5.5.1. Classifying walk, bike and motorized modes by fuzzy logic theory

**** Fuzzy logic model***

Because motorcycle can show values of acceleration and speed fairly akin to those of walk, bike, bus and car (see Figure 5.7). Thus, we decided to group motorcycle, bus and car into a motorized group.

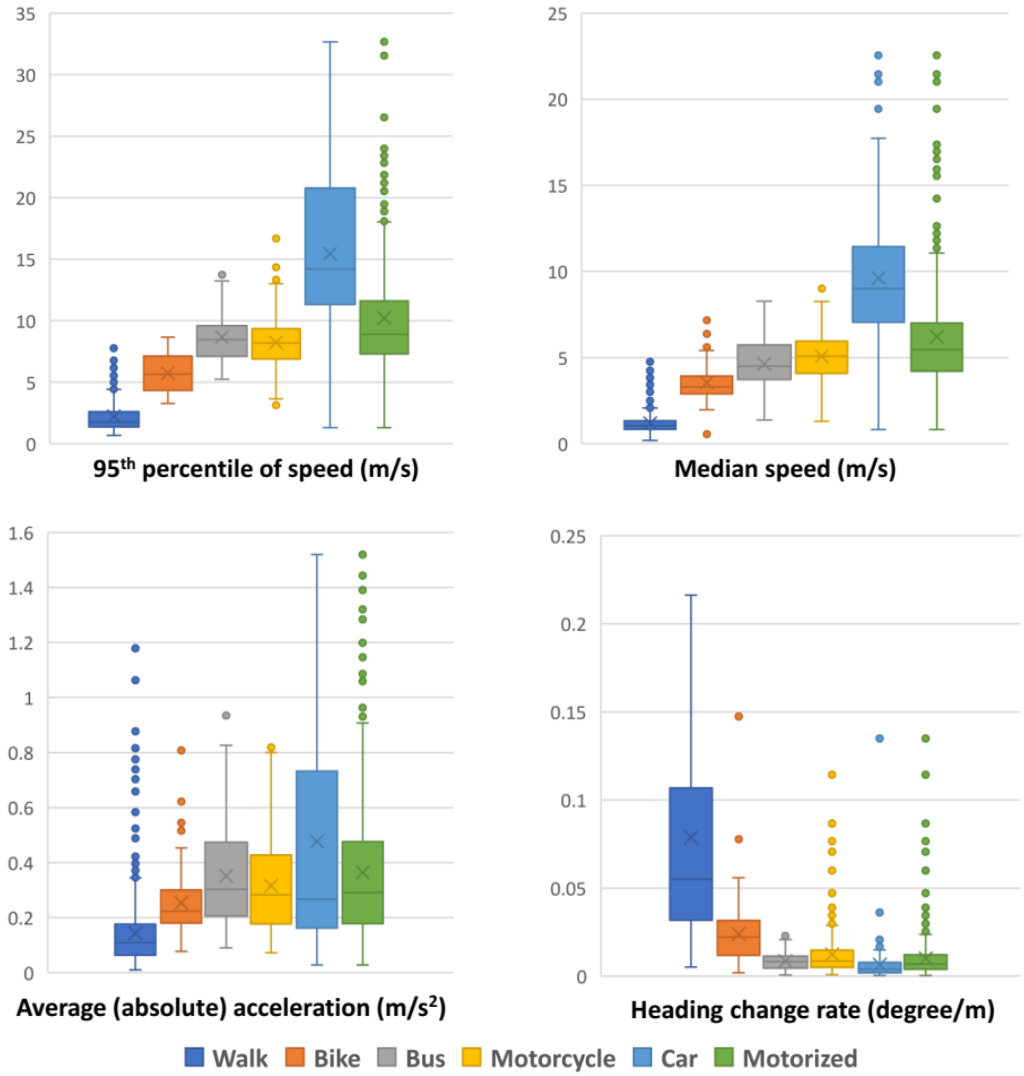


Figure 5 - 7. Boxplots of variables by travel modes

With fuzzy logic, speed and acceleration variables are generally sufficient for detecting walk, bicycle, and motorized segments (Rasmussen et al., 2015). However, the classification problem in Hanoi was more complex because motorcycle segments' acceleration and speed profiles were similar to those of walking and bike ones (see Figure 5-7). In particular, the overlapping of the 95th percentile of speed, median speed, and average acceleration between motorized and non-motorized modes was significant at low ranges (e.g., from 2 m/s to 8 m/s for the 95th percentile of speed); therefore, the heading change rate, which was the average change of heading per meter (Dabiri and Heaslip, 2018), was added. The heading between two points was calculated from their coordinates using equations presented by Dabiri and Heaslip (2018). As can be seen in Figure 5-7, the heading change rates were typically high for walking and typically low for motorized modes, which can be explained by the fact that motorized modes kept strictly to roads, but pedestrians did not always walk in a straight line from point to point. The heading change rates for bicycles were higher than those of motorized modes but lower than those of walking.

To limit the complexity of the rules in the fuzzy logic model, heading change rate was used as a supplement in case the speed and acceleration profiles between modes were ambiguous. On the basis of the trapezoidal membership functions (see Figure 5-8) and rules (see Table 5-4), each segment received three probabilities corresponding to walk, bicycle, and motorized modes. The mode with the highest probability was attributed to the segment.

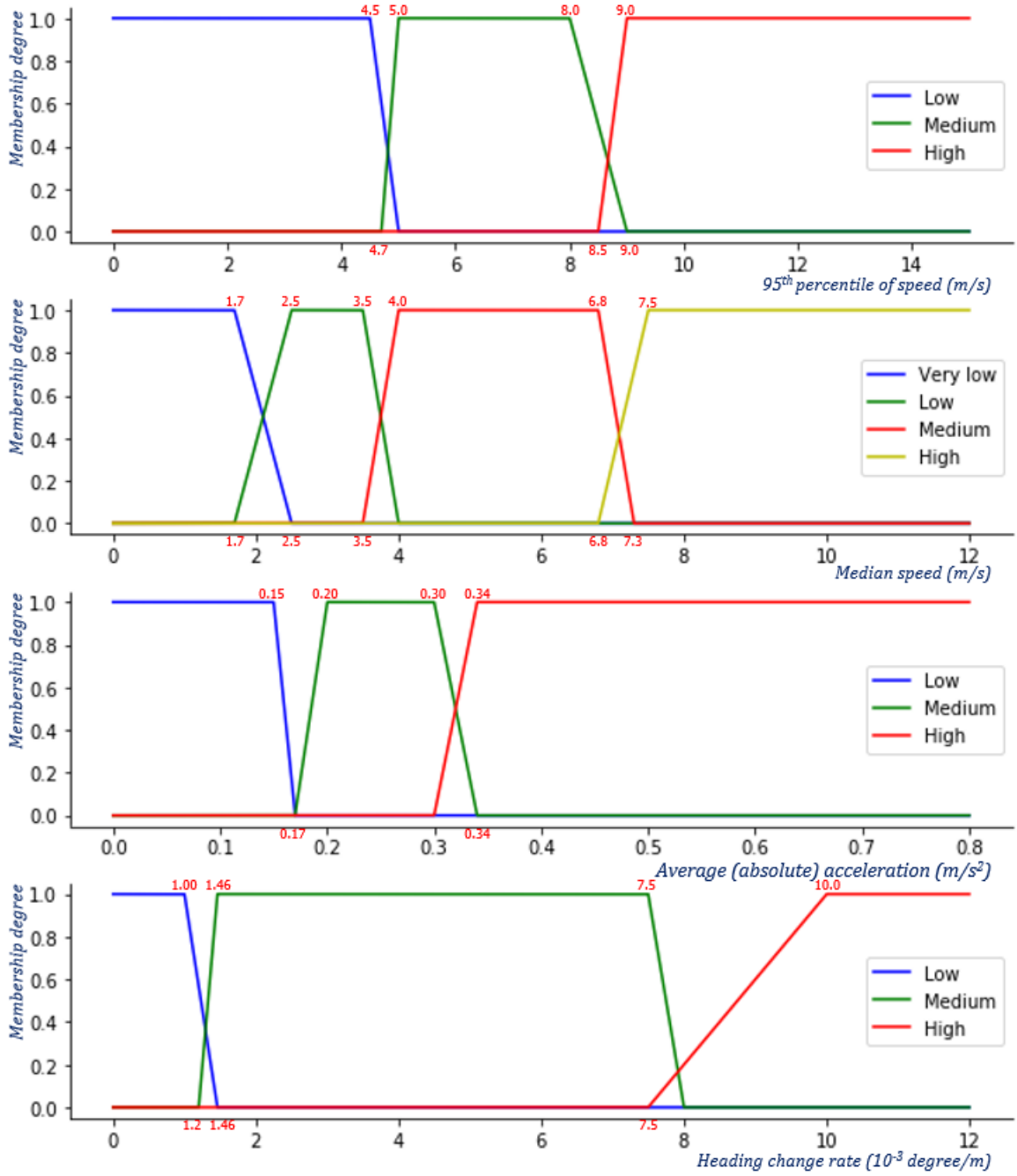


Figure 5 - 8. Membership functions

Table 5 - 4. Fuzzy rules of fuzz logic-based model to detect walk, bike and motorized modes

Rule	Input				Output
	95th percentile of speed	*Median speed*	*Average (absolute) acceleration*	*Heading change rate*	*Mode*
1	Low	Very low	-	-	WALK
2	Low	Low	Low	Low	MOTORIZED
3	Low	Low	Low	Medium	BIKE
4	Low	Low	Low	High	WALK
5	Low	Low	Medium	Low	BIKE
6	Low	Low	Medium	Medium	BIKE
7	Low	Low	Medium	High	WALK
8	Low	Low	High	Low	MOTORIZED
9	Low	Low	High	Medium	BIKE
10	Low	Low	High	High	WALK
11	Low	Medium	Low	-	BIKE
12	Low	Medium	Medium	-	MOTORIZED
13	Low	Medium	High	-	MOTORIZED
14	Low	High	Low	-	BIKE
15	Low	High	Medium	-	MOTORIZED
16	Low	High	High	-	MOTORIZED
17	Medium	Very low	Low	-	WALK
18	Medium	Very low	Medium	Low	BIKE
19	Medium	Very low	Medium	Medium	WALK
20	Medium	Very low	Medium	High	WALK
21	Medium	Very low	High	Low	MOTORIZED
22	Medium	Very low	High	Medium	WALK
23	Medium	Very low	High	High	WALK
24	Medium	Low	Low	Low	MOTORIZED
25	Medium	Low	Low	Medium	BIKE
26	Medium	Low	Low	High	WALK
27	Medium	Low	Medium	Low	MOTORIZED
28	Medium	Low	Medium	Medium	BIKE
29	Medium	Low	Medium	High	WALK
30	Medium	Low	High	Low	MOTORIZED
31	Medium	Low	High	Medium	BIKE
32	Medium	Low	High	High	MOTORIZED
33	Medium	Medium	Low	Low	MOTORIZED
34	Medium	Medium	Low	High	BIKE
35	Medium	Medium	Medium	Low	MOTORIZED
36	Medium	Medium	Medium	Medium	BIKE
37	Medium	Medium	Medium	High	MOTORIZED
38	Medium	Medium	High	-	MOTORIZED
39	Medium	High	Low	-	BIKE
40	Medium	High	Medium	-	MOTORIZED
41	Medium	High	High	-	MOTORIZED
42	High	-	-	-	MOTORIZED

**** Rule-based model***

To evaluate the competence of the fuzzy logic-based model, we developed a rule-based algorithm to identify walk, bike and motorized modes. Two variables were employed were 95th percentile speed and median speed. Rules set up should cover at least 50% of the data of each input variable for each mode. In addition, rules should produce the recall of every class over 50%. In case there was an overlap between modes, a priority was given to the mode whose share was larger in the sample. Rules were proposed as follows:

- If a segment has 95th percentile of speed smaller than 3.5 m/s and median speed smaller than 2 m/s, it will be labeled walk.

- If a segment has 95th percentile of speed smaller than 6 m/s and median speed smaller than 4 m/s and is not walk, it will be labeled bike.

- The remainder of segments will be motorized.

5.5.2. Rule-based bus detection

The outcomes of the first step were the walk, bicycle, and motorized segments. The second step analyzed motorized segments to detect bus segments.

Ambiguity between bus and car segments is a well-known challenge for mode detection. It can be addressed successfully with the use of GIS data along with actual or real-time operational data (Feng and Timmermans, 2019; Gong et al., 2012; Rasmussen et al., 2015; Semanjski et al., 2017; Stenneth et al., 2011; Nour et al., 2016; Marra et al., 2019). Bus detection in Hanoi was more complex than that in previous studies owing to four reasons as follows; (1) limited external data sources, (2) the distribution of bus stops, (3) the (fairly) frequent occurrence of bus bunching at peak times, and (4) the passing of bus stops where no boarding and alighting of passengers took place.

- *First*, the implementation of map-matching algorithms to evaluate the consistency between actual paths and bus routes (Rasmussen et al., 2015; Semanjski et al., 2017; Tsui and Shalaby, 2006) and the use of the buses' real-time or actual operational information (Marra et al., 2019; Stenneth et al., 2011) to enhance the detection were infeasible because only coordinates of stops were available.

- *Second*, to reduce the duration of bus trips, many bus stops are located near intersections so that the red-light phases can be used for boarding and alighting. However, this makes non-movement confusing as it can be owing to waiting for the traffic lights or to collecting passengers. Furthermore, bus stops are regularly distributed in front of points of interest, such as markets, universities, hospitals, and residential areas. Consequently, many non-bus segments also have either origins or destinations near bus stops.

- *Third*, over 70% of 120 bus routes provide connections to the central business districts, resulting in the overcrowding of buses there at peak hours. At a stop, there may be two, three, or even four buses in a row. Owing to bus bunching, boarding and alighting may not take place exactly at the stop, and a bus may pass the stop slowly. Figure 5-9 shows an example where three buses in a row allow passengers to get on and off at the same time. The boarding and alighting of Bus 3 are recorded quite far from the location of the bus stop. After boarding and alighting are completed, Bus 3 passes the bus stop slowly. Thus, there are no GPS points indicating that Bus 3 stopped at the bus stop.

- *Fourth*, if there are no passengers waiting to board or alight, a bus tends to ignore a stop to save time and avoid blocking the traffic behind it. However, it is still able to stop immediately to satisfy any sudden requests.

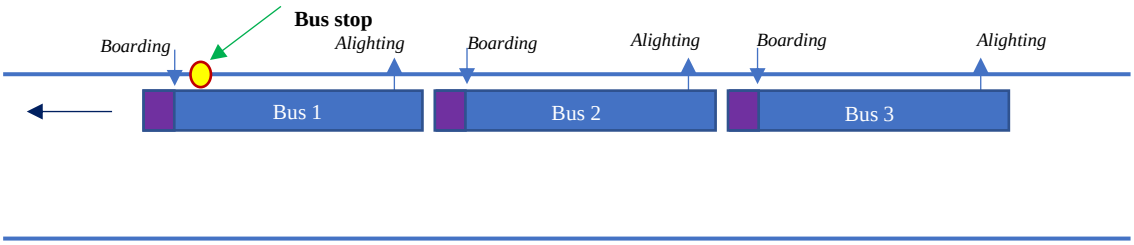


Figure 5 - 9. Example of boarding and alighting in case of bus bunching

Points two to four above show that it was unreliable to detect bus segments using the proximity of both the origin and the destination to bus stops. Points three and four showed that the stationary state of buses at stops would not be sufficient to identify the vast majority of bus segments.

The authors therefore decided to extend the approach introduced by Nour et al. (2016), which is based on the rate of stops adjacent to transit stations. First, the inverse of the average stop rate (i.e., the average distance between stops) was employed. Second, unlike Nour et al. (2016), to avoid the loss of the opportunity to detect bus segments, particularly short ones, bus stops in the proximity of intersections were retained. Third, stopping at every bus stop and passing of stops slowly were considered. Fourth, Nour et al. (2016) determined the threshold of the stop rate to detect bus segments in their data; thus, the threshold may be valid for their sample only. In this study, with stopping at and slow movement past bus stops taken into consideration, almost all stops on a segment that a bus had moved on were detected and included. The average distance between bus stops of a bus segment should be compatible with a threshold of average distance between stops on the citywide network.

The bus detection method proposed had three steps, as follows:

**** Step 1: Finding the list of stops a vehicle passed slowly or stopped at***

First, the radius from a bus stop in which to search for slow movement or stopping was defined as 100 m. To estimate the speed of slow movement, 15 bus segments were randomly selected. Each of them was plotted on a map so that all the stops the bus should stop at could be known. The speed threshold of slow movement was the median value of instantaneous speed levels of all points that were within the 100 m buffer from each bus stop. In this way, the threshold of 2.5 m/s was determined.

For each segment, to count the number of stops a vehicle passed slowly or stopped at, the distance to the nearest bus stop of each point whose instantaneous speed was under 2.5 m/s was noted. If the distance was smaller than 100 m, the corresponding bus stop was retained. The result of searching for all points of a segment was a list of bus stop candidates. In the list, duplicates were deleted. Any stop, whose distance to its previous stop in the list was under 350 m (i.e., the minimum distance between two stops on a route in the bus network), was eliminated.

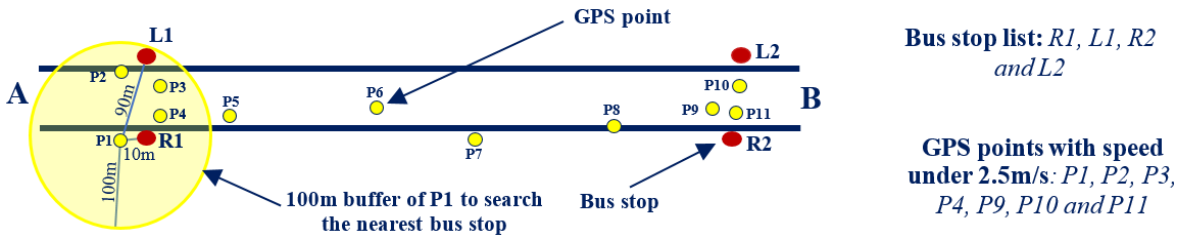


Figure 5 - 10. An example of searching stops a bus passed slowly or stopped at

Figure 5-10 shows an example in which a person on a bus that moves from A to B and either stops at or passes slowly the two stops R1 and R2. In the list of potential stops the following criteria are applied;

a. R1 is added twice because GPS points P1 and P4 have instantaneous speeds under 2.5 m/s, and the distances to R1 are the smallest (under 100 m);
b. L1 is added twice (i.e., corresponding to P2 and P3);
c. L2 is added once (i.e., corresponding to P10); and
d. R2 is added twice (i.e., corresponding to P9 and P11.

Therefore, to count the number of bus stops on the A–B segment;

a. All duplicates of L1, R1, L2, and R2 are eliminated;
b. L1 is removed because its distance to R1 is under 350 m; and
c. L2 is removed because its distance to R1 is under 350 m.

As a result of the above, the final list comprises R1 and R2.

**** Step 2: Calculating the average distance between stops***

The average distance between bus stops for a segment was calculated using Equation 5-1 below. Hereafter, "bus stop" refers to those designated boarding and alighting points that a bus passed slowly or stopped at. The length of the segment was the sum of the distances between consecutive points that belong to the segment.

$$Average\ distance\ between\ bus\ stops = \frac{Length\ of\ segment}{Number\ of\ bus\ stops - 1} \quad (5\text{-}1)$$

**** Step 3: Determining whether a segment is a bus segment***

- Estimating the threshold of distance between two consecutive stops on bus network

The task was to seek a threshold of average distance between stops on the whole bus network to determine whether a segment was travelled by bus. All distances between two consecutive stops on the same bus route and the same direction were noted.

For example, in Figure 5-10, the distances between L1 and L2 along with R1 and R2 were valid, whereas the distances of L1–R1, L2–R2, L1–R2, and L2–R1 were disregarded. To cover most cases, the 95th percentile value that is higher than 95% of the other distances was chosen. Because of the different distributions of bus stops in different areas, the 95th percentile of the distance between two consecutive stops varies. As indicated in section 3.1, Hanoi comprises two main areas. The first encompasses central business districts with a dense distribution of bus stops. In the other area, the stop distribution is sparser. To calculate the 95th percentile distance in the dense area, all distances between two consecutive stops of the same bus route were examined. The 95th percentile distance in the sparse area was calculated in the same way. The 95th percentile distance of mixed area (i.e., both areas together) was calculated by considering stops on the entire bus network. The values of the 95th percentile of distance between two consecutive stops in the dense area, mixed area, and sparse area were 1,200 m, 1,650 m, and 2,000 m, respectively.

- Giving label to segment

On the basis of the list of stops, segments can be divided into three types: dense segments, sparse segments, or mixed segments (see Figure 5-11).

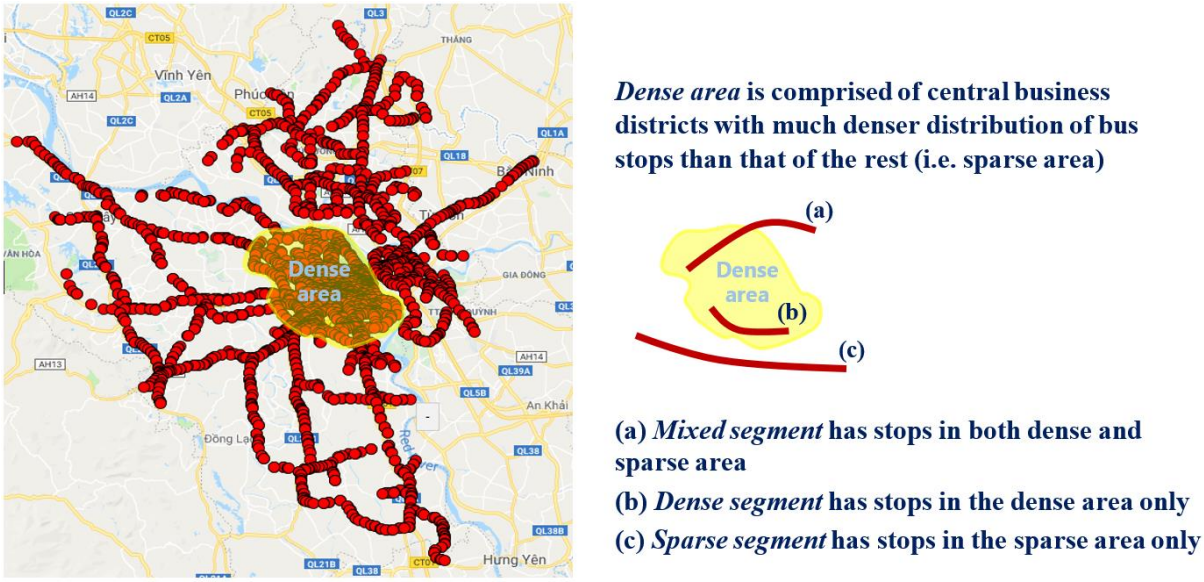


Figure 5 - 11. Distribution of bus stops in Hanoi and classification of segments based on the spatial relationship between their routes and areas

A dense segment was associated as a bus segment if the average distance between its stops (calculated by Equation 1) was smaller than the 95th percentile of the distance between two consecutive stops in the dense area (i.e., 1,200 m). If this condition was not met, it was associated as a non-bus segment. Similarly, sparse segments and mixed segments by bus were determined in the same way.

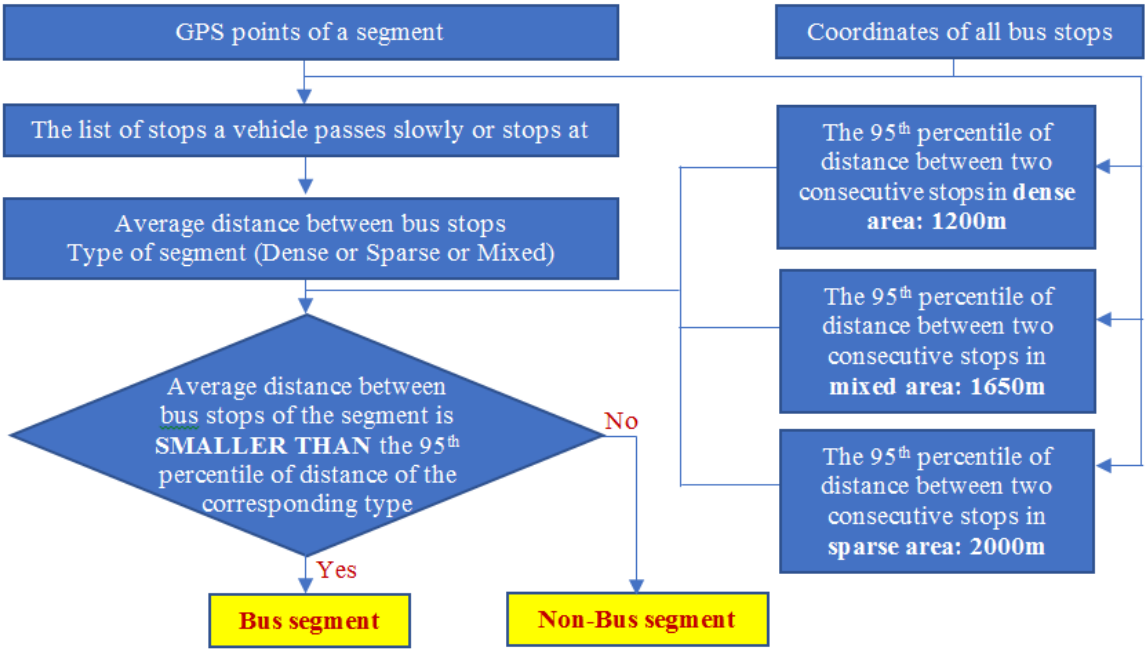


Figure 5 - 12. Flowchart of bus segment detection

5.5.3. Discrimination of motorcycle from car by Random Forest

The outcome of the second step is bus segments and car/motorcycle segments. This step concentrates on discriminating motorcycle from car.

Confusion between the car and motorcycle modes was generally significant. There was lesser confusion where the car ran at high speed or accelerated/decelerated at a high rate; however, in most

situations, they showed very similar speed and acceleration profiles owing to moving in urban areas. To distinguish between them, the model must use many variables, which was not suitable for deterministic and probabilistic methods. Car—and especially motorcycle—segments accounted for significant percentages of the sample (see Figure 5-4), so a learning-based mode was developed.

**** Random Forest***

Random Forest (RF) is a standard and favorite non-parametric prediction tool first introduced by Breiman, (2001). It is an ensemble of numerous decision trees, thus to master RF, understanding of Decision Tree (DT) is the prerequisite.

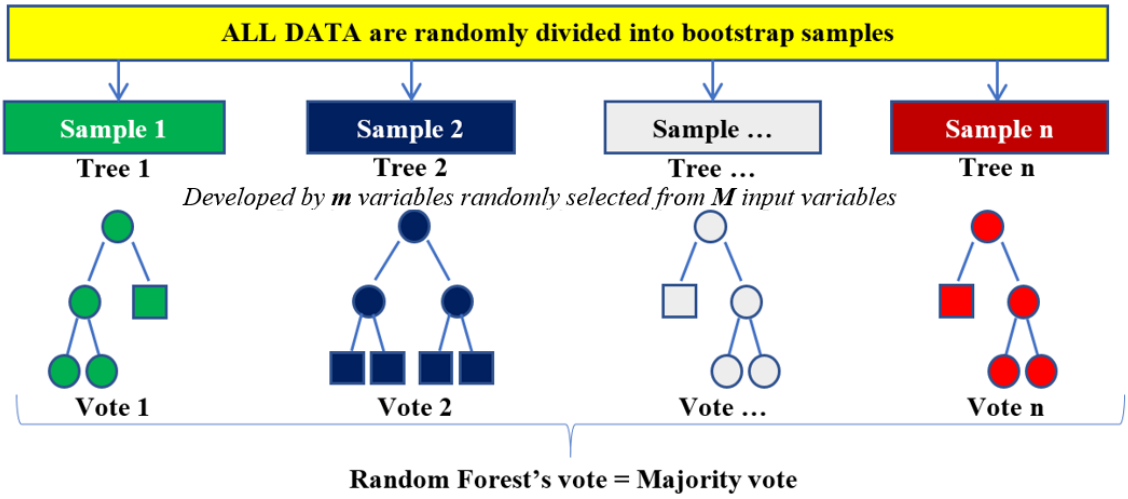


Figure 5 - 13. Random Forest's structure

DT is a type of supervised learning algorithm that mimics the top-down human thinking process in a simple way by a series of True/False questions in concert. Basic concepts in DT are, as follows:

- *Root node* represents entire sample that is further divided into homogeneous sets. Root node lies at the top of DT.

- *Terminal node/ End node/ Leaf* cannot be split due to encompassing a homogeneous set of a particular class. Leaf lies at the bottom of DT.

- *Parent node* is divided into nodes called (its) *child nodes*. Child nodes that share the same parent node are called *sibling nodes*.

- *Splitting* is a procedure of dividing a (parent) node into two (child) nodes. Splitting process continues until the homogeneous sets are found or stopping criteria are satisfied. Common criteria are the maximum length of the longest path from a root to a leaf (i.e. tree depth), the maximum number of observations in either a leaf or a node.

- *Internal node* lies between a root node and leaves. It represents a scenario based on a dichotomous criterion. Each *branch* represents the outcome of the test (True or False). Each leaf node represents a class. The path from the root to a leaf node represents a classification rule. For example: If *medium speed* $\leq$ 3m/s and *max speed* $\leq$ 3.9m/s, then the class is *walk* (see Figure 5.14).

DT is categorized based on the type of target variables, that is, categorical DT and continuous DT. For travel mode detection, the type used is categorical DT.

A DT model is constructed based on splitting parent nodes into child nodes until finding out terminal nodes. This is the process of breaking groups into smaller groups until reaching the homogeneous groups.

Table 5 - 5. Example of mode segments classified by decision tree in Figure 5-14

ID segment	Distance (km)	Max speed (m/s)	Median speed (m/s)	Mode
1	0.5	3	1.5	walk
2	0.7	3.5	1.8	walk
3	0.3	4.5	2.5	walk
4	1.1	2.5	1.1	walk
5	5.5	10	6	car
6	9.8	20	12	car
7	2.5	16	9	car
8	7.4	9	5	car
9	3	11	5.5	motorcycle
10	8	7	4	motorcycle
11	5	15	8.5	motorcycle
12	1.5	6	3.5	bike
13	1.1	4.3	2.2	bike
14	3.2	8	4	bike

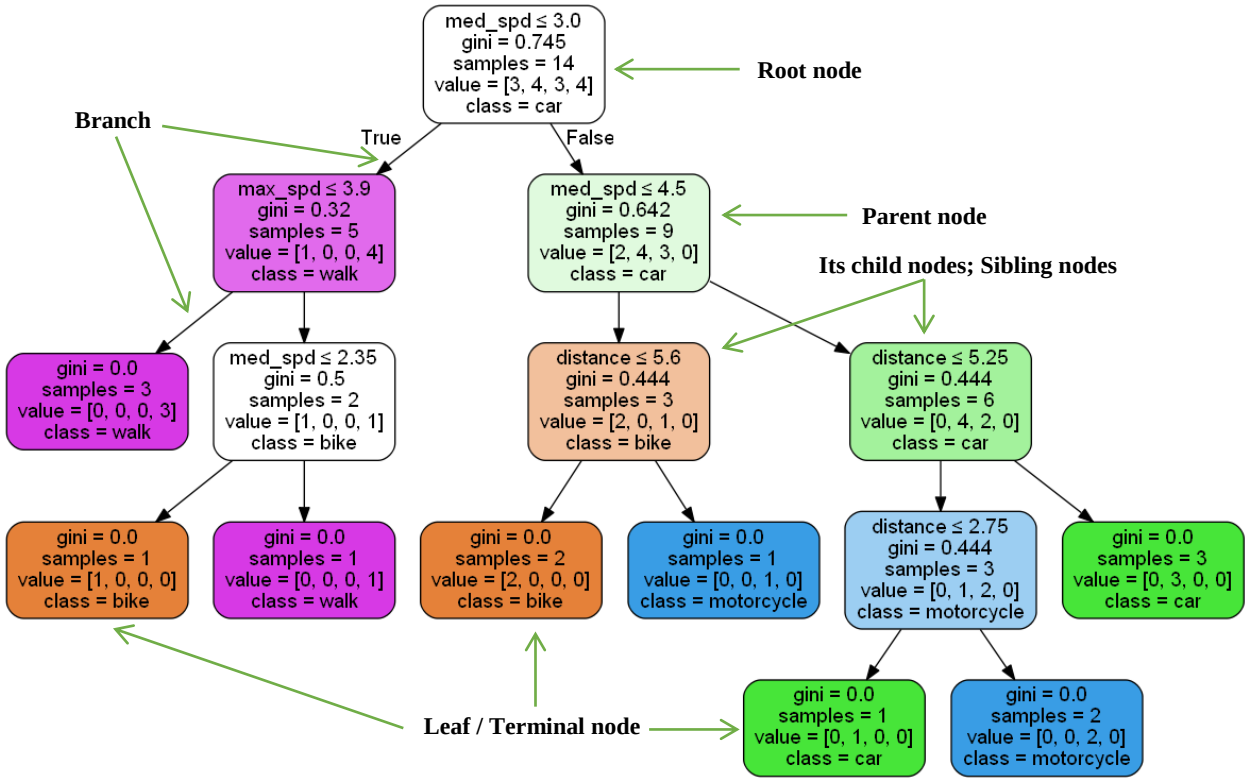


Figure 5 - 14. Visualization of the decision tree model classifying mode segments in Table 5-5

To describe and explain DT, we made an example of classifying modes by distance (km), max speed (m/s) and median speed (m/s) (see Table 5-5 and Figure 5-14). It is important to note that data here are not data from Hanoi survey. At the root node, the whole sample of 14 segments (3 bike, 4 car, 3 motorcycle, 4 walk) is checked by the criterion of median speed to be partitioned into (1) segments with this speed equal to or under 3m/s and (2) segments with this speed over 3m/s. Then the two new sub-samples are sequentially divided until homogeneous sets are achieved.

Whilst the computer iterates through scenarios respective to criteria of variables, it gauges Gini coefficient to determine the best variable for the split. Gini coefficient (or Gini index or Gini impurity) is a metric of how often a randomly chosen observation from the set would be falsely classified in case it is randomly classified according to the distribution of classes in the subset. Gini coefficient is called Gini impurity because it measures the degree of impurity of the resultant child noes. Its values range between 0 and 1. The 0 level means all observations at the node belong to a particular class or the node includes one class only, and thus further splitting is the worst. The 1 level means observations at the node are randomly distributed across different classes, and thus further splitting is the best. The variable with the lowest Gini coefficient is chosen as the root because it is the strongest predictive variable with the lowest probability of incorrect classification. If Gini coefficient is over 0, more splitting are undertaken until a node is completely pure with a zero Gini coefficient. In case the criterion splitting is valid, at each node, the variable that results in the greatest reduction in Gini coefficient is chosen. The weighted total Gini coefficient decreases as moving down the tree.

$$I_G(n) = 1 - \sum_{i=1}^{J}(p_i)^2 \quad (5\text{-}2)$$

Where n is the note considered

J is a set of classes

i is a particular class belonging to J

p is the probability of an object being misclassified to a particular class

Considering the example in Figure 5-14, the root node' Gini coefficient of 0.745 means there is a 74.5% chance of falsely classifying an observation chosen randomly based on the distribution of modes in the sample. The value of 0.745 is calculated as follows:

$$I_{root} = 1 - \left(\left(\frac{3}{14}\right)^2 + \left(\frac{4}{14}\right)^2 + \left(\frac{3}{14}\right)^2 + \left(\frac{4}{14}\right)^2\right) = 0.745$$

By similar way, the Gini indexes of the root's left node and the root's right node are 0.32 and 0.642, respectively. The weighted Gini coefficient at second level (0.527) is smaller than that of the first level (i.e. the root node's Gini value).

$$I_{left\ child\ node} = 1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{0}{5}\right)^2 + \left(\frac{0}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right) = 0.32$$

$$I_{right\ child\ node} = 1 - \left(\left(\frac{2}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{3}{9}\right)^2 + \left(\frac{0}{9}\right)^2\right) = 0.642$$

$$I_{second\ layer} = \frac{N_{left}}{N_{parent}} * I_{left\ child\ node} + \frac{N_{right}}{N_{parent}} * I_{right\ child\ node} = \frac{5}{14} * 0.32 + \frac{9}{14} * 0.642 = 0.527$$

The DT's biggest advantage is understandable and self-explanatory in case of being compacted. Besides, it is a non-parametric method that does not require any assumptions about classifier structure. The input variables are both nominal and numeric. However, DT tends to become large and over-complex to fit closely particularities of training data but fails to generalize the data well. DT has high variance, small variations in datasets lead to new trees with completely different structures (Song and Lu, 2015). As a result, it performs poorly on unseen/novel datasets. This is called overfitting problem, which is a common pitfall but it is difficult to avoid or handle effectively it.

Overfitting in DT can be alleviated by two main techniques. The first is pruning that is the opposite process of splitting and conducted by removing nodes containing much less additional information than others. By this way, the complexity of tree is lessened and thus the magnitude of concentration on

particularities is reduced. However, to achieve both higher accuracy and better avoid overfitting, development of RF is the favorite way (Statnikov et al., 2008).

RF is an ensemble of numerous decision trees with two key concepts, that is, (1) random sampling of training data whilst building trees and (2) random subsets of features whilst splitting nodes, which enables decreasing the variance without increasing the bias

Specifically, during the training process, multiple random sub-samples of observations from the training data, with replacement are drawn to build trees. Because each tree learns from different data, they are fairly uncorrelated with others. Besides, each tree can learn unique characteristics of classes from the sub-sample, so their combination can create a more stable and stronger classifier.

Randomly selecting sub-set of variables in splitting nodes is towards de-correlating trees. In each split of a tree, RF considers a small number of variables rather than all. This is important because there are usually a few stronger variables than others. If all trees are under its influence, they will possibly have similar structures and thus highly correlated.

As for evaluating a RF model trained, for creating each tree in previous steps a bootstrapped data set is used by removing some observations and duplicating some of others. So, there are the observations that do not include in the training set of the tree. They are called out-of-bag (OOB) dataset. Each observation of the OOB set will be passed all the trees that do not contain it in their training sets. The class given to the observation by RF is the most nominated one (i.e. receiving the most votes from trees). A RF model is evaluated by the prediction error rate (i.e. the rate of wrongly classified OBB observations).

For another choice, the whole training set can be divided into two sub-sets, that is, training set and validation set. The training set is actually used to train RF whilst the validation set is used to validate the RF developed by the error rate. The main difference between the OBB error rate and the validation error rate is related to the trees used to generate the RF predictions. Whilst the OBB error rate is estimated on a subset of trees not encompassing the OBB sample in their bootstrapped training set, the validation error rate is calculated by all trees of RF and these trees have not ever seen the validation data.

All in all, by means of the randomness in both selecting data and selecting variables to train RF's trees together with the voting mechanism, RF avoids well overfitting to generate satisfactory prediction results.

**** Classifying car and motorcycle by Random Forest***

Random forest models have been documented in a number of mode detection studies (Gong et al., 2018; Marra et al., 2019; Stenneth et al., 2011). In this study, a random forest model was implemented using Python and the scikit-learn library. In addition to the 95th percentile of speed, median speed, and average absolute acceleration, variables including the 95th percentile of absolute acceleration and segment length were used as inputs to the model. To train the model, 75% of the 1,245 motorcycle segments and 75% of the 587 car segments were used. Thus, the data used to evaluate the hierarchical process comprised 758 walk segments, 104 bicycle segments, 97 bus segments, 311 motorcycle segments, and 147 car segments.

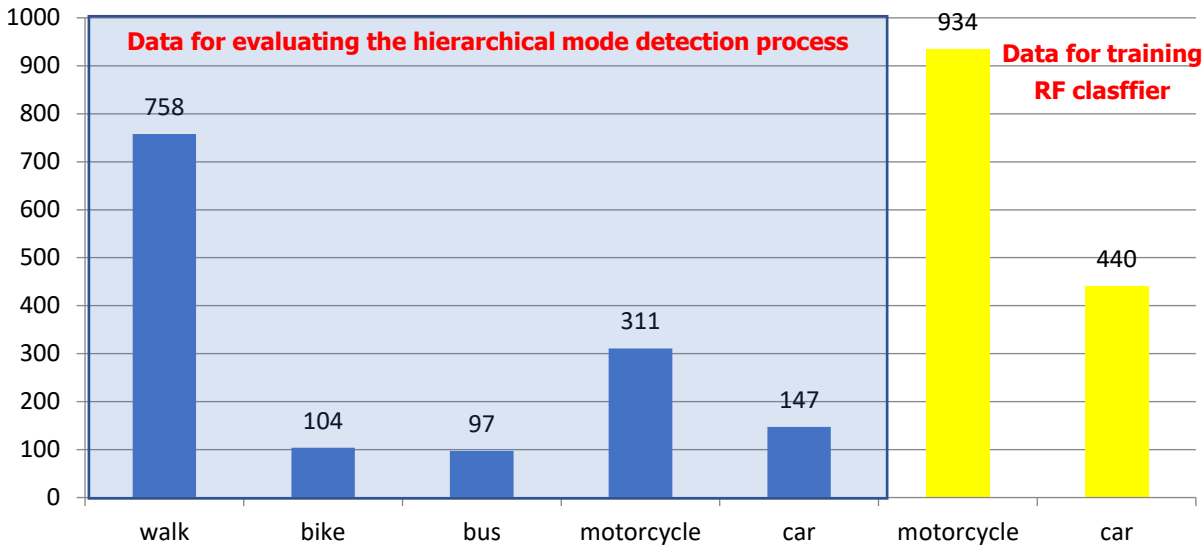


Figure 5 - 15. Data for evaluating the hierarchical process adopted

5.5.4. Assessing the mode classification results

Mode detection is a multi-class classification problem. To evaluate a classifier, the most often used measures are *precision*, *recall*, *accuracy* and *Fscore*.

Precision is the rate between the number of cases correctly predicted as mode *i* and the total number of cases predicted as mode *i* by the model.

Recall is the rate between the number of cases correctly predicted as mode *i* and the total number of cases actually being mode *i* in the data.

F-score of predicting mode *i* is the harmonic mean of the corresponding precision and the corresponding recall.

Accuracy is the rate between the number of cases correctly predicted and the total number of cases in data.

$$precision_i = \frac{the\ number\ of\ segments\ correctly\ classified\ as\ mode\ i}{the\ number\ of\ segments\ classified\ as\ mode\ i} \quad (5\text{-}3)$$

$$recall_i = \frac{the\ number\ of\ segments\ correctly\ classified\ as\ mode\ i}{the\ number\ of\ segments\ actually\ being\ mode\ i} \quad (5\text{-}4)$$

$$Fscore_i = \frac{2 * precision_i * recall_i}{precision_i + recall_i} \quad (5\text{-}5)$$

$$accuracy = \frac{the\ number\ of\ segments\ correctly\ classified}{the\ total\ number\ of\ segments} \quad (5\text{-}6)$$

Among measures, precision and recall are calculated for modes and thus evaluate the prediction ability to each mode. It is better to combine them instead of use one solely to make evaluation. A model that gains high precision but low recall is not good because it omits many true cases of mode *i*. By contrast, a model that achieves high recall but low precision is also bad in that in order to detect almost all true cases of mode *i* it accepts to misclassify cases of other modes as mode *i*. A good model should have high and balanced levels of both precision and recall. Precision and recall are usually presented in concert by a confusion matrix, also known as error matrix and being a special kind of contingency table with imputed cases of modes in columns and true cases of modes in rows.

Given that considering two models, comparing their capacities of detecting a particular mode by precision and recall simultaneously would be complex. For example, model 1 has higher precision but

model 2 has higher recall. Fscore fits this case because it is formed equally by both precision and recall. The better model should have higher Fscore.

Different from three mentioned-above metrics, accuracy is the critical one to evaluate the model's prediction results and thus comparing a model with others.

Values of four measures range between 0 and 1. The closer to one they are, the better prediction.

5.6. RESULTS AND DISCUSSIONS

5.6.1. Identifying walk, bike and motorized modes

**** Rule-based method***

Applying rule-based method for 2791 segments shows the overall accuracy of 87%, which demonstrates that the model addressed the classification problem quite well. Yet, it is necessary how well it classified particular modes.

Table 5 - 6. Results of fuzz logic-based and rule-based models to detect walk, bike and motorized modes

		Predicted				**Total**	**Recall**	**Fscore**
		Method	*Walk*	*Bike*	*Motorized*			
Reported	*Walk*	Fuzzy logic	717	18	23	758	*94.6%*	*96.3%*
		Rule-based	632	109	17	758	*83.4%*	*89.1%*
	Bike	Fuzzy logic	4	79	21	104	*76.0%*	*75.6%*
		Rule-based	0	60	44	104	*57.7%*	*27.5%*
	Motorized	Fuzzy logic	10	8	537	555	*96.8%*	*94.5%*
		Rule-based	28	164	1737	1929	*90.0%*	*93.2%*
Total		Fuzzy logic	731	105	581	1417	-	-
		Rule-based	660	333	1798	2791	-	-
Precision		Fuzzy logic	*98.1%*	*75.2%*	*92.4%*	-	**Accuracy**	**94.1%**
		Rule-based	*95.8%*	*18.0%*	*96.6%*	-	**Accuracy**	**87.0%**

Motorized segments gained the highest recall of 90% followed by walk with the 83% recall. By contrast, only 58% of bike segments were correctly classified. The precision levels of walk and motorized modes were over 95%; yet, the figure for bike was very low at only 18%. Hence, the high overall accuracy of the model came from motorized and walk that made up the largest percentage in the mode share whilst bike had to be sacrificed its number of correct classification.

The greatest confusion can be seen between motorized and the bike mode. 42% of bike segments were misclassified as motorized ones. This is quite strange because rule-based method is powerful enough to obtain high success in discriminating motorized modes from non-motorized modes (Safi et al., 2016). The possible reason is that traffic in Hanoi is mixed and congestion fairly frequently takes place in some areas at peak time, leading speed of motorized means and the bike mode to be close together. Second, motorized travel of Hanoi citizens depends heavily upon the mode of motorcycles that is not a significant mode in developed countries, thus not included in mode detection list. The contribution of motorcycle will be analyzed more rigorously in Sub-section 5.6.3.

**** Fuzzy logic method***

As indicated in the last paragraph of the Sub-section 5.5.3, fuzzy logic model ran on data of 1417 segments because 1245 motorized segment of the whole 2791 were kept separately to train RF.

Compared with the rule-based method, fuzzy logic performed clearly better with overall accuracy of 94%. Because the accuracy levels will decrease in the following steps, thus the 94% accuracy of fuzzy logic method allowed us to dream of the final accuracy for 5-mode classification at approximately 90%.

For walk, fuzzy logic method generated high levels of both precision and recall, that is, 98.1% and 94.6%, respectively.

Classification of bike segments outperformed that by rule-based model with the precision and recall being around 75%.

When it comes to motorized modes, rule-based model achieved higher precision (96.6% vs. 92.4%) but lower recall (90% vs. 96.8%), leading its Fscore of 93.2% to be slightly smaller than 94.5% of the fuzzy logic model. The recall of motorized mode is extremely important because it had great impact on the classification of bus, car and motorcycle segments. With only 18 out of 555 motorized segments being wrongly detected, the motorized mode prediction outcome was promising.

Accordingly, the fuzzy logic method obviously fitted the classification of walk, bike and motorized modes in Hanoi much more than deterministic rules.

5.6.2. Identifying bus

Performance of the rule-based classifier was far worse than that of the fuzzy logic model, thus motorized segments found by the probabilistic model were processed to identify bus.

Table 5 - 7. Results of rule-based method to detect bus segments

		Predicted				**Total**	**Recall**	**Fscore**
		Walk	*Bike*	*Bus*	*Car/Motorcycle*			
Reported	Walk	717	18	3	20	758	*94.6%*	*96.3%*
	Bike	4	79	2	19	104	*76.0%*	*75.6%*
	Bus	0	0	92	5	97	*94.8%*	*87.2%*
	Car/Motorcycle	10	8	17	423	458	*92.4%*	*91.5%*
Total		731	105	114	467	1417		
Precision		*98.1%*	*75.2%*	*80.7%*	*90.6%*		**Accuracy**	**92.5%**

As can be seen in Table 5-7, the high recall level of 94.8% showed that bus-detection rules identified almost all of the actual bus segments (92 of 97 cases). However, a number of other segments were falsely detected as bus segments—possibly owing to taking passing of bus stops slowly into consideration—resulting in a precision level of 80.7% corresponding to the bus. Despite this limitation, however, these precision and recall levels were comparable with those of studies using real-time information or high-quality GIS data (Feng and Timmermans, 2019; Gong et al., 2018; Marra et al., 2019; Rasmussen et al., 2015; Semanjski et al., 2017; Tsui and Shalaby, 2006). In comparison with the model of Nour et al. (2016), the bus-detection method of this study achieved a much higher recall (94.8% vs. 84.7%) and thereby a larger F-score (87% vs. 84%).

The accuracy of this step decreased from 94.1 to 92.5% and we kept hoping that the eventual accuracy can gain the 90% target.

Bus identification was greatly affected by both the threshold of distance to search for the nearest bus stop and the threshold of speed representing the slow movement. The sensitivity of the inferences when

(1) fixing the speed at 2.5 m/s coupled with changing the distance, and (2) fixing the distance at 100 m coupled with changing the speed were tested, respectively.

**** Testing the sensitivity of bus detection results to the distance thresholds***

We determined the ranges of distance based on those deployed before. The minimum level of 30m was reported by (Nour et al., 2016), 75m in (Gong et al., 2012) for bus detection. For metro segment detection, the distance levels were over 150m (Bohte and Maat, 2009; Patterson and Fitzsimmons, 2016). We did not believe that a very high level of searching distance was effective thus we set up the maximum distance of this test at 150m.

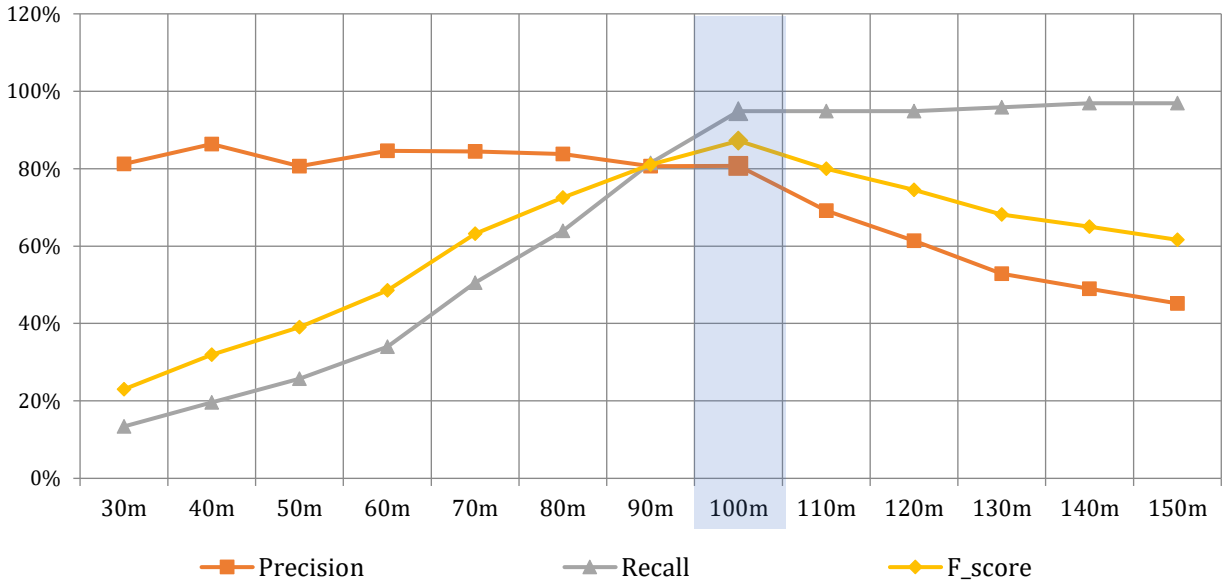


Figure 5 - 16. Sensitivity in case the distance changes and the speed fixes

In the case of a 2.5 m/s speed threshold (Figure 5-16), at the smallest level of 30 m, the ability to detect bus segments was unsatisfactory, with a recall of only 18% yielded. Between 30 m and 100 m, the higher the distance was, the higher the recall was. This emphasized that buses would not stop exactly at bus stops. In this range, the precision values nearly levelled out. Between 110 m and 150 m, the increase in recall levels was insignificant, and the precision decreased dramatically with distance. In fact, at 100 m, the recall reached the near-maximum level of approximately 95%, with 92 of the 97 bus segments being correctly detected. The changes in precision and recall were reflected by changes in the F-score. The highest F-score was achieved at 100 m, from which we concluded that 100 m was the best distance threshold when the speed was 2.5 m/s. The 100 m distance is suitable for the 2.5 m/s speed possibly because a bus tended to slow down before each bus stop.

**** Testing the sensitivity of bus detection results to the speed thresholds***

By the same way, we measured how bus detection is sensitive to changes in the speed threshold in case of fixing the searching distance at 100m.

We set up the ranges of speed between 0.7m/s and 4m/s to test the bus identification performances. Very low speed level refers to the stable status of bus. The 0.7m/s level was employed in (Nour et al., 2016). Figure 5-17 reveals that the consideration of stopping (i.e., very low speeds, such as 0.7 m/s) did not lead to accurate detection of bus segments, with a recall of 63% yielded. As the speed increased from 2 m/s to 2.7m/s, more bus segments were successfully recognized, albeit with nearly unchanged precision. From 2.9 m/s, precision dropped once recall was nearly stable, leading to a decrease in the F-score. The F-score reached its maximum value of 87.2% for a speed of 2.5 m/s.

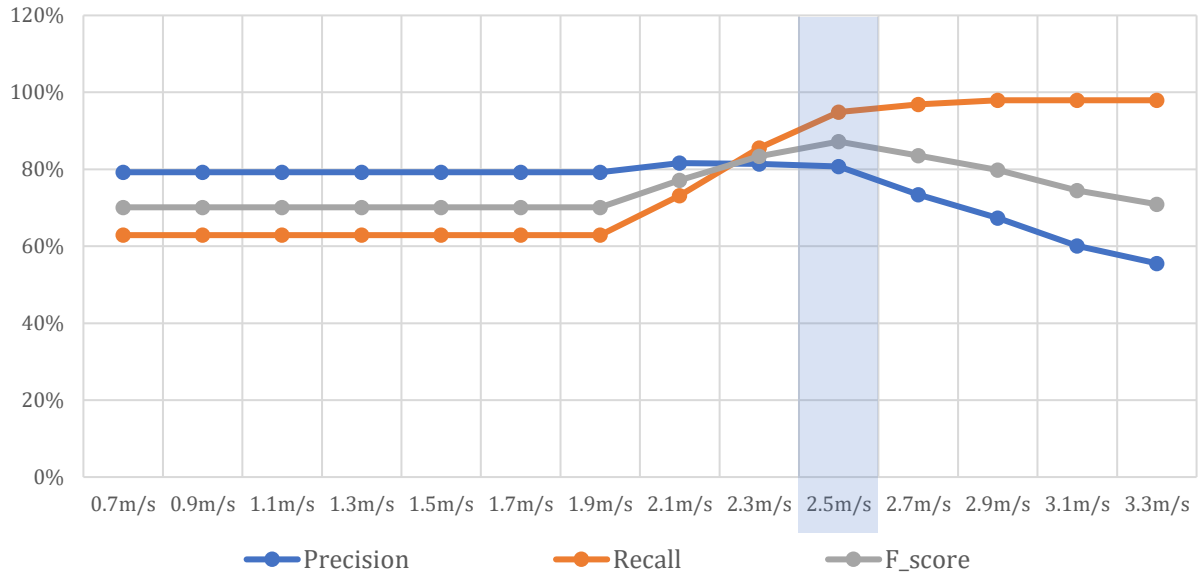


Figure 5 - 17. Sensitivity in case the speed changes and the distance fixes

These analyses have demonstrated that it was appropriate to consider stopping at and passing bus stops slowly together with the thresholds chosen (2.5 m/s and 100 m).

5.6.3. Identifying car and motorcycle

As can be seen in Table 5-8, 89.1% of total segments were successfully inferred, which slightly worse than our expected level of 90%. However, the 89.1% level is comparable to those of many previous studies constructing either all-in-one or hierarchical processes (Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Marra et al., 2019; Nour et al., 2016; Rasmussen et al., 2015; Stenneth et al., 2011; Tsui and Shalaby, 2006; Xiao et al., 2015).

The confusion matrix of the hierarchical process gave us interesting points related to the abilities to detect modes.

Motorcycles were shown to be the primary source of misclassification. Of 118 segments labeled as cars, 13 were actually motorcycle. Apart from walk, motorcycle was the only mode misclassified as bicycle, (8 segments). Nineteen bicycle segments (18% of the total bicycle segments) were falsely classified as motorcycle segments. The ambiguity of the motorcycle mode with other modes reduced the correct number of inferences, which was reflected in its precision level of only 77.7%, about 10% lower than its recall level of 87.1%.

Table 5 - 8. Confusion matrix of mode detection generated by the hierarchical process adopted

		Predicted						***Recall***	**Fscore**
		Walk	Bike	Bus	Motorcycle	Car	*Total*		
Reported	Walk	717	18	3	20	0	*758*	***94.6%***	***96.3%***
	Bike	4	79	2	19	0	*104*	***76.0%***	***75.6%***
	Bus	0	0	92	4	1	*97*	***94.8%***	***87.2%***
	Motorcycle	7	8	12	271	13	*311*	***87.1%***	***82.1%***
	Car	3	0	5	35	104	*147*	***70.7%***	***78.5%***
	Total	*731*	*105*	*114*	*349*	*118*	*1417*	-	-
Precision		***98.1%***	***75.2%***	***80.7%***	***77.7%***	***88.1%***	-	**Accuracy**	**89.1%**

Car segments were also confused with motorcycle segments. Of the 147 car segments, 35 were labeled as motorcycle segments, resulting in a recall of 70.7%, the lowest value for all the modes. However, the confusion between car and bus, despite being a well-known issue in literature, was minor.

5.6.4. Comparing the hierarchical process adopted with other processes

A simple hierarchical process and an all-in-one process (see Table 4) were developed to assess the performance of the proposed hierarchical process. The former used rules related to speed and distance to the nearest bus stops. The latter was based on a random forest model.

The accuracy level of the simple hierarchical process was low at 61.3%, compared with 79.1% for the all-in-one process and 89.1% for the process proposed in this study. The poor performance of the rule-based process was anticipated because such a process was too simple to deal with the challenge of classifying five modes.

The all-in-one process had very low F-score values for bus and bicycle segments because it had a significant bias against modes having minor percentages of the dataset. In other words, the all-in-one process failed to deal with the imbalanced data problem. In contrast, the process proposed in this work generated comparable F-score values for all modes, and these values were higher than those of the random forest model.

The F-score values of bus segments for the rule-based process (0.17) and the random forest-based process (0.16) were much lower than that of the proposed process (0.87), which highlighted how much bus detection could be improved by considering the average distance between bus stops that a bus passed slowly or stopped at. Furthermore, the rule-based and random forest processes did not detect cars and motorcycles as well as the hierarchical process proposed here did, which emphasized the confusion between motorcycles and cars.

Table 5 - 9. Description and prediction results of simple hierarchical process and all-in-one process

Process	**Description**				**F-score**	**Accuracy**
RULE-BASED (SIMPLE HIERARCHICAL)	*95th percentile of speed*	*Median speed*	*Proximity to bus stops*	*Mode*	*Walk: 0.89* *Bike: 0.27* *Bus: 0.17* *Motorcycle: 0.62* *Car: 0.64*	*61.3%*
Step 1	< 3.5	< 2.0	-	Walk		
Step 2	< 6.0	< 4.0	-	Bike		
Step 3	< 15.0	≥ 3.5	Yes	Bus		
Step 4	> 12.0	≥ 6.0	-	Car		
Step 5	The remainder of segments			Motorcycle		
RANDOM FOREST (ALL-IN-ONE)	*Features:* 95th percentile of speed, median speed, proximity to bus stops (0 if no and 1 if yes), heading change rate, low speed rate, 95th percentile of acceleration, average (absolute) acceleration. *Splitting data*: at the rate of 75% vs. 25%				*Walk: 0.93* *Bike: 0.25* *Bus: 0.16* *Motorcycle: 0.80* *Car: 0.69*	*79.1%*

Note: - Proximity to bus stop refers to the distances from both origin and destination of a segment to the nearest stops within 75m.

- Detailed results can be seen in Appendix 4.

5.7. SUMMARY

Making transportation mode inferences has been a common goal of GPS-data-based research owing to the absence of trip characteristics in logs. This study has addressed a difficult challenge, with the inclusion of motorcycles, a major travel mode, in data collected in Hanoi.

First, a hierarchical process was developed to classify walk, bicycle, and motorized modes using a fuzzy logic algorithm. In addition to acceleration and speed specific variables, heading change rate was used to enhance the classifier's power. Bus segments were then distinguished from other motorized segments upon extension of the work of Nour et al. (2016). Specifically, an average distance between stops at which the bus passed slowly or stopped at was compared with those estimated from the bus network. The advantages of this method are that it could detect almost all of the bus segments by the coordinates of stops only, and this was easily understandable because it originated from the actual operation of bus services. To limit other modes being misclassified as bus, it was necessary to carefully choose thresholds of speed and distance. The distance should be determined before the speed is estimated from the sample. Finally, a random forest model was developed to detect motorcycle and car segments.

The proposed hierarchical process performed well, with an accuracy of 89.1%. The main source of confusion was the mode of motorcycle. The most frequent ambiguities were between motorcycles and cars, not between cars and buses. This is typical not only for Hanoi but also for cities in a number of developing countries where travel depends heavily on two-wheeled motorized vehicles. This study was an effort to extend the list of modes and the geographical scope of research into imputing travel modes from GPS data.

Although thoroughly developed, the process was validated only on a sample that was biased toward persons working and studying at a university in Hanoi. Thus, the inferences would, to some extent, benefit from the homogeneity of travel patterns of the participants. Moreover, the Hanoi urban transport system does not include any metro lines at present. These limitations emphasize the need to adapt and test the process using a more diverse sample with more travel options. Future research could be conducted to enhance identification of the motorcycle mode.

Chapter 6: ENHANCING PURPOSE IMPUTATION WITHOUT USING GIS DATA

6.1. INTRODUCTION

In the previous chapter, travel modes are imputed from GPS data in Hanoi. To complete the travel profile, this chapter aims at developing a trip purpose detection model.

It contributes to the literature by enhancing purpose detection with no use of GIS data, which is distinct from the vast majority of earlier studies. It connects with Chapter 5 by using mode prediction results as variables to detect trip purposes. Notably, this idea is in line with processing GPS data in (Shen and Stopher, 2014b) but has never been experimented before. New feature(s) related to users' personal locations (i.e. of residence and non-residence frequently visited), which are able to be extracted from multi-day GPS data, are tested.

As for the structure, the following is the synthesis of existing purpose identification works to emphasize gaps which this chapter makes an effort to address. Section 6.3 describes data preparation and purpose list determination whilst Section 6.4 presents a method to develop a RF model and optimize its hyper-parameters along with variables used. Predictions results and their interpretation are content of Section 6.5. The last Section summarizes this chapter.

6.2. SYNTHESIZING PURPOSE IMPUTATION STUDIES

In the past, collection of mobility data had been dependent entirely on conventional techniques like paper and pen, (e-)mail and computer-assisted telephone interviewing (Armoogum et al., 2014). Unfortunately, self-reported survey methods are subject to lack of reliability due to the human memory limit and the habit of rounding time (Forrest and Pearson, 2005; Kelly et al., 2013) not to mention low response rate due to the high burden on participants and big time gaps between periodic national or regional travel household investigations (Chen et al., 2016; Ortúzar et al., 2011). Owing to the discontinuation of the restricted use of Global Positioning System (GPS) for non-military goals, GPS has brought about a technological revolution in travel surveys (Auld et al., 2009; Shen and Stopher, 2014b). Its data that are accurate and detailed in terms of both time and space are collected continuously and passively during periods of many days, even of several years (Bohte and Maat, 2009; Nguyen et al., 2019a; Thomas et al., 2018; Zheng et al., 2010). More and more GPS-based surveys have been carried out thanks to the ubiquitous spread of smartphones (Cottrill et al., 2013; Patterson and Fitzsimmons, 2016; Thomas et al., 2018; Xiao et al., 2016). However, positioning data are not comprised of purpose information that is one among the most important trip characteristics and essential for travel behavior research. This deficiency has induced the investment of many scientific efforts (Bohte and Maat, 2009; Chen et al., 2010; Cui et al., 2018; Feng and Timmermans, 2015; McGowen and McNally, 2007; Montini et al., 2014; Oliveira et al., 2014; Reumers et al., 2013; P. Stopher et al., 2008a; Wolf et al., 2001; Xiao et al., 2016; Yazdizadeh et al., 2019) in the field of trip purpose inference from GPS data.

Depending on GIS data (i.e. land use and points of interest) is the key to inferring well activity types in almost all earlier studies (Bohte and Maat, 2009; Chen et al., 2010; Cui et al., 2018; Feng and Timmermans, 2015; McGowen and McNally, 2007; Oliveira et al., 2014; P. Stopher et al., 2008a; Wolf et al., 2001; Xiao et al., 2016; Yazdizadeh et al., 2019), except for (Montini et al., 2014; Reumers et al., 2013). This is such a research gap because the reliance has prevented researchers from implementing purpose

prediction in areas inadequately geo-coded. In this study, we aim at extending (Reumers et al., 2013) by deploying familiar features related to time, user, and actual travel mode. Besides, we adopted and tested novel features related to mode predicted and frequently visited places extracted from multi-day GPS data.

The literature is synthesized according to three main aspects, including algorithm, feature and purpose list. The last paragraphs are devoted to reviewing Reumers et al., (2013) that has inspired this chapter and Montini et al., (2014) that classify purposes without the use of GIS data.

Regarding algorithms, pre-defined ad-hoc rules were developed in the fancy (Bohte and Maat, 2009; P. Stopher et al., 2008a; Wolf et al., 2001). They functioned well on a test by data of 13 participants (Wolf et al., 2001); however, performing poorly with larger-scale data. They correctly predicted approximately 60% of activities of 66 respondents in (P. Stopher et al., 2008a) and 43% of 1104 participants' data in (Bohte and Maat, 2009). Chen et al., (2010) achieved the accuracy of under 80% by a multinomial logit model. After 2010, studies focused on creating supervised learning models that were flexible enough to infer activity types from big spatial data. Tree-based methods generally have been preferred. Decision tree succeeded in classifying around 75% of cases (McGowen and McNally, 2007; Oliveira et al., 2014; Reumers et al., 2013). Recently, Random Forest (RF), an improvement of decision tree, has been the most heavily used (Feng and Timmermans, 2015; Montini et al., 2014; Yazdizadeh et al., 2019) and achieved the highest accuracy level of 96.8% (Feng and Timmermans, 2015) in the literature. Enhanced versions of the neural network by particle swarm optimization (Xiao et al., 2016) and determining parameters between nodes of different layers based on bayesian probability distribution (Cui et al., 2018) generated the accuracy levels of over 90%. Other methods (i.e. bayesian network, naïve bayes, support vector machine, artificial neural network with back propagation, K-nearest neighbors) were formed to generate baseline levels to emphasize the superiority of main models adopted (Cui et al., 2018; Feng and Timmermans, 2015; Xiao et al., 2016). Once developing learning-based models, authors usually tailored values of hyper-parameters that are involved in configuration of algorithms and specified before training process begins (Cui et al., 2018; Montini et al., 2014; Xiao et al., 2016). Typical examples are the number of trees in the forest (Montini et al., 2014) and the number of hidden neurons (Xiao et al., 2016).

Regarding features, there are four main groups related to GIS data, activity, trip and participant. Apart from (Montini et al., 2014; Reumers et al., 2013), the rest (Bohte and Maat, 2009; Chen et al., 2010; Cui et al., 2018; Feng and Timmermans, 2015; McGowen and McNally, 2007; Oliveira et al., 2014; P. Stopher et al., 2008a; Wolf et al., 2001; Xiao et al., 2016; Yazdizadeh et al., 2019) depended upon being aware of the potential locations of a trip destination to determine its purpose in well geo-coded areas like Zurich (Switzerland), Canberra and Sydney (Australia), New York (US), Shanghai (China). Time indicators (e.g. start time, duration and day of week) are typical for activities and thus widely employed (Chen et al., 2010; Feng and Timmermans, 2015; McGowen and McNally, 2007; Montini et al., 2014; Oliveira et al., 2014; P. Stopher et al., 2008a; Wolf et al., 2001; Xiao et al., 2016; Yazdizadeh et al., 2019). Transportation mode-related features have been useful for predictions (Feng and Timmermans, 2015; Montini et al., 2014; Oliveira et al., 2014; Xiao et al., 2016; Yazdizadeh et al., 2019). They were actual mode extracted directly from ground truth data. Notably, purpose imputation is the subsequent step of mode detection in processing GPS data (Shen and Stopher, 2014b). In this sense, mode predicted rather than actual mode should be used to identify trip purpose. User information has been worthy features (Bohte and Maat, 2009; Cui et al., 2018; McGowen and McNally, 2007; Montini et al., 2014; Oliveira et al., 2014; Shen and Stopher, 2014b; Wolf et al., 2001; Xiao et al., 2016; Yazdizadeh et al., 2019); yet, respondents would be unwilling to provide their personal locations (Feng and Timmermans, 2015).

Regarding purpose lists, they obviously varied across the studies. Whilst 11 types were classified in (Feng and Timmermans, 2015), only 5 were included in (McGowen and McNally, 2007). Generally, the size of 6 purposes were primarily preferred (Cui et al., 2018; Reumers et al., 2013; Xiao et al., 2016; Yazdizadeh et al., 2019). Non-home and non-work purposes can be combined in different ways and thus the same terminologies could have dissimilar interpretations in studies. For example, "visit" was a part of "recreation" in (Cui et al., 2018) but "visit" played as an independent type together with "recreation" in (Bohte and Maat, 2009; Feng and Timmermans, 2015). Interestingly, if addresses of work and school were not available, "work" and "education" were grouped into one category (Chen et al., 2010; McGowen and McNally, 2007; Xiao et al., 2016). Types making up minor percentages were ignored. In short, the component and details of a purpose list have been decided flexibly based on the availability of input information and the proportions of each purpose.

Differing from other authors, Reumers et al., (2013) inferred activities in a very limited condition. They employed start time and duration to identify 6 classes, which has been interesting because their method was independent of GIS data and thus transferable. Only classifications of work and home labels were satisfactory. Although the precision levels of the rest (e.g. shopping, social and bring-get) were low at around 40%, their algorithm achieved the 76% accuracy level thanks to greatly being biased towards majority classes (i.e. home and work) dominating over minority classes (i.e. other purposes). In this sense, the accuracy did not tell the complete truth of the model's power and performance. Detecting minor classes should be enhanced.

Similar to Reumers et al., (2013), Montini et al., (2014) predicted purposes without the use of GIS-related variables. However, the 2014 work's emphasis was on segments instead of trips. Mode transfer that is a purpose of a trip segment rather than of traveling between two significant locations was considered. This purpose accounted for 34.9% of the 1403-activity test set[5], much higher than home (26.7%) and work/education (12.6%). Mainly because of the nearly perfect identification of mode transfer with the recall of 99%, their model reached an accuracy of 84.4%. Work and home were detected satisfactorily thanks to the availability of their addresses. Prediction of minority classes (e.g. recreation, business, pick-up/drop-off) was limited with recall levels of between 20% and 40%, which was in line with Reumers et al., (2013).

This chapter focuses on extending the work of Reumers et al., (2013) because our attention is to trip purpose along with the absence of home's and work's addresses in data.

6.3. PREPARING DATA AND DETERMINING PURPOSE LIST

6.3.1. Data preparation

In data sent by Trivector, GPS trajectories were partitioned into consecutive activity and travel segments (see Table 6-1 and Figure 6-1 for example). For activities, the center longitudes and latitudes that had been estimated by in-built algorithms of TravelVU were provided in the dataset. We used these activities' coordinates to make further analyses. TravelVU's algorithms were beyond the research scope.

Similar to preparing data for detecting travel modes, all data outside Hanoi were removed. As a result, we had 63-person data of 652 days, which was equivalent to approximately 10.4 valid days per user on average.

[5] Estimated by us based on confusion matrix (table 1) in (Montini et al., 2014)

Table 6 - 1. Example of all segments/trips of a person during a day

Phone ID	Segment ID	Start time	Type	Label given	Duration (min)	Center latitude	Center longitude	Trip ID
653166392	5460770	2019-04-08 08:19:25 Monday	TRAVEL	WALK	1.783333333			1
653166392	5460771	2019-04-08 08:21:12 Monday	TRAVEL	MOTORCYCLE	43.01666667			1
653166392	5460772	2019-04-08 09:04:13 Monday	ACTIVITY	BUSINESS	27.25	21.0214166	105.8561609	
653166392	5460773	2019-04-08 09:31:28 Monday	TRAVEL	MOTORCYCLE	7.866666667			2
653166392	5460774	2019-04-08 09:39:20 Monday	ACTIVITY	BUSINESS	43.76666667	21.0103409	105.8493492	
653166392	5460775	2019-04-08 10:23:06 Monday	TRAVEL	MOTORCYCLE	11.25			3
653166392	5460778	2019-04-08 10:34:21 Monday	TRAVEL	WALK	2.3			3
653166392	5460779	2019-04-08 10:36:39 Monday	ACTIVITY	WORK	350.0666667	20.9800737	105.8420233	
653166392	5461808	2019-04-08 16:26:43 Monday	TRAVEL	MOTORCYCLE	30.08333333			4
653166392	5461809	2019-04-08 16:56:48 Monday	ACTIVITY	PICK-UP	13.21666667	20.9719314	105.7873996	
653166392	5462053	2019-04-08 17:10:01 Monday	TRAVEL	MOTORCYCLE	2.966666667			5
653166392	5462054	2019-04-08 17:12:59 Monday	ACTIVITY	PICK-UP	7.433333333	20.9719382	105.7846888	
653166392	5462055	2019-04-08 17:20:25 Monday	TRAVEL	MOTORCYCLE	6.916666667			6
653166392	5462056	2019-04-08 17:27:20 Monday	ACTIVITY	HOME	826.4666667	20.9709348	105.7951602	

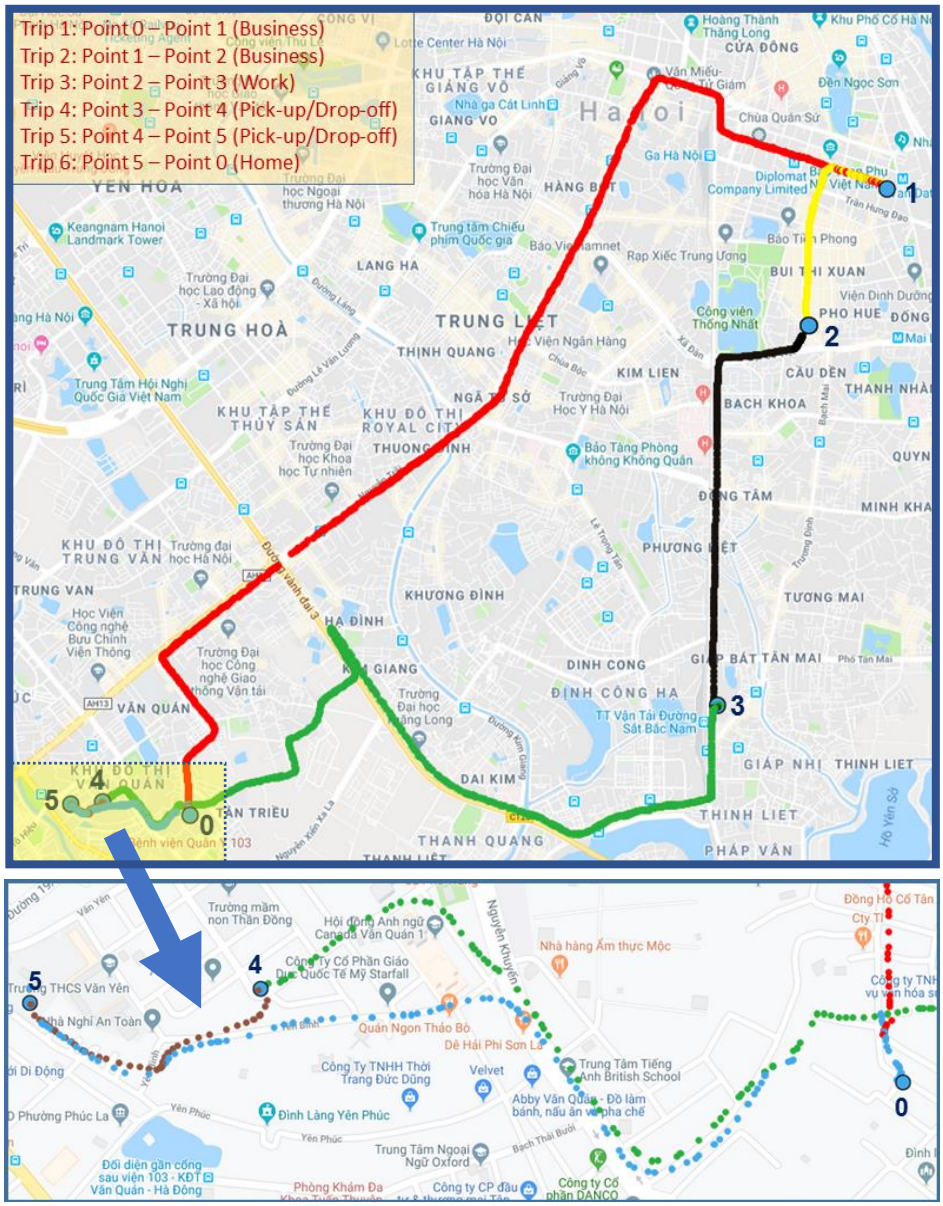


Figure 6 - 1. Map version of the day tour indicated in Table 6 - 1

In that the app was sensitive to the change between movement and non-movement, an activity at a location was usually split into many segments. An example can be seen in Figure 6-2 where a person travelled and occurred several times at different blocks of a university. In this case, all consecutive segments within the university perimeter were merged into one activity with the accumulated duration and the average coordinate.

Because we did not have polygon-shaped GIS data, we carefully visualized all consecutive activity segments annotated identically to decide whether to join them or not. This step was vital because it determined activities' time profiles.

Figure 6 - 2. Example of merging activities in the same geographical location
A user reported three "work" activities at P1, P2, P3 (blue circles) that were within the area of University of Transport and Communications from 14h50 before leaving from P3 at 17h20. Therefore, P1, P2 and P3 were merged into one activity P (red star) that lasted from 14h50 to 17h20 and whose coordinate was an average of those of P1, P2 and P3

6.3.2. Determining trip purpose

The purpose list of this study was comprised of (1) *home*, (2) *work/education*, (3) *shopping/eating*, (4) *pick-up/drop-off*, (5) *visit/leisure*, (6) *business*.

- Work and educational activities were similar with significantly longer duration than non-home purposes and the high frequency of occurring in the morning. Moreover, we did not have work and school addresses; therefore, grouping them was acceptable.

- Shopping and eating out were arranged into one category because having breakfast at street vendors, diners in tandem with going (daily) shopping in the morning to buy food for all day is a common habit in Hanoi. Besides, some could spend parts of days to do shopping and have lunch or drink coffee at department stores and malls.

- Pick-up and drop-off were grouped since they usually lasted during a short time and took place at specific time windows.

- Business acted as a type because some users were sale staff or repairers who usually from their offices to meet customers or go to places to maintain or repair machines. It usually occurred within the normal working hours albeit with a shorter duration than that of work/education type.

- Visit and leisure activities varied in time. This group included activities users labeled as entertainment, visit and hobby.

All activities that either did not belong to the six chosen types or lasted for smaller than 60s were disregarded. Eventually, 2596 activities were eligible for purpose imputation. Among them, *home* was the dominant activity type with the number of 935 accounting for 36%. Its figure and share exceeded the counterparts of the second most common purpose that is *shopping/eating* with 454 cases (17.5%) but not *work/education*. It may be surprising; yet, having a logic with the occupational characteristic of the sample. Specifically, 27% of participants having dynamic jobs (see Sub-section 3.1.3) likely labeled their working activities outside their main workplace as *business*, which was reflected by the *business* type's share of 13.3%. If grouping *work/education* with *business*, its share would have been high at 28.7%. The high share of *shopping/eating* could be explained that inhabitants in Hanoi have the habit of shopping for meals every day rather than shopping once for some days. Additionally, they usually have breakfast outside home not to mention fairly frequently having small suppers after finishing working/educational time on street vendors or small restaurants. The percentage of *pick-up/drop-off* was at 6% with 157 activities, smaller than all other purposes' figures; however, we believed that they would be sufficient for a construction of a learning-based classifier.

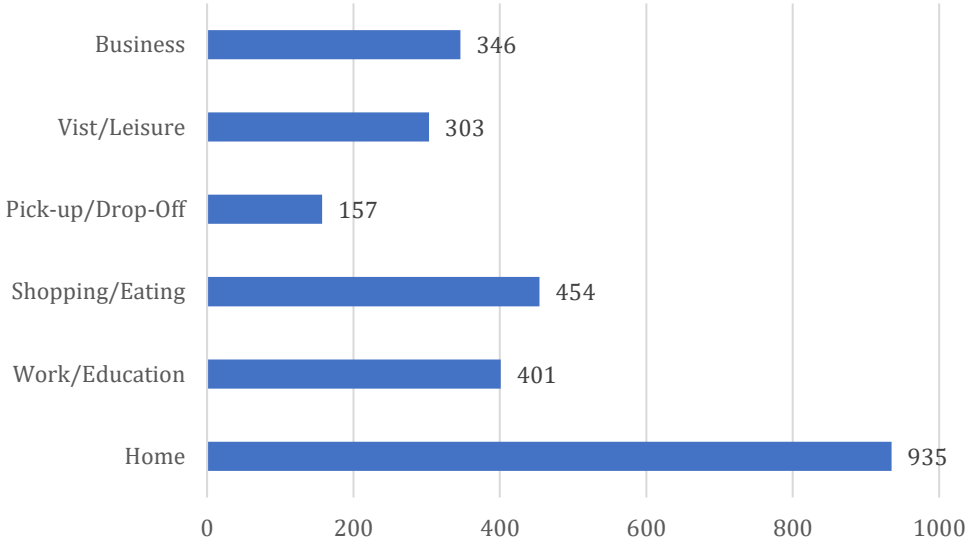


Figure 6 - 3. Numbers of activities

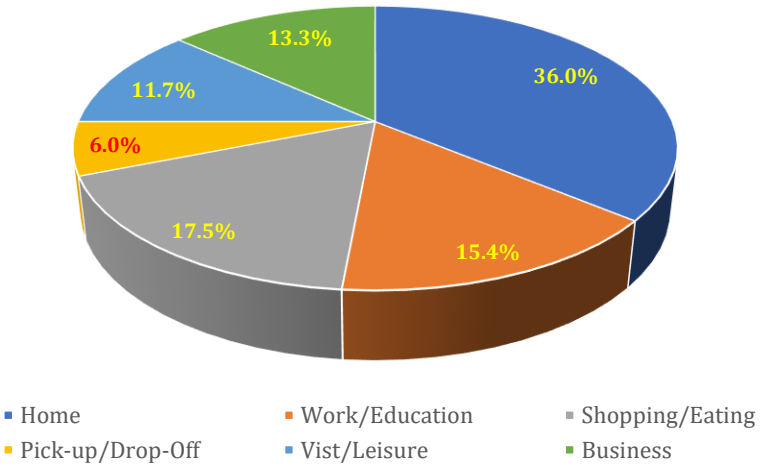


Figure 6 - 4. Shares of activities

6.3.3. Activities' time profiles

We considered two time-related attributes of purposes, that is, duration and start time (see appendix 6 and appendix 7 for detailed information).

As can be seen in Figure 6-5, *working/education* had the peak time in the morning between 7:00-10:00 whilst *home* occurred the most in the evening, after 17:00, explained by the fact users came back home after working/educational time. 16:00-18:00 was the busiest time of picking up or dropping off some ones. About 60% of *pick-up/drop-off* activities were made during this period. *Shopping/eating* type had three peak times per day, that is, 7:00-9:00, 11:00-13:00 and 17:00-20:00. Similarly, *business* activities were more frequently recorded in 8:00-11:00 and 14:00-17:00. The peak time of *visit/leisure* kind was in the evening from 18:00 to 22:00.

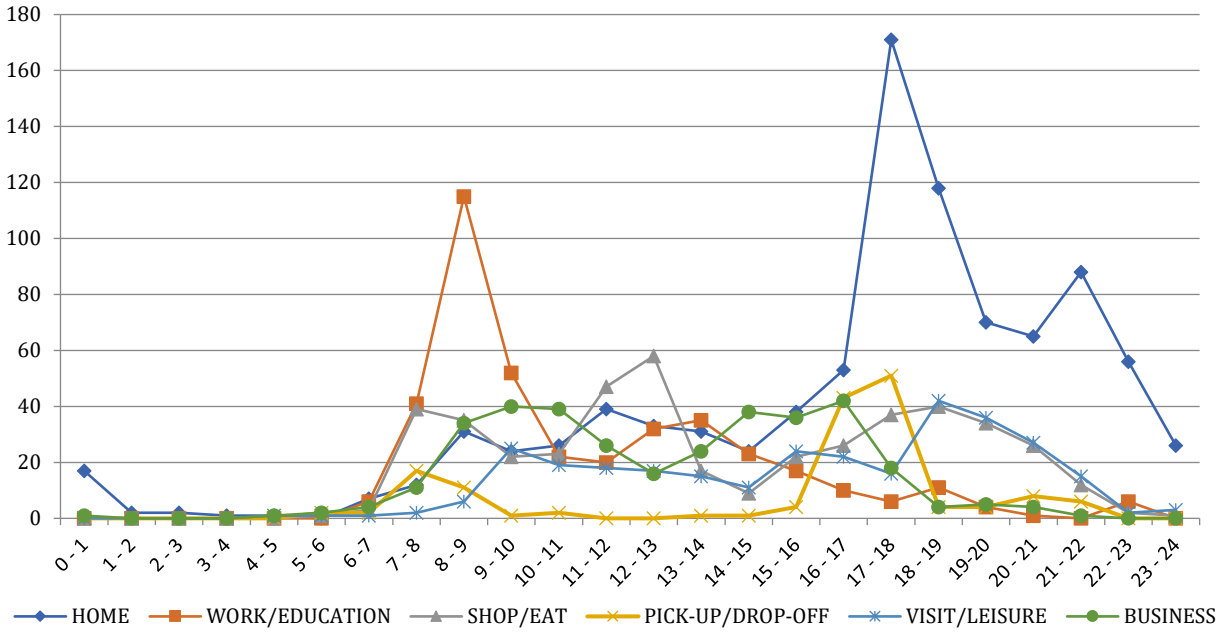


Figure 6 - 5. Numbers of activities based on start time

As regards duration in Figure 6-6, about half *home* activities lasted for over 12 hours whilst vast the majority of *working/education* ones' duration ranged from 2 to 4.5 hours or from 8 to 10 hours. The duration around 3-4 hours pertained to the fact user started working in the morning and stopping at noon or (re-)started working afternoon (e.g. 13h) and finishing at about 17:00. The duration of from 8 to 9.5 hours would correspond to the fact smartphones were kept inside the workplace from starting time in the morning (e.g. 8:00) to the finishing time (e.g. 17h). A number of *visit/leisure* activities had duration from 2 to 4 hours. Generally, activities whose duration that fell into these two ranges (i.e. 2 to 4.5 hours and from 8 to 9.5 hours) would be either *work* or *home* or partially *visit/leisure*, and thus misclassification between them was unavoidable. Almost all *business* activities lasted for under 2 hours whereas a few *work/education* activities had a duration of less than 2 hours. Therefore, *business* and *work/education* can be effectively differentiated from each other by duration. The ambiguity between *business*, *visit/leisure*, *shopping/eating* and *pick-up/drop-off* activities was considerable because they were short ones. Notably, approximately 90% of *pick-up/drop-off* activities lasted within 30 minutes.

By the above analyses, duration seems to be more useful for detecting purposes than start time. In case both of them were used, home and work were expected to be identified quite well thanks to their difference in the peak of start time compared with together and their difference in duration compared with

other purposes. However, the level of misclassification between *home* and *work/education* would be significant. It was difficult to achieve a high success of detecting other activity types.

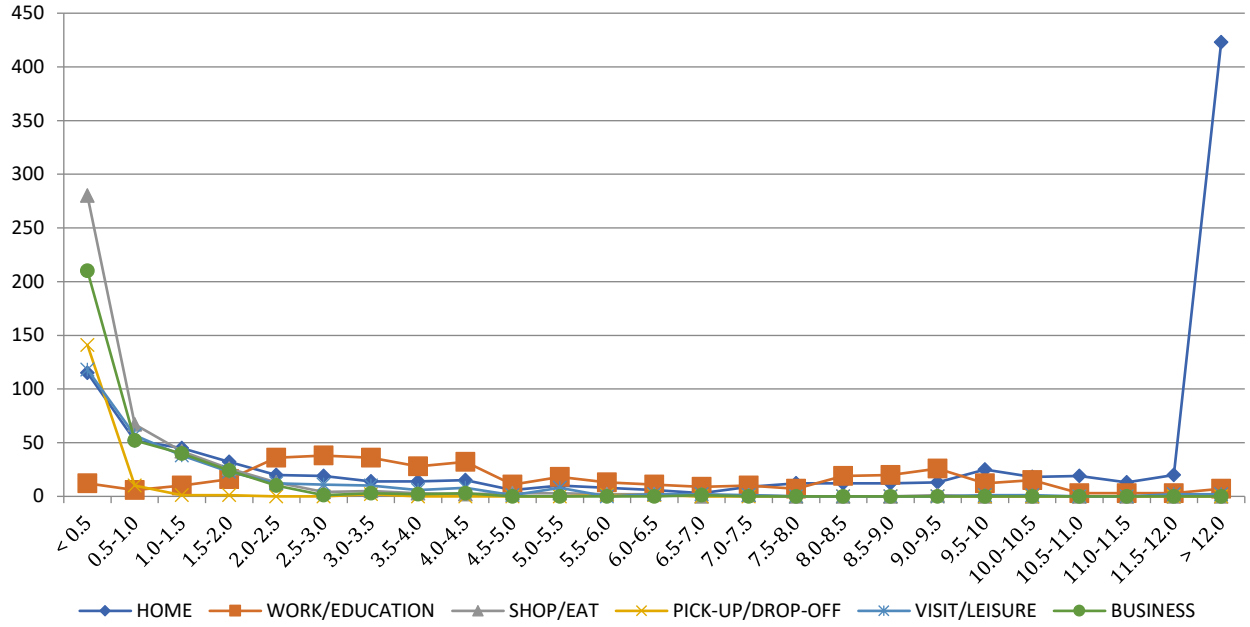


Figure 6 - 6. Numbers of activities based on duration

6.4. METHOD

6.4.1. Random Forest and tuning its hyper-parameters

We chose Random Forest (RF) for this chapter because it has been one of the most pragmatic approaches to purpose imputation (Feng and Timmermans, 2015; Montini et al., 2014; Yazdizadeh et al., 2019). Being a standard and favorite non-parametric prediction tool first introduced in (Breiman, 2001), it is an ensemble of numerous decision trees and operates based on the randomness. All trees learn from samples selected randomly from the original data with replacement (i.e. bootstrapping). To split each node of a decision tree, a subset of input features randomly withdrawn from all features is used. To respond a classification problem, decision trees' votes are aggregated to form the final prediction decision of RF. Thanks to the randomness and the voting mechanism, RF avoids better overfitting, at least than decision tree.

RF is capable of estimating the importance of features; however, it tends to inflate magnitude of continuous variables and categorical variables with more categories than others (Strobl et al., 2007). Therefore, it is suitable for choosing relevant features only. To evaluate objectively and comprehensively both the contribution of features and the power of predictive models, we optimized model settings based on the corresponding features used. This is supported by the conclusion of (Probst et al., 2019) that emphasized that RF is characterized by a number of hyper-parameters (HPs) whose values are dependent largely on the input data thus needs to be tuned carefully case by case to enhance the predictive ability.

We utilized *Python* and *Scikit-Learn library* to tailor HPs. Considered HPs are as follows.

- The first is the number of features used for each split. Its optimal value is primarily influenced by the number of relevant features. If the number is large, it should be determined small to cover the less influential features useful for detecting minor classes. By contrast, it should be set high to find out the most relevant ones (Goldstein et al., 2011).

- The second is the number of trees in a forest. More trees would contribute to better performance at the expense of more costly computation. Its optimal value depends on other HPs and the dataset (Probst et al., 2019).

- The third is the maximum number of splits until the leaf, thus specifies the maximum (vertical) depth of a tree. A higher depth allows a tree to learn better data; however, possibly making tree focus on particularities rather than generalizing data.

- The fourth is the minimum number of observations in a leaf that is a terminal node in a decision tree. A smaller end node may lead the model to be more likely to fit noise. By contrast, determining too high threshold can lead the model to fail to capture the underlying trend of the data, frequently called as the under-fitting problem.

- The last is the minimum number of observations in a node to cause further split.

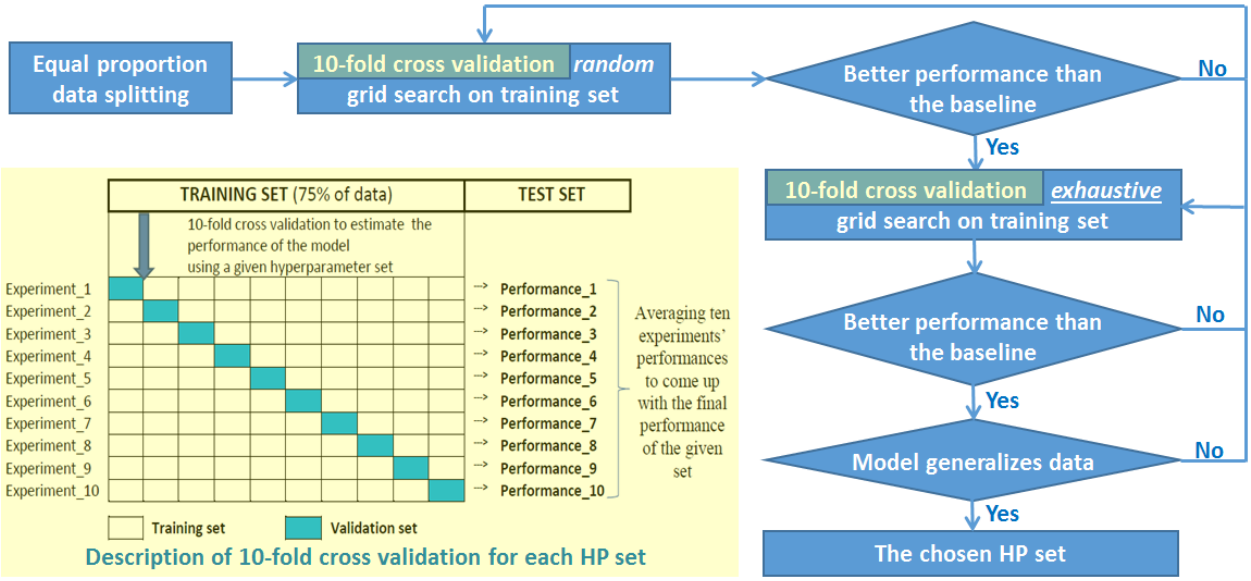


Figure 6 - 7. Flowchart of tuning hyper-parameters

Figure 6-7 shows the tuning process, first data were divided into training and test sets by the ratio of 75% and 25% of total data of each purpose, respectively. Afterwards, ten-fold cross validation random grid search was carried out on training set. Regarding ten-fold cross validation, the training set was further split into ten folds. A model specified by a HP set candidate was iteratively fitted 10 times, each time it was trained by data of nine folds and evaluated by the rest, called validation set. At the end of the 10-time iteration procedure, we averaged 10 performances conducted on 10 different validation sets to generate the score of the HP set. Performance metric selected was the accuracy level. Regarding random grid search, we defined a range of values for each HP based on experience and the literature (Probst et al., 2019). The combination of values created the grid. HP sets were randomly chosen from the grid. Performing 10-fold cross validation with sets enabled to find out the best whose score was the highest. The best set was used for the model that next was trained by training set before evaluated by the test set. In case the test performance was better than that produced by default HP values in the *Scikit-Learn library*, 10-fold cross validation exhaustive grid search was undertaken. For each HP in this step, its values were determined around the best found in random search. The search exhaustively ran across all HP sets possibly taken from the grid. Once the model based on the best set found in exhaustive search showed higher accuracy on test set than that of the baseline model, we next considered its generalization. Whilst its accuracy on training set was far better than that on test set, it was neglected due to the serious overfitting problem. For real data, overfitting is unavoidable but should be limited, for example (Feng and Timmermans, 2019) reported the

5.1% difference between training and test accuracy levels. Here, the maximum accepted level of 6% was determined. In case a search failed to give a good set in terms of generalization and/or comparison with the baseline, the grid was modified. If changes in exhaustive grid could not bring about expected outcome, random search was re-conducted to seek new potential ranges.

6.4.2. Feature selection

**** Time-related features***

Two time-related indicators of activities that were duration and start time were used as continuous variables. As analyzed in Sub-section 6.3.3, they represented unique characteristics of some purpose types (e.g. *home* and *work/education*); therefore, being useful for the inference model.

**** Participant-related features***

Gender, job, age and child in household were used as binary variables to limit the model complexity, and thus the risk of overfitting. The variable named "job dynamic" was typical for whether frequently working outside main office or not. This was expected to be fruitful for identifying business activities well.

**** Transportation mode-related features***

Travel mode information used here was the main mode of a trip. A trip may encompass many segments with different modes. If at least one segment of a trip is by bus, the mode of trip is bus. If at least one segment is by car and no segment is bus, the mode is car. If a trip is not by car and bus and has at least one segment by motorcycle, the mode is motorcycle. The mode is bicycle if the trip is not motorized and there is at least one segment by bicycle. The mode is walk if all segments are on foot. For example, trip 1 in Table 6-1 had two travel segments by walk and motorcycle, thus its mode was motorcycle.

We considered two types of mode information.

- The first that was extracted directly from ground truth has been widely used previously (Feng and Timmermans, 2015; Montini et al., 2014; Oliveira et al., 2014; Xiao et al., 2016; Yazdizadeh et al., 2019).

- The second that has never been employed before was results of the hierarchical mode prediction developed and presented in the previous chapter.

The hierarchical model classified correctly 81% of total trips.

**** Home-related feature***

It was a binary one relying on the spatial relationship between an activity position and home location. We estimated the home location of a user by considering the origin of the first trip and the destination of the last trip in each day. Travel day in mobility survey is not the normal day lasting from 0:00 to 24:00. It should be characterized in order to collect very early and very late trip of a participant to complete tours. Therefore, it is a pre-assigned 24-hour period from 4:00 of a day through 4:00 of the following day.

The first trip of the first day and the last trip of the last day were ignored because the installation and/or uninstallation would be implemented out of home. We calculated the frequency for each of origins and destinations. The frequency of staying at a point was the total number of other points that were close to it. Closeness meant the distance between two points being smaller than 100 meters. Finally, the coordinate of home was the location having the highest frequency. Its coordinate was an average of coordinates of points adjacent to it.

**** The most frequently visited non-home place-related feature***

It was a binary one relying on the spatial relationship between an activity position and location of a place that was not home and more frequently visited than other places. To estimate its coordinate, we removed all home points before counting the visit frequency of each of the remainder by the identical way applied for home location estimate. The location with the highest frequency was selected and its coordinate was an average of those of points close to it.

The return to previously visited locations with regularity is common in human patterns (Hasan et al., 2013). The most frequently visited non-home places would vary among persons. They may be workplace, school or market. The exact answer can be found by observing effects of this feature on prediction performances.

Table 6 - 2. Description of features

Group	Feature	Description
Time	Start time	Difference in minutes between midnight and the start of the activity
	Duration	Duration of the activity in minutes
Participant	Gender	1: Male; 0:Female
	Age	1: Up to 22 years old; 0: Over 22 years old
	Job	1: (Self-)Employed or student or pupil; 0:Otherwise
	Child in household	1: At least 1; 0: Otherwise
	Dynamic job	1: Frequently working outside the main office; 0: Otherwise
Actual mode	Walk	1: Trip made on foot. 0: Otherwise
	Bike	1: Trip made by bike. 0: Otherwise
	Bus	1: Trip made by bus. 0: Otherwise
	Motorcycle	1: Trip made by motorcycle. 0: Otherwise
	Car	1: Trip made by car. 0: Otherwise
Predicted mode	Walk	1: Trip predicted on foot. 0: Otherwise
	Bike	1: Trip predicted by bike. 0: Otherwise
	Bus	1: Trip predicted by bus. 0: Otherwise
	Motorcycle	1: Trip predicted by motorcycle. 0: Otherwise
	Car	1: Trip predicted by car. 0: Otherwise
Close to home		1: Distance between the activity location and home under 150m; 0: Otherwise
Close to the most frequently visited non-home place		1: Distance between the activity location and it under 150m; 0: Otherwise

6.4.3. Assessment metrics

Similar to travel mode detection, purpose imputation is a type of classification problem; therefore, metrics including precision, recall, accuracy and Fscore can be used. They would be useful tools to evaluate models using different variables and the same data. However, to compare a model adopted in this chapter with that in the literature, they are insufficient. Precision and recall are for evaluating how well individual purposes are detected but not a model is. Accuracy is under great impacts of major purposes. For example, in data that includes 90% home activities the model can reach a 90% accuracy by inferring all activities in data as home. Therefore, accuracy should be used with the macro Fscore for the model, which is estimated by the average Fscore values of purposes (see Equation 6-1).

$$F^{model} = \frac{\sum_{i=1}^{n} F^{purpose\ i}}{n} \quad (6\text{-}1)$$

6.5. RESULTS AND DISCUSSIONS

We constructed 6 models corresponding to 6 different scenarios of using features (see Table 6-3). For all models, the difference in accuracy levels on training and test sets were around 4%-6%, meaning that overfitting was fairly well controlled. Table 6-4 shows an example where *model_6* that was input by all features except for mode predicted had the training accuracy level of 81% whilst the test one was lower at 75%.

Table 6 - 3. Comparison of models using different features

Model	**Features**						**Performance**	
	Time	*User*	*Predicted mode*	*Actual model*	*Home*	*The most frequently visited non-home place*	*Accuracy*	F^{model}
model_1	Yes	-	-	-	-	-	60.1%	56.0%
model_2	Yes	Yes	-	-	-	-	62.4%	58.1%
model_3	Yes	Yes	Yes	-	-	-	63.6%	60.1%
model_4	Yes	Yes	-	Yes	-	-	64.3%	61.7%
model_5	Yes	Yes	-	Yes	Yes	-	73.0%	66.0%
model_6	Yes	Yes	-	Yes	Yes	Yes	75.0%	68.8%
(Reumers et al., 2013)*	Yes	-	-	-	-	-	76.0%	49.6%
* F^{model} was estimated by us based on F^{class} reported in (Reumers et al., 2013)								

As can be seen in Table 6-3, 60% of all activities in the test set were successfully identified by start time and duration solely *(model_1)*. The addition of user's information *(model_2)* induced the 2.3% improvement. Interestingly, adding either actual *(model_3)* or predicted *(model_4)* mode-related features led to better performance albeit with a stronger effect from mode extracted from ground truth, that is 64,3% and 63.6%, respectively. By means of using the home closeness feature, *model_5* showed a superior performance with an increase in accuracy level by 8.7% to 73% compared with *model_4*. This stemmed from the fact that many more *home* activities were correctly imputed. In *model_6*, 2% of total cases more were successfully detected with the use of the non-home frequently visited place-based feature, which demonstrated that our idea on this feature was worthy.

The mentioned-above changes in model performances emphasized that features related to time and home location were the strongest influential ones. The contribution of user- and actual mode-based features were fairly similar and significant to the improvement of the overall accuracy, which is in line with their heavy use before (Cui et al., 2018; Feng and Timmermans, 2015; McGowen and McNally, 2007; Montini et al., 2014; Shen and Stopher, 2014b; Xiao et al., 2016). The predicted mode information possibly due to including wrong imputation did not enhance the performance as much as the truth. Next, we analyzed the changes in precision and recall of purposes corresponding to the addition of features.

Table 6 - 4. Training and test results of model_6

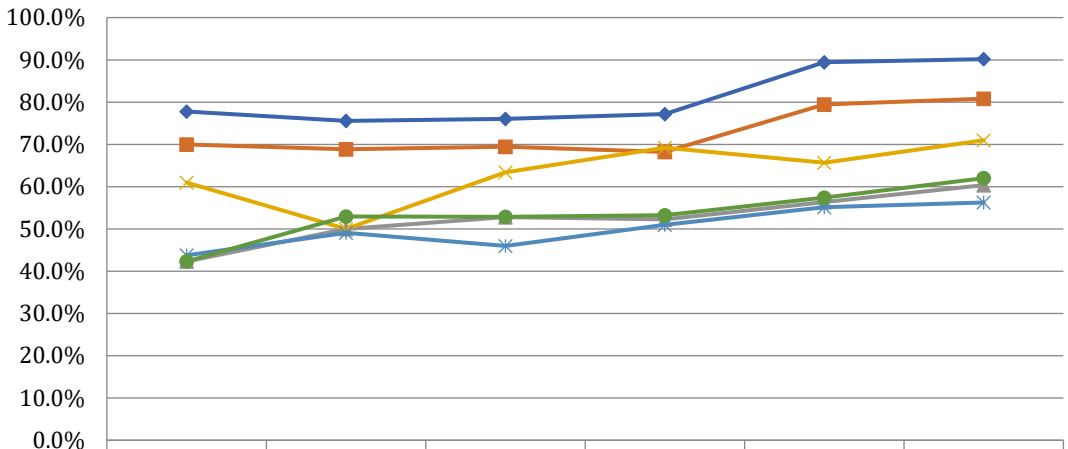
		Predicted						Recall	F^{class}
		Home	Work /education	Shopping /eating	Pick-up /drop-off	Visit /leisure	Business		
PREDICTION RESULTS OF TRAINING SET									
Reported	Home	673	7	11	2	3	5	*96.0%*	*95.3%*
	Work/education	6	261	12	1	13	8	*86.7%*	*85.4%*
	Shopping/eating	9	13	230	17	35	36	*67.6%*	*68.4%*
	Pick-up/drop-off	5	1	27	75	4	6	*63.6%*	*69.1%*
	Visit/leisure	15	22	32	1	126	31	*55.5%*	*59.3%*
	Business	3	6	21	3	17	210	*80.8%*	*75.5%*
Precision		*94.7%*	*84.2%*	*69.1%*	*75.8%*	*63.6%*	*70.9%*	Accuracy: **80.9%**	F^{model}: **75.5%**
PREDICTION RESULTS OF TEST SET									
Reported	Home	220	4	5	1	2	2	*94.0%*	*92.1%*
	Work/education	5	80	6	0	3	6	*80.0%*	*80.4%*
	Shopping/eating	8	2	67	6	16	15	*58.8%*	*59. 6%*
	Pick-up/drop-off	6	0	6	22	2	3	*56.4%*	*62.9%*
	Visit/leisure	2	10	15	1	36	12	*47.4%*	*51.4%*
	Business	3	3	12	1	5	62	*72.1%*	*66.7%*
Precision		*90.2%*	*80.8%*	*60.4%*	*71.0%*	*56.3%*	*62.0%*	Accuracy: **75.0%**	F^{model}: **68.8%**

Figure 6-8 and Figure 6-9 show that with start time and duration, detection of home and work were reasonable with both recall and precision levels of over 70%. The rest was identified much worse with the precision values being smaller than 50%, except for *pick-up/drop-off* whose precision and recall were being 60%. This may result from the fact that most cases of this purpose had the shortest duration (i.e. up to 30 mins) and two peak periods per day, as noted in Sub-section 6.3.3.

The recall of *business* in *model_2* significantly rose in comparison with that in *model_1*, leading to an increase in its precision. This would come from the addition of the variable "dynamic job". User's information improved the classification of *social/visit* whereas affected negatively that of *pick-up/drop-off* whose precision and recall decreased by 9% and 3.5%, respectively. In this sense, more actual *pick-up/drop-off* cases were falsely inferred and more cases of other purposes were misclassified as *pick-up/drop-off* simultaneously.

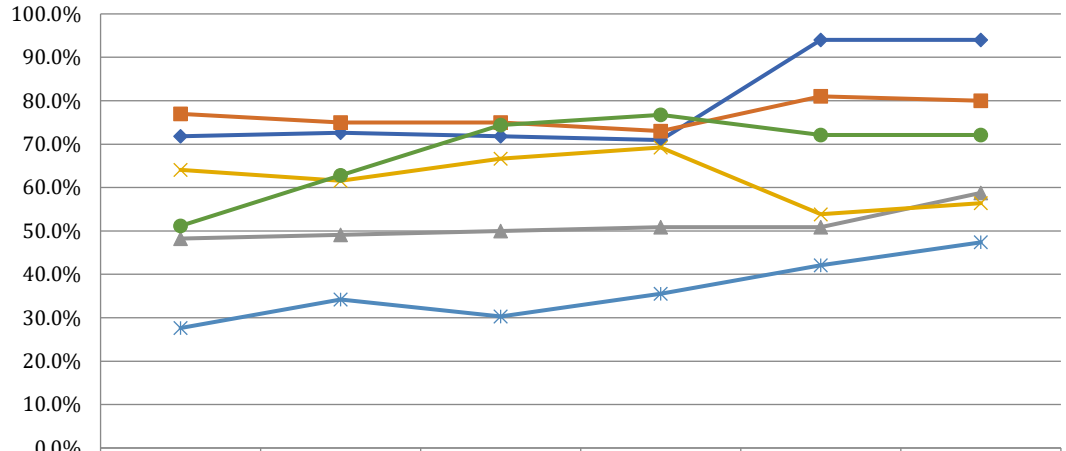
The use of predicted mode information in *model_3* caused the strongest effects on the detection of *business* and *visit/leisure* types albeit with opposite sides. In contrast to the 11% increase in recall of *business*, a decrease from 34.2% to 30.3% was reported for the *visit/leisure*. This would result from the fact that 20% of *visit/leisure* trips had mode wrongly recognized. The precision of *business* was nearly stable demonstrated many more cases were falsely classified as it.

In case actual mode was used (*model_4*) instead of predicted mode (*model_3*), recall of *business* improved from 62.8% (*model_2*) to 76.7%, which was higher than the increased level of *model_3* versus *model 2*. The recall of *visit/leisure* in *model 4* slightly climbed compared with that in *model_2*. Therefore, falsely identified mode affected mainly *visit/leisure* whilst mode information enhanced dramatically prediction of *business*.



	Model_1	Model_2	Model_3	Model_4	Model_5	Model_6
Home	77.8%	75.6%	76.0%	77.2%	89.4%	90.2%
Work/education	70.0%	68.8%	69.4%	68.2%	79.4%	80.8%
Shopping/eating	42.3%	50.0%	52.8%	52.3%	56.3%	60.4%
Pick-up/Drop-off	61.0%	50.0%	63.4%	69.2%	65.6%	71.0%
Visit/leisure	43.8%	49.1%	46.0%	50.9%	55.2%	56.3%
Business	42.3%	52.9%	52.9%	53.2%	57.4%	62.0%

Figure 6 - 8. Precision levels of purposes in six models



	Model_1	Model_2	Model_3	Model_4	Model_5	Model_6
Home	71.8%	72.6%	71.8%	70.9%	94.0%	94.0%
Work/education	77.0%	75.0%	75.0%	73.0%	81.0%	80.0%
Shopping/eating	48.2%	49.1%	50.0%	50.9%	50.9%	58.8%
Pick-up/Drop-off	64.1%	61.5%	66.7%	69.2%	53.8%	56.4%
Visit/leisure	27.6%	34.2%	30.3%	35.5%	42.1%	47.4%
Business	51.2%	62.8%	74.4%	76.7%	72.1%	72.1%

Figure 6 - 9. Recall levels of purposes in six models

Given that adding the home-based feature (*model_5*), the recall and precision of *home* went up impressively, emphasizing that the home-detection algorithm functioned well. Counterparts of *visit/leisure* and *work/education* considerably increased. As indicated in Sub-section 6.3.3, *visit/leisure* and *work/education* had ambiguous time profiles with *home*, and thus would be misclassified as home in *model_4*. In *model_5*, with the use of the home-specific variable many of them would be correctly discriminated from home because they were far from users' home locations. A fall from 69.2% to 53.9% in recall of *pick-up/drop-off* could result from the proximity of kindergartens and/or school of users' children to home.

In *model_6*, the use of the frequently visited place-based feature induced a rise in recalls of *shopping/eating*, *visit/leisure* and *pick-up/drop-off* whereas those of *home* and *work/education* stabilized. It can be explained that the majority of the most frequently visited non-home locations are markets instead of workplace, which in line with the fact there were more *shopping/eating* activities than *work/education* activities in the sample. Therefore, this feature directly and substantially fostered *shopping/eating* prediction. Along with this, a number of *pick-up/drop-off* and *visit/leisure* activities avoided being misclassified as *shopping/eating*. Notably, although the recall of *business* levelled off, its precision increased clearly, as many *shopping/eating* activities that had been falsely identified as *business* in *model_5* were correctly inferred in *model_6*.

As regards the comparison between *model_6* and the model of (Reumers et al., 2013), the accuracy of the former (75%) was lower than 76% of the latter; however, the accuracy was not the appropriate metric here. In *model_6*, F^{class} for home and work were higher than counterparts in (Reumers et al., 2013). Similarly, all non-home and non-work purposes were predicted better with the precision levels being between 56% and 71% (see Table 6-4) compared with those being around 40% in (Reumers et al., 2013). Consequently, F^{model} of *model_6* was obviously higher, that is, 68.8% versus 49.6% (see Table 6-3).

Interestingly, the precision and recall of work and home in *model_6* were comparable to those in studies collecting personal locations of participants (Montini et al., 2014; Yazdizadeh et al., 2019). Yet, imputation performances of other purposes were outperformed by counterparts generated by recent learning-based algorithms enriched with GIS data (Cui et al., 2018; Feng and Timmermans, 2015; Xiao et al., 2016).

6.6. SUMMARY

This chapter rigorously presents the process of choosing features and tuning HPs of RF models to enhance trip purpose imputation from GPS data without GIS data. This is such an extension of (Reumers et al., 2013).

Time and home-related features were the strongest influential ones. Transportation mode predicted increased the accuracy, which encouraged integrating mode detection and purpose imputation into a continuous process. The feature of frequently visited non-home place was handy for detecting purpose without GIS information. Conclusion of importance of feature associated to this place and predicted mode are novel and useful for the literature.

The proposed method does not require users to provide personal locations that may be sensitive and unachievable due to privacy concern. Additionally, it depends completely on internal features that were obtained from the sample solely. Therefore, it is highly transferable and able to be applied for comparing performances of detecting trip purposes in different areas simultaneously. Better classification, particularly for place-specific ones like shopping, eating out, visiting is obtainable with the fusion of GIS data. In this situation, our contributions to the use of predicted mode and the frequently visited non-home place are still informative and applicable. Generally, purpose imputation without GIS is suitable for considering mainly work/education together with home labels and accepting the reasonable predictions of other classes.

A remaining issue is the importance of mode-related features. The mode list used encompassed walk, bicycle, motorcycle, bus and car with a dominant role of motorcycle but without train and metro that have been often indicated previously (Bohte and Maat, 2009; Chen et al., 2010; Nguyen et al., 2019a; P. Stopher et al., 2008a; Yazdizadeh et al., 2019). Thus, we wish to repeat this study with data including more travel options and/or in different areas.

Chapter 7: CHALLENGES TO GPS-BASED SURVEYS AND RECOMMENDATIONS FOR IMPROVING QUALITY

7.1. INTRODUCTION

On the basis of empirical findings of the four previous chapters, Chapter 7 aims at summarizing challenges to develop inference models to derive a description of daily mobility from GPS data in general and for cities of developing countries with Hanoi (Vietnam) as a case study. The analyses of existing problems enable to find partial responses to the role of GPS-based survey in mobility investigation. Equally important is proposing recommendations for boosting the development of models to derive trip attributes from GPS data.

The following of this chapter is structured into three main parts. Section 7.2 presents challenges to develop imputation models whilst Section 7.3 discusses the role of GPS-based surveys in travel data collection. Section 7.4 proposes solutions to improve GPS-based survey's quality.

7.2. CHALLENGES TO DEVELOP INFERENCE MODEL

GPS-based survey is a multifaceted field; hence, with experiments only in Rhone-Alpes (France) and Hanoi (Vietnam), it is impossible to cover all its existing challenges and issues. Here, we focus on those preventing us from successfully constructing powerful mode and purpose identification algorithms.

7.2.1. General challenges

GPS data, because being the stream of numerous discrete points with coordinates and timestamps, require sophisticated algorithms to translate them into meaningful knowledge about trips' characteristics. The process of constructing and applying models would be as presented in Figure 7-1. Initially, GPS data are collected passively in either smartphone- or dedicated device-based surveys that are then followed by prompted recall surveys in order to gather ground truth information. By comparing prediction results with ground truth data, researchers can arrive at conclusions of either keeping calibrating models or applying models for data at a larger scale. In case the model is used to analyze new data, these data are usually not together with ground truth because the model is believed in being powerful enough to translate trajectories into travel patterns. To evaluate the potential of GPS-based survey and also the model, the results derived from GPS data by the model are compared with data collected in national/regional (household) travel surveys.

We followed this process to develop a fuzzy logic model enriched with GIS data in Chapter 4. Whilst running on a non-ground truth GPS data set (i.e. TOMOS data), it generated a reasonable prediction result of both trip detection and transportation mode detection in comparison with travel patterns extracted from data of the EDR-RA survey. Also based on fuzzy logic theory, a model was created to discriminate walk, bike and motorized modes from Hanoi data. The accuracy of 94% is good; however, it decreased in case of considering individual modes (i.e. bus, car and motorcycle) rather than motorized means. This

would remind us of the low transferability of methods between research areas, thus more tests should be carried out.

A problem of both mode and purpose imputation fields is imbalanced data that are constituted mainly by data of several major classes like home in purpose detection and car in mode detection. Imbalanced data probably lead inference models to focus on predicting major classes, even ignoring the precision/recall of minor classes, which can be seen clearly in the performance of RF mode identification model in Sub-section 5.6.4. Specifically, it successfully detected 79% of cases but only 15.4% (see Appendix 4) of the actual bike segments whose percentage in the sample was only 3.7% (see Section 5.4). Due to the imbalanced data problem that is common, the use of machine learning algorithms, which are widely applied and able to produce (very) high prediction results, would be inappropriate if sufficient data are not collected. In this sense, making the choice of model type depends largely upon data collected.

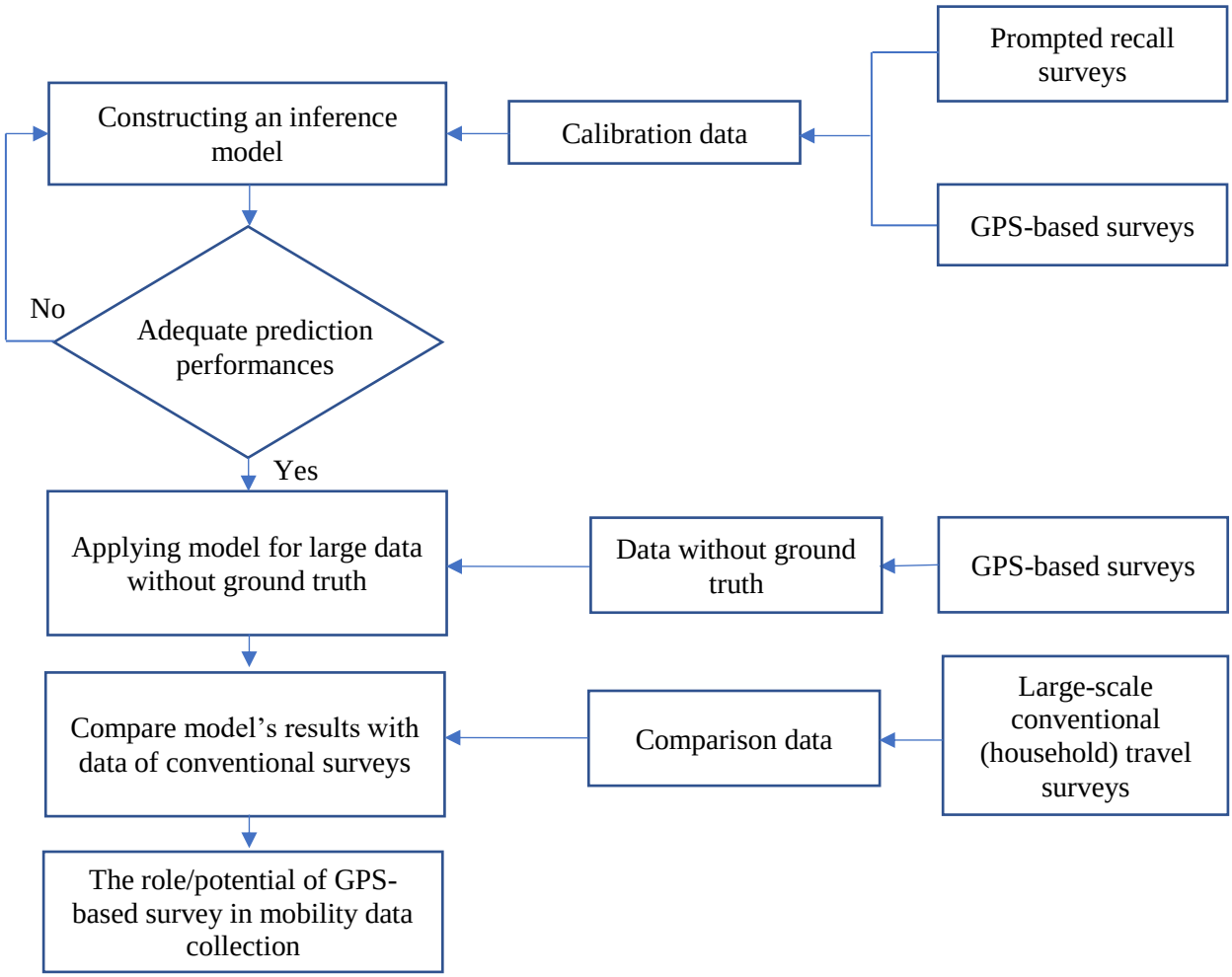


Figure 7 - 1. Process of developing and applying inference models

Developing imputation models relies heavily on the availability of external data sources. In Chapter 6, we implemented purpose identification without using GIS data because simply these data are not available. As a result, the proposed model was trained by variables extracted from the sample only. The lack of external source limits the model's prediction competence regarding non-home and non-work activities.

Another impediment to build up powerful classifiers is the limited quality of ground truth data. As presented in Section 3.2, by common sense and the logic between vehicle ownership and mode use, we pinpointed wrong validations to ask users to correct them or remove them in the Hanoi survey. The wrong

validations were associated with both modes and purposes. Bike and bus segments were falsely confirmed whilst the business purpose was misclassified as work type. Because ground truth is not only the input to train a model but also the standard to assess the prediction results, if ground truth data are not precise, the model constructed by them can show very high accuracy on evaluation but probably predicts poorly in practice.

7.2.2. Typical challenges in developing countries

The experience in conducting and analyzing data of the Hanoi survey helps us to realize barriers to use smartphones to observe travel in developing countries' cities.

The main and typical obstacle is the lack of the app dedicated for or carefully prepared for the research area. In the Hanoi survey, we used TravelVU that has been developed by calibration data of volunteers living in European countries. The transport contexts in these countries are obviously different from the counterparts in developing countries, Hanoi for example. The main travel mode in European cities is car, so it is probable that the largest amount of data to train in-built algorithms is of car whilst data of motorcycle would be minor. Bike here is encouraged by prioritized facilities like dedicated cycling lanes, leading traveling by bike to reach a fairly high speed. Whereas, the motorcycle is the most common in Hanoi and have average speed levels in urban areas ranging between 25 and 40km/h, similar to those of bike in European cities. As a result, numerous recommendations of the app for the potential mode of segments were bike but they were actually motorcycle. The frequent occurrence of wrong recommendations caused both increasing workloads of correcting data and users' unpleasant feelings, which would more or less affect negatively the quality of confirmations and the participation willingness. Besides, users faced particular difficulties whilst interacting and communicating with the app by foreign languages.

Privacy concern may prevent conducting GPS-based surveys completely or partially. For those strictly and seriously worrying about the risk of privacy invasion, they directly refused to install the app. In Hanoi, some partially concerned about their privacy, which was reflected by refusing to provide home addresses and possibly ignoring validating data.

Omitting data is an issue. This a common problem for any mobility observation-dedicated apps (Montini et al., 2015). For GPS wearable devices, missing data can come from warm and cold start issues. However, for smartphone, it is complex to uncover exact reasons and thus fix them timely. In a smartphone, many apps rather than only the mobility survey app use resources (e.g. memory and battery) at the same time to run. Additionally, reasons may vary between smartphone brands and versions surveyed. Users check a trip based on not only its characteristics (e.g. duration, speed) but also its previous trips. Consequently, missing data makes a user lose clues to recall or even give wrong confirmations owing to depending completely on the order of trips in his/her memory.

In Hanoi, GIS data of the city generally and of public transport network particularly are not accessible or not available, requiring more manual participation to develop algorithms independent on GIS data. In Chapter 6, if polygon-shaped data were available, we would have constructed an algorithm that used them to check and merge activities taking place at the same place and labeled identically.

7.3. VIEWS ON THE ROLE OF GPS-BASED SURVEY IN MOBILITY DATA COLLECTION

In the initial era of modern technology with the prominent role of GPS, the scientific society showed very positive views on the potential of GPS-based survey. It was believed to be not only a useful complement to but also a complete alternative to conventional methods like CATI or face-to-face interviews.

Many technological efforts have been made to progress GPS-based surveys. Typical evidence is the revolution of devices to collect data, that is, from on-board devices to handheld devices and smartphones now. However, GPS-based survey is far from the full replacement to conventional methods. Some reasons for it have been examined and reported, such as the unstable spatial accuracy of GPS data (Vij and Shankari, 2015), missing a part of or the whole trips (T. T. Nguyen et al., 2017) and difficulties in collecting data of special groups with disabilities or old ages (Neven et al., 2018; Safi et al., 2017).

In this thesis, the hierarchical process detected correctly 89.1% of travel segments according to mode whilst the RF model inferred successfully 75% of activity segments. These accuracy levels demonstrate that there are gaps between self-reported confirmations of users and predictions of models. Hypothesizing that ground truth information provided is perfect, more efforts need to be invested to have better classifiers. Unfortunately, ground truth encompasses false validations at different levels; limiting the prediction ability of models built on it in real.

Researchers always profit from good input conditions to gain high detection performances; however, the practice is more complex. In fact, there is no available mode information to help well to identify activities. Mode is hide information in GPS. So, it should be recognized and input to purpose imputation. Clearly, purpose prediction by mode identified was not good as that by actual mode. Another example is related to the unavailability of GIS data. In real, GIS data cannot be reached in any places. Without this datum, purpose imputation is only satisfactory for work and home.

The lists of mode and purpose derived from GPS data are aggregated to gain high accuracy. Undeniably at disaggregated prediction levels, GPS data would be sufficient for knowing travel behaviors like identifying purpose distribution of home and work activities from bike data in (Usyukov, 2017). However, it would be impossible to have information about various mode/activity types as in conventional surveys.

All in all, we view that efforts are still concentrating on developing algorithms to derive travel characteristics from GPS data. It is difficult for GPS-based survey to act alone to provide very good knowledge about both mode and purpose simultaneously, even compared with conventional techniques. GPS data would be the most suitable support for conventional investigations in order to capture travel behaviors in a much more detailed and accurate way. In case the disaggregated classification is accepted and the sample is large enough, the estimation of trip attributes from GPS data is a good reference; however, there is a need for carefully developing inference models.

7.4. RECOMMENDATIONS FOR ENHANCING QUALITY OF MOBILITY SURVEY USING GPS

To address the challenges mentioned above and thus improve the accuracy along with the potential of practical use of inference models, a number of points should be taken into considerations, as follows:

- The focus now is to develop inference models. Data to calibrate, train and test models should be carefully collected.

- The integration of mode and purpose into a continuous process is feasible and possibly generate reasonable results. Separating purpose detection from mode imputation enables to achieve a satisfactory prediction performance but it may be opposite to procedures of processing GPS data.

- GPS-based surveys by smartphone are a useful tool for collecting data not only in cities of developed countries but also in urban areas in developing countries. Most attention should be paid to the app. In order to deploy it to collect data of large sample, surveyors may need to collect data to train its in-built algorithms. The scale of calibration data survey does not need to be so large but needs to be precise. If doing it successfully, they can limit wrong recommendations and indeed reduce users' burden in terms of validating data.

- Language barrier should be cared about whilst using an app for countries whose citizens are not fluent in the language(s) offered by the app. English would be an alternative; however, it is easier and perhaps more effective for users to provide their validations if their mother tongue is available.

- Technical issues related to high battery consumption and omitting to record traces should be alleviated.

- To extend the coverage of GPS surveys to cities in Global South, it is necessary to deal with the vague indicators (e.g. speed, acceleration) between motorcycles and other motorized modes, especially car. GIS data may not be as good as those in cities of the Global North, thus new and innovative solutions are expected. The less dependent upon external sources that are specific for particular areas a model is, the more highly applicable and transferable it is.

- Besides the utilization of all-in-one process with powerful supervised learning algorithms, the development of hierarchical processes based on taking advantages of different method types (i.e. rule-based, probability-based and learning-based also) allows reaching the balance between accuracy and interpretability.

Chapter 8: CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

GPS is one of the prominent technologies of today and tomorrow. Its applications in travel data collection have made a revolution and encouraged thoughts about its potentials to supplement and towards completely replacing conventional methods. However, GPS-based surveys have their own limitations that have prevented their data from being analyzed and used independently to be aware of travel characteristics of participants.

Based on the literature review in Chapter 2, the biggest drawbacks of GPS data are lack of trip attributes, that is, mode and purpose. In the following chapters, the thesis concentrated on developing methods to infer travel modes and purposes from GPS data in different contexts.

Chapter 3 describes two datasets used for this thesis. The first was collected by wearable loggers in Rhone-Alpes, France whilst the second was gathered by smartphone in Hanoi, Vietnam.

In Chapter 4, a fuzzy logic-based model together with the GIS datum-based rules were developed to analyze a GPS datum set without ground truth collected in Rhone-Alpes, France. The results emphasized the reasonable prediction of models proposed and the better ability to trace behaviors of GPS than CATI.

In Chapter 5, a different datum set collected in Hanoi with the corresponding ground truth and the dominant percentage of motorcycle was used to develop a hierarchical process of travel mode inference. By deploying the fuzzy logic model proposed in Chapter 4 as a first step, the hierarchy classified satisfactorily with the 89.1% accuracy. However, the confusion between motorcycle and car should be enhanced.

In Chapter 6, the mode prediction results in Chapter 5 were used as input variables to derive trip purposes in Hanoi data. Because GIS data were unavailable, locations of home and the most frequently visited non-home places were estimated to enhance the predictions. The results highlighted that using mode predicted improved the purpose prediction. Without GIS data, purposes still can be detected at reasonable levels. More importantly, the findings of Chapter 5 and Chapter 6 contributed to gain a view on and a prospect of getting the most important trip characteristics (i.e. mode and purpose) by a continuous process.

In Chapter 7, challenges to develop mode and purpose inference models were synthesized based on findings of previous chapters. Difference in the quality of ground truth and the availability of external data sources in various contexts has been the main barriers keeping GPS-based survey far from completely replacing conventional techniques. The emphasis should be on how to combine GPS and self-reported data to gain comprehensive and adequate pictures of travel behaviors. Besides, using smartphones with well-prepared apps can be a good solution to collect data in cities of the Global South.

Regarding future research directions, inference models were developed for data collected in Hanoi where there was no metro service at the survey time. Thus, it is worth applying and improving models by data comprised of more travel options. Besides, more sophisticated and complex remedies need to be created to address better the confusion between motorcycle and car. Conclusions about the potential integration of mode and purpose detection into a continuous process should be tested in different backgrounds, with the availability of GIS data for example.

BIBLIOGRAPHY

Abdulazim, T., Abdelgawad, H., Habib, K.M.N., Abdulhai, B., 2013. Using Smartphones and Sensor Technologies to Automate Collection of Travel Data. Transp. Res. Rec. J. Transp. Res. Board 2383, 44–52. https://doi.org/10.3141/2383-06

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. J. Urban Technol. 17, 3–27. https://doi.org/10.1080/10630731003597306

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol., Big Data in Transportation and Traffic Engineering 58, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., Hickman, M., 2018. Public transport trip purpose inference using smart card fare data. Transp. Res. Part C Emerg. Technol. 87, 123–137. https://doi.org/10.1016/j.trc.2017.12.016

Armoogum, J., Bonsall, P., Browne, M., Christensen, L., Cools, M., Cornelis, E., Diana, M., Guilloux, T., Harder, H., Hegner Reinau, K., Hubert, J.-P., Kagerbauer, M., Kuhnimhof, T., Madre, J.-L., Moiseeva, A., Polak, J., Schulz, A., Tébar, M., Vidalakis, L., 2014. Survey Harmonisation with New Technologies Improvement (SHANTI). IFSTTAR.

Armoogum, J., Dill, J., 2015. Workshop Synthesis: Sampling Issues, Data Quality & Data Protection. Transp. Res. Procedia, Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia 11, 60–65. https://doi.org/10.1016/j.trpro.2015.12.006

Armoogum, J., Ellison, A., Kalter, M.-J.O., 2018a. Workshop Synthesis: Representativeness in surveys: challenges and solutions. Transp. Res. Procedia 32, 224–228. https://doi.org/10.1016/j.trpro.2018.10.041

Armoogum, J., Tébar, M., Christian, B., Garcia, C., Nguyen, M.H., Rendina, F., 2018b. Rapport de synthèse : Méthodologie afin de mesurer la mobilité régionale : élaboration d'une enquête régionale. IFSTTAR, Champs-sur-Marne.

Auld, J., Williams, C., Mohammadian, A. (Kouros), Nelson, P., 2009. An automated GPS-based prompted recall survey with learning algorithms. Transp. Lett. Int. J. Transp. Res. 1, 59–79.

Bassett, D.R., Pucher, J., Buehler, R., Thompson, D.L., Crouter, S.E., 2008. Walking, cycling, and obesity rates in Europe, North America, and Australia. J. Phys. Act. Health 5, 795–814.

Bayart, C., Bonnel, P., 2015. How to Combine Survey Media (Web, Telephone, Face-to-Face): Lyon and Rhône-alps Case Study. Transp. Res. Procedia, Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia 11, 118–135. https://doi.org/10.1016/j.trpro.2015.12.011

Bayart, C., Bonnel, P., 2012. Combining web and face-to-face in travel surveys: comparability challenges? Transportation 39, 1147–1171. https://doi.org/10.1007/s11116-012-9393-x

Behrens, R., Freedman, M., McGuckin, N., 2009. The Challenge of Surveying "Hard to Reach" Groups: Synthesis of a Workshop, in: Transport Survey Method: Keeping Up With a Changing World. Emerald.

Biljecki, F., Ledoux, H., Oosterom, P. van, 2013. Transportation mode-based segmentation and classification of movement trajectories. Int. J. Geogr. Inf. Sci. 27, 385–407. https://doi.org/10.1080/13658816.2012.692791

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Böcker, L., Dijst, M., Prillwitz, J., 2013. Impact of Everyday Weather on Individual Daily Travel Behaviours in Perspective: A Literature Review. Transp. Rev. 33, 71–91. https://doi.org/10.1080/01441647.2012.747114

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. Transp. Res. Part C Emerg. Technol. 17, 285–297. https://doi.org/10.1016/j.trc.2008.11.004

Boltze, M., Tuan, V.A., 2016. Approaches to Achieve Sustainability in Traffic Management. Procedia Eng. 142, 205–212. https://doi.org/10.1016/j.proeng.2016.02.033

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Bricka, S., Bhat, C.R., 2006. A Comparative Analysis of GPS-Based and Travel Survey-based Data. http://www.metapress.com/content/x6643h3856243663/.

Buehler, R., Pucher, J., Merom, D., Bauman, A., 2011. Active Travel in Germany and the U.S.: Contributions of Daily Walking and Cycling to Physical Activity. Am. J. Prev. Med. 41, 241–250. https://doi.org/10.1016/j.amepre.2011.04.012

Burbidge, S., Goulias, K., 2009. Active travel behavior. Transp. Lett. 1, 147–167. https://doi.org/10.3328/TL.2009.01.02.147-167

Burkhard, O., Becker, H., Weibel, R., Axhausen, K.W., 2020. On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data. Transp. Res. Part C Emerg. Technol. 114, 99–117. https://doi.org/10.1016/j.trc.2020.01.021

Byon, Y.-J., Abdulhai, B., Shalaby, A., 2009. Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices. J. Intell. Transp. Syst. 13, 161–170. https://doi.org/10.1080/15472450903287781

Castree, N., Kitchin, R., Rogers, A., 2013. A Dictionary of Human Geography, 1st ed. Oxford University Press. https://doi.org/10.1093/acref/9780199599868.001.0001

Chaix, B., Méline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., Lewin, A., Labadi, K., Kestens, Y., 2013. GPS tracking in neighborhood and health studies: A step forward for environmental exposure assessment, a step backward for causal inference? Health Place 21, 46–51. https://doi.org/10.1016/j.healthplace.2013.01.003

Chen, C., Batty, M., van Vuren, T., 2015. Editorial. Transportation 42, 537–540. https://doi.org/10.1007/s11116-015-9614-1

Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transp. Res. Part C Emerg. Technol. 46, 326–337. https://doi.org/10.1016/j.trc.2014.07.001

Chen, C., Gong, H., Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transp. Res. Part Policy Pract. 44, 830–840. https://doi.org/10.1016/j.tra.2010.08.004

Chen, C., Liao, C., Xie, X., Wang, Y., Zhao, J., 2019. Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. Pers. Ubiquitous Comput. 23, 53–66. https://doi.org/10.1007/s00779-018-1175-9

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Chung, E.-H., Shalaby, A., 2005. A Trip Reconstruction Tool for GPS-based Personal Travel Surveys. Transp. Plan. Technol. 28, 381–401. https://doi.org/10.1080/03081060500322599

Cools, M., Moons, E., Creemers, L., Wets, G., 2010. Changes in Travel Behavior in Response to Weather Conditions: Do Type of Weather and Trip Purpose Matter? Transp. Res. Rec. J. Transp. Res. Board 2157, 22–28. https://doi.org/10.3141/2157-03

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., Zegras, P., 2013. Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore. Transp. Res. Rec. J. Transp. Res. Board 2354, 59–67. https://doi.org/10.3141/2354-07

Cui, Y., Meng, C., He, Q., Gao, J., 2018. Forecasting current and next trip purpose with social media data and Google Places. Transp. Res. Part C Emerg. Technol. 97, 159–174. https://doi.org/10.1016/j.trc.2018.10.017

Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp. Res. Part C Emerg. Technol. 86, 360–371. https://doi.org/10.1016/j.trc.2017.11.021

Das, R.D., Winter, S., 2018. A fuzzy logic based transport mode detection framework in urban environment. J. Intell. Transp. Syst. 22, 478–489. https://doi.org/10.1080/15472450.2018.1436968

Deng, Z., Ji, M., 2010. Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach, in: Traffic and Transportation Studies 2010. Presented at the Seventh International Conference on Traffic and Transportation Studies (ICTTS) 2010, American Society of Civil Engineers, Kunming, China, pp. 768–777. https://doi.org/10.1061/41123(383)73

Devillaine, F., Munizaga, M., Trépanier, M., 2012. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. Transp. Res. Rec. J. Transp. Res. Board 2276, 48–55. https://doi.org/10.3141/2276-06

Draijer, G., Kalfs, N., Perdok, J., 2000. Global Positioning System as Data Collection Method for Travel Research. Transp. Res. Rec. J. Transp. Res. Board 1719, 147–153. https://doi.org/10.3141/1719-19

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., Das, K., 2017. Real-time trip purpose prediction using online location-based search and discovery services. Transp. Res. Part C Emerg. Technol. 77, 96–112. https://doi.org/10.1016/j.trc.2017.01.020

Fang, C., Liu, H., Luo, K., Yu, X., 2017. Process and proposal for comprehensive regionalization of Chinese human geography. J. Geogr. Sci. 27, 1155–1168. https://doi.org/10.1007/s11442-017-1428-y

Feng, T., Timmermans, H.J.P., 2019. Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty. Transp. Plan. Technol. 42, 274–292. https://doi.org/10.1080/03081060.2019.1576384

Feng, T., Timmermans, H.J.P., 2016. Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. Transp. Plan. Technol. 39, 180–194. https://doi.org/10.1080/03081060.2015.1127540

Feng, T., Timmermans, H.J.P., 2015. Detecting activity type from gps traces using spatial and temporal information. Eur. J. Transp. Infrastruct. Res. 15, 662–674.

Feng, T., Timmermans, H.J.P., 2013. Transportation mode recognition using GPS and accelerometer data. Transp. Res. Part C Emerg. Technol. 37, 118–130. https://doi.org/10.1016/j.trc.2013.09.014

Forman, G., 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. J Mach Learn Res 3, 1289–1305.
Forrest, T., Pearson, D., 2005. Comparison of Trip Determination Methods in Household Travel Surveys Enhanced by a Global Positioning System. Transp. Res. Rec. J. Transp. Res. Board 1917, 63–71. https://doi.org/10.3141/1917-08
Furletti, B., Cintia, P., Renso, C., Spinsanti, L., 2013. Inferring Human Activities from GPS Tracks, in: Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13. ACM, New York, NY, USA, pp. 5:1–5:8. https://doi.org/10.1145/2505821.2505830
Gautama, S., Atzmueller, M., Kostakos, V., Gillis, D., Hosio, S., 2017. Observing Human Activity Through Sensing, in: Loreto, V., Haklay, M., Hotho, A., Servedio, V.D.P., Stumme, G., Theunis, J., Tria, F. (Eds.), Participatory Sensing, Opinions and Collective Awareness. Springer International Publishing, Cham, pp. 47–68. https://doi.org/10.1007/978-3-319-25658-0_3
General Statistics Office, 2017. Statistical Yearbook of Vietnam. Statistical Publishing House, Vietnam.
Gibson, C., 2009. Human Geography, in: International Encyclopedia of Human Geography. Elsevier, pp. 218–231. https://doi.org/10.1016/B978-008044910-4.00275-3
Goldstein, B.A., Polley, E.C., Briggs, F.B.S., 2011. Random forests for genetic association studies. Stat. Appl. Genet. Mol. Biol. 10, 32. https://doi.org/10.2202/1544-6115.1691
Golob, T.T., Meurs, H., 1986. Biases in response over time in a seven-day travel diary. Transportation 13, 163–181. https://doi.org/10.1007/BF00165546
Gong, H., Chen, C., Bialostozky, E., Lawson, C.T., 2012. A GPS/GIS method for travel mode detection in New York City. Comput. Environ. Urban Syst., Special Issue: Geoinformatics 2010 36, 131–139. https://doi.org/10.1016/j.compenvurbsys.2011.05.003
Gong, L., Kanamori, R., Yamamoto, T., 2018. Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. Travel Behav. Soc. https://doi.org/10.1016/j.tbs.2017.03.004
Gong, L., Liu, X., Wu, L., Liu, Y., 2016. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartogr. Geogr. Inf. Sci. 43, 103–114. https://doi.org/10.1080/15230406.2015.1014424
Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. Procedia - Soc. Behav. Sci., The 9th International Conference on Traffic and Transportation Studies (ICTTS 2014) 138, 557–565. https://doi.org/10.1016/j.sbspro.2014.07.239
Hall, R. (Ed.), 2003. Handbook of Transportation Science, 2nd ed. 2003 edition. ed. Springer, Boston.
Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal Patterns of Urban Human Mobility. J. Stat. Phys. 151, 304–318. https://doi.org/10.1007/s10955-012-0645-0
Hashemi, M., Karimi, H.A., 2016. A weight-based map-matching algorithm for vehicle navigation in complex urban networks. J. Intell. Transp. Syst. 20, 573–590. https://doi.org/10.1080/15472450.2016.1166058
Hashemi, M., Karimi, H.A., 2014. A critical review of real-time map-matching algorithms: Current issues and future directions. Comput. Environ. Urban Syst. 48, 153–165. https://doi.org/10.1016/j.compenvurbsys.2014.07.009
Ho, C.Q., Mulley, C., 2013. Multiple purposes at single destination: A key to a better understanding of the relationship between tour complexity and mode choice. Transp. Res. Part Policy Pract. 49, 206–219. https://doi.org/10.1016/j.tra.2013.01.040
Hsu, T.-P., Sadullah, A.F.M., Nguyen, X.D., 2003. A comparison study on motorcycle traffic development in some Asian countries – case of Taiwan, Malaysia and Vietnam (Prepared for: The Eastern Asia Society for Transportation Studies (EASTS)).
Hu, X., An, S., Wang, J., 2018. Taxi Driver's Operation Behavior and Passengers' Demand Analysis Based on GPS Data. J. Adv. Transp. 2018, 1–11. https://doi.org/10.1155/2018/6197549
Huynh, D., 2020. Making Megacities in Asia: Comparing National Economic Development Trajectories, SpringerBriefs in Regional Science. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-15-0660-4
Huynh, D.T., Gomez-Ibañez, J., 2017. Vietnam, in: Pojani, D., Stead, D. (Eds.), The Urban Transport Crisis in Emerging Economies. Springer International Publishing, Cham, pp. 267–282. https://doi.org/10.1007/978-3-319-43851-1_13
Jiang, S., Ferreira, J., Gonzalez, M.C., 2017. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. IEEE Trans. Big Data 3, 208–219. https://doi.org/10.1109/TBDATA.2016.2631141
Johnson, R., Kuby, P., 2008. Elementary statistics, 10th ed. ed. Thomson Brooks/Cole, Belmont, CA.
Kelly, P., Krenn, P., Titze, S., Stopher, P., Foster, C., 2013. Quantifying the Difference Between Self-Reported and Global Positioning Systems-Measured Journey Durations: A Systematic Review. Transp. Rev. 33, 443–459. https://doi.org/10.1080/01441647.2013.815288
Krause, C.M., Zhang, L., 2019. Short-term travel behavior prediction with GPS, land use, and point of interest data. Transp. Res. Part B Methodol. 123, 349–361. https://doi.org/10.1016/j.trb.2018.06.012

Lee, R.J., Sener, I.N., III, J.A.M., 2016. An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. Transp. Lett. 8, 181–193. https://doi.org/10.1080/19427867.2015.1106787
Lee, S.G., Hickman, M., 2014. Trip purpose inference using automated fare collection data. Public Transp. 6, 1–20. https://doi.org/10.1007/s12469-013-0077-5
Liao, L., Fox, D., Kautz, H., 2007. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. Int. J. Robot. Res. 26, 119–134. https://doi.org/10.1177/0278364907073775
Liu, C., Susilo, Y.O., Karlström, A., 2017. Weather variability and travel behaviour – what we know and what we do not know. Transp. Rev. 37, 715–741. https://doi.org/10.1080/01441647.2017.1293188
Lu, Y., Zhu, S., Zhang, L., 2012. A Machine Learning Approach to Trip Purpose Imputation in GPS-Based Travel Surveys. Presented at the 4th Conference on Innovations in Travel Modeling, Florida, United States.
Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud. 7, 1–13. https://doi.org/10.1016/S0020-7373(75)80002-2
Marchal, P., Madre, J.-L., Yuan, S., 2011. Postprocessing Procedures for Person-Based Global Positioning System Data Collected in the French National Travel Survey 2007–2008. Transp. Res. Rec. J. Transp. Res. Board 2246, 47–54. https://doi.org/10.3141/2246-07
Marra, A.D., Becker, H., Axhausen, K.W., Corman, F., 2019. Developing a passive GPS tracking system to study long-term travel behavior. Transp. Res. Part C Emerg. Technol. 104, 348–368. https://doi.org/10.1016/j.trc.2019.05.006
McGowen, P.T., McNally, M.G., 2007. Evaluating the Potential To Predict Activity Types from GPS and GIS Data. Presented at the Transportation Research Board 86th Annual MeetingTransportation Research Board.
Meng, C., Cui, Y., He, Q., Su, L., Gao, J., 2017. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data, in: 2017 IEEE International Conference on Big Data (Big Data). Presented at the 2017 IEEE International Conference on Big Data (Big Data), pp. 1319–1324. https://doi.org/10.1109/BigData.2017.8258062
Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., Axhausen, K.W., 2015. Comparison of Travel Diaries Generated from Smartphone Data and Dedicated GPS Devices. Transp. Res. Procedia, Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia 11, 227–241. https://doi.org/10.1016/j.trpro.2015.12.020
Montini, L., Rieser-Schüssler, N., Horni, A., Axhausen, K., 2014. Trip Purpose Identification from GPS Tracks. Transp. Res. Rec. J. Transp. Res. Board 2405, 16–23. https://doi.org/10.3141/2405-03
Motlagh, O., Tang, S.H., Ismail, N., Ramli, A.R., 2009. A Review on Positioning Techniques and Technologies: A Novel AI Approach. J. Appl. Sci. 9, 1601–1614. https://doi.org/10.3923/jas.2009.1601.1614
Murakami, E., Wagner, D.P., 1999. Can using global positioning system (GPS) improve trip reporting? Transp. Res. Part C Emerg. Technol. 7, 149–165. https://doi.org/10.1016/S0968-090X(99)00017-0
Murakami, E., Watterson, W.T., 1990. Developing a household travel survey for the Puget Sound Region. Transp. Res. Rec. J. Transp. Res. Board 1285, 40–48.
Neven, A., Schutter, I.D., Wets, G., Feys, P., Janssens, D., 2018. Data Quality of Travel Behavior Studies: Factors Influencing the Reporting Rate of Self-Reported and GPS-Recorded Trips in Persons with Disabilities. Transp. Res. Rec. 2672, 662–674. https://doi.org/10.1177/0361198118772952
Nguyen, M.H., Armoogum, J., Garcia, C., 2019a. Mode-Based Comparison of Data in Mobility Surveys using GPS and Telephone. Presented at the 98th TRB Annual Meeting, Washington, D.C.
Nguyen, M.H., Ha, T.T., Le, T.L., Nguyen, T.C., 2017. Challenges to Development of Bus System Evidence from a Comparative Analysis of Surveys in Hanoi, in: Transportation for a Better Life: Mobility and Road Safety Managements. Presented at the Atrans Annual Conference, Bangkok, Thailand, pp. 1–10.
Nguyen, M.H., Ha, T.T., Tu, S.S., Nguyen, T.C., 2019b. Impediments to the bus rapid transit implementation in developing countries – a typical evidence from Hanoi. Int. J. Urban Sci. 0, 1–20. https://doi.org/10.1080/12265934.2019.1577747
Nguyen, M.H., Pojani, D., 2018. Chapter Two - Why Do Some BRT Systems in the Global South Fail to Perform or Expand?, in: Shiftan, Y., Kamargianni, M. (Eds.), Preparing for the New Era of Transport Policies: Learning from Experience, Advances in Transport Policy and Planning. Academic Press, pp. 35–61. https://doi.org/10.1016/bs.atpp.2018.07.005
Nguyen, T.C., 2016. Quality improvement and bus system development in Hanoi vision to 2025.
Nguyen, T.T., 2013. Mise au point d'une méthode de collecte de données de mobilité en utilisant des récepteurs GPS qui soit comparable avec les enquêtes classiques et applicable dans les pays du Sud. UNIVERSITE PARIS 1, France.
Nguyen, T.T., Armoogum, J., Madre, J.-L., Pham, T.H.T., 2017. GPS and travel diary: Two recordings of the same mobility, in: ISCTSC,11th International Conference on Transport Survey Methods. Esterel, Canada, p. 13p.
Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P., 2014. Supporting large-scale travel surveys with smartphones – A practical approach. Transp. Res. Part C Emerg. Technol., Special Issue with Selected Papers from Transport Research Arena 43, Part 2, 212–221. https://doi.org/10.1016/j.trc.2013.11.005

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7, e37027. https://doi.org/10.1371/journal.pone.0037027
Nour, A., Hellinga, B., Casello, J., 2016. Classification of automobile and transit trips from Smartphone data: Enhancing accuracy using spatial statistics and GIS. J. Transp. Geogr. 51, 36–44. https://doi.org/10.1016/j.jtrangeo.2015.11.005
OECD, 2018. OECD Urban Policy Reviews: Viet Nam, OECD Urban Policy Reviews. OECD Publishing. https://doi.org/10.1787/9789264286191-en
Oliveira, M.G.S., Vovsha, P., Wolf, J., Mitchell, M., 2014. Evaluation of Two Methods for Identifying Trip Purpose in GPS-Based Household Travel Surveys. Transp. Res. Rec. J. Transp. Res. Board 2405, 33–41. https://doi.org/10.3141/2405-05
Ortúzar, J.D.D., Armoogum, J., Madre, J., Potier, F., 2011. Continuous Mobility Surveys: The State of Practice. Transp. Rev. 31, 293–312. https://doi.org/10.1080/01441647.2010.510224
Patterson, Z., Fitzsimmons, K., 2016. DataMobile: Smartphone Travel Survey Experiment. Transp. Res. Rec. J. Transp. Res. Board 2594, 35–43. https://doi.org/10.3141/2594-07
Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568. https://doi.org/10.1016/j.trc.2010.12.003
People's Committee of Hanoi, 2016. Master public transport plan in Hanoi to 2020 and vision to 2025.
Pham, T.H.T., 2016. Apports et difficultés d'une collecte de données à l'aide de récepteurs GPS pour réaliser une enquête sur la mobilité. Université Paris-Est, France.
Pojani, D., 2020. Planning for Sustainable Urban Transport in Southeast Asia Policy Transfer, Diffusion, and Mobility. Springer, Cham.
Prelipcean, A.C., Gidófalvi, G., Susilo, Y.O., 2017. Transportation mode detection – an in-depth review of applicability and reliability. Transp. Rev. 37, 442–464. https://doi.org/10.1080/01441647.2016.1246489
Probst, P., Wright, M.N., Boulesteix, A., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9. https://doi.org/10.1002/widm.1301
Quddus, M., Washington, S., 2015. Shortest path and vehicle trajectory aided map-matching for low frequency GPS data. Transp. Res. Part C Emerg. Technol. 55, 328–339. https://doi.org/10.1016/j.trc.2015.02.017
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. Transp. Res. Part C Emerg. Technol. 15, 312–328. https://doi.org/10.1016/j.trc.2007.05.002
Rasmussen, T.K., Ingvardson, J.B., Halldórsdóttir, K., Nielsen, O.A., 2015. Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: A case study from the Greater Copenhagen area. Comput. Environ. Urban Syst. 54, 301–313. https://doi.org/10.1016/j.compenvurbsys.2015.04.001
Reumers, S., Liu, F., Janssens, D., Cools, M., Wets, G., 2013. Semantic Annotation of Global Positioning System Traces: Activity Type Inference. Transp. Res. Rec. J. Transp. Res. Board 2383, 35–43. https://doi.org/10.3141/2383-05
Richardson, A.J., Ampt, E.S., Meyburg, A.H., 1996. Nonresponse Issues in Household Travel Surveys, in: Transportation Research Board Conference Proceedings. Presented at the Conference on Household Travel Surveys: New Concepts and Research NeedsTransportation Research Board; Federal Highway Administration; Federal Transit Administration; and Bureau of Transportation Statistics.
Safi, H., Assemi, B., Mesbah, M., Ferreira, L., 2017. An empirical comparison of four technology-mediated travel survey methods. J. Traffic Transp. Eng. Engl. Ed., Special Issue: Driver Behavior, Highway Capacity and Transportation Resilience 4, 80–87. https://doi.org/10.1016/j.jtte.2015.12.003
Safi, H., Assemi, B., Mesbah, M., Ferreira, L., 2016. Trip Detection with Smartphone-Assisted Collection of Travel Data. Transp. Res. Rec. J. Transp. Res. Board 2594, 18–26. https://doi.org/10.3141/2594-03
Safi, H., Assemi, B., Mesbah, M., Ferreira, L., Hickman, M., 2015. Design and Implementation of a Smartphone-Based Travel Survey. Transp. Res. Rec. 2526, 99–107. https://doi.org/10.3141/2526-11
Schönfelder, S., Samaga, U., 2003. Where do you want to go today – More observations on daily mobility. Presented at the 3rd Swiss Transport Research Conference, Switzerland.
Schuessler, N., Axhausen, K., 2009. Processing Raw Data from Global Positioning Systems Without Additional Information. Transp. Res. Rec. J. Transp. Res. Board 2105, 28–36. https://doi.org/10.3141/2105-04
Schüssler, N., 2010. Accounting for similarities between alternatives in discrete choice models based on high-resolution observations of transport behaviour (Doctoral Thesis). ETH Zurich. https://doi.org/10.3929/ethz-a-006278872
Semanjski, I., Gautama, S., Ahas, R., Witlox, F., 2017. Spatial context mining approach for transport mode recognition from mobile sensed big data. Comput. Environ. Urban Syst. 66, 38–52. https://doi.org/10.1016/j.compenvurbsys.2017.07.004
Shafique, M.A., Hato, E., 2015. Use of acceleration data for transportation mode prediction. Transportation 42, 163–188. https://doi.org/10.1007/s11116-014-9541-6
Shen, L., Stopher, P.R., 2014a. Using SenseCam to pursue "ground truth" for global positioning system travel surveys. Transp. Res. Part C Emerg. Technol. 42, 76–81. https://doi.org/10.1016/j.trc.2014.02.022

Shen, L., Stopher, P.R., 2014b. Review of GPS Travel Survey and GPS Data-Processing Methods. Transp. Rev. 34, 316–334. https://doi.org/10.1080/01441647.2014.903530
Shen, L., Stopher, P.R., 2013a. Should we change the rules for trip identification for GPS travel records? Presented at the Transport and the new world city: 36th Australasian Transport Research Forum (ATRF), Brisbane, Queensland, Australia, p. 11.
Shen, L., Stopher, P.R., 2013b. A process for trip purpose imputation from Global Positioning System data. Transp. Res. Part C Emerg. Technol. 36, 261–267. https://doi.org/10.1016/j.trc.2013.09.004
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
Song, Y.-Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27, 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319. https://doi.org/10.1186/1471-2105-9-319
Stead, D., Pojani, D., 2017. The Urban Transport Crisis in Emerging Economies: A Comparative Overview, in: Pojani, D., Stead, D. (Eds.), The Urban Transport Crisis in Emerging Economies. Springer International Publishing, Cham, pp. 283–295. https://doi.org/10.1007/978-3-319-43851-1_14
Stenneth, L., Wolfson, O., Yu, P.S., Xu, B., 2011. Transportation mode detection using mobile phones and GIS information, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '11. Presented at the the 19th ACM SIGSPATIAL International Conference, ACM Press, Chicago, Illinois, p. 54. https://doi.org/10.1145/2093973.2093982
Stopher, P., FitzGerald, C., Xu, M., 2007. Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation 34, 723–741. https://doi.org/10.1007/s11116-007-9126-8
Stopher, P., FitzGerald, C., Zhang, J., 2008a. Search for a global positioning system device to measure person travel. Transp. Res. Part C Emerg. Technol., Emerging Commercial Technologies 16, 350–369. https://doi.org/10.1016/j.trc.2007.10.002
Stopher, P., FitzGerald, C., Zhang, J., 2008b. Search for a global positioning system device to measure person travel. Transp. Res. Part C Emerg. Technol. 16, 350–369. https://doi.org/10.1016/j.trc.2007.10.002
Stopher, P., Jiang, Q., FitzGerald, C., 2005. Processing GPS Data from Travel Surveys, in: 28th Australasian Transport Research Forum. Toronto, Canada.
Stopher, P.R., Daigler, V., Griffith, S., 2018. Smartphone app versus GPS Logger: A comparative study. Transp. Res. Procedia, Transport Survey Methods in the era of big data:facing the challenges 32, 135–145. https://doi.org/10.1016/j.trpro.2018.10.026
Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? Transp. Res. Part Policy Pract. 41, 367–381. https://doi.org/10.1016/j.tra.2006.09.005
Stopher, P.R., Kockelman, K., Greaves, S.P., Clifford, E., 2008. Reducing Burden and Sample Sizes in Multiday Household Travel Surveys. Transp. Res. Rec. 2064, 12–18. https://doi.org/10.3141/2064-03
Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8. https://doi.org/10.1186/1471-2105-8-25
Sugeno, M., 1985. Industrial Applications of Fuzzy Control. Elsevier Science Inc., New York, NY, USA.
Thomas, T., Geurs, K.T., Koolwaaij, J., Bijlsma, M., 2018. Automatic Trip Detection with the Dutch Mobile Mobility Panel: Towards Reliable Multiple-Week Trip Registration for Large Samples. J. Urban Technol. 25, 143–161. https://doi.org/10.1080/10630732.2018.1471874
Tsui, S., Shalaby, A., 2006. Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. Transp. Res. Rec. J. Transp. Res. Board 1972, 38–45. https://doi.org/10.3141/1972-07
Usyukov, V., 2017. Methodology for identifying activities from GPS data streams. Procedia Comput. Sci., 8th International Conference on Ambient Systems, Networks and Technologies, ANT-2017 and the 7th International Conference on Sustainable Energy Information Technology, SEIT 2017, 16-19 May 2017, Madeira, Portugal 109, 10–17. https://doi.org/10.1016/j.procs.2017.05.289
van Hassel, D., van der Velden, L., de Bakker, D., Batenburg, R., 2017. Age-related differences in working hours among male and female GPs: an SMS-based time use study. Hum. Resour. Health 15, 84. https://doi.org/10.1186/s12960-017-0258-4
Vij, A., Shankari, K., 2015. When is big data big enough? Implications of using GPS-based surveys for travel demand analysis. Transp. Res. Part C Emerg. Technol. 56, 446–462. https://doi.org/10.1016/j.trc.2015.04.025
Vincenty, T., 1975. Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. Surv. Rev. 23, 88–93. https://doi.org/10.1179/sre.1975.23.176.88
Vu, H.T., Ha, T.T., 2016. Developing Public Transport Systems - Experience from Practical Activities of Hanoi Bus System. J. Transp. - Minist. Transp. 11/2016, 80–83.
Wang, B., Gao, L., Juan, Z., 2017. A trip detection model for individual smartphone-based GPS records with a novel evaluation method. Adv. Mech. Eng. 9, 168781401770506. https://doi.org/10.1177/1687814017705066

Wang, P., Liu, G., Fu, Y., Zhou, Y., Li, J., 2017. Spotting Trip Purposes from Taxi Trajectories: A General Probabilistic Model. ACM Trans. Intell. Syst. Technol. TIST 9, 29:1–29:26. https://doi.org/10.1145/3078849
Wolf, J., Bachman, W., Oliveira, M.S., Auld, J., Mohammadian, A. (Kouros), Vovsha, P., 2014a. Applying GPS Data to Understand Travel Behavior, Volume I: Background, Methods, and Tests. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/22370
Wolf, J., Bachman, W., Oliveira, M.S., Auld, J., Mohammadian, A. (Kouros), Vovsha, P., National Cooperative Highway Research Program, Transportation Research Board, National Academies of Sciences, Engineering, and Medicine, 2014b. Applying GPS Data to Understand Travel Behavior, Volume II: Guidelines. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/23436
Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. Transp. Res. Rec. J. Transp. Res. Board 1768, 125–134. https://doi.org/10.3141/1768-15
Wolf, J., Loechl, M., Thompson, M., Arce, C., 2003a. Trip Rate Analysis in GPS-Enhanced Personal Travel Surveys, in: Transport Survey Quality and Innovation. Emerald, Bingley.
Wolf, J., Oliveira, M., Thompson, M., 2003b. Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey. Transp. Res. Rec. J. Transp. Res. Board 1854, 189–198. https://doi.org/10.3141/1854-21
Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., Axhausen, K., 2004. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. Transp. Res. Rec. J. Transp. Res. Board 1870, 46–54. https://doi.org/10.3141/1870-06
Wu, L., Yang, B., Jing, P., 2016. Travel Mode Detection Based on GPS Raw Data Collected by Smartphones: A Systematic Review of the Existing Methodologies. Information 7, 67. https://doi.org/10.3390/info7040067
Xiao, G., Juan, Z., Gao, J., 2015a. Travel Mode Detection Based on Neural Networks and Particle Swarm Optimization. Information 6, 522–535. https://doi.org/10.3390/info6030522
Xiao, G., Juan, Z., Gao, J., 2015b. Travel Mode Detection Based on Neural Networks and Particle Swarm Optimization. Information 6, 522–535. https://doi.org/10.3390/info6030522
Xiao, G., Juan, Z., Zhang, C., 2016. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transp. Res. Part C Emerg. Technol. 71, 447–463. https://doi.org/10.1016/j.trc.2016.08.008
Xiao, G., Juan, Z., Zhang, C., 2015c. Travel mode detection based on GPS track data and Bayesian networks. Comput. Environ. Urban Syst. 54, 14–22. https://doi.org/10.1016/j.compenvurbsys.2015.05.005
Yang, X., Stewart, K., Tang, L., Xie, Z., Li, Q., 2018. A Review of GPS Trajectories Classification Based on Transportation Mode. Sensors 18, 3741. https://doi.org/10.3390/s18113741
Yazdizadeh, A., Patterson, Z., Farooq, B., 2019. An automated approach from GPS traces to complete trip information. Int. J. Transp. Sci. Technol. 8, 82–100. https://doi.org/10.1016/j.ijtst.2018.08.003
Yuan, S., 2010. Méthodes d'analyse de données GPS dans les enquêtes sur la mobilité des personnes : les données manquantes et leur estimation. UNIVERSITE PARIS 1, France.
Yue, Y., Lan, T., Yeh, A.G.O., Li, Q.-Q., 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. Travel Behav. Soc. 1, 69–78. https://doi.org/10.1016/j.tbs.2013.12.002
Zadeh, L.A., 1973. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Trans. Syst. Man Cybern. SMC-3, 28–44. https://doi.org/10.1109/TSMC.1973.5408575
Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8, 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X
Zadeh, L.A., Aliev, R.A., 2018. Fuzzy Logic Theory and Applications: Part I and Part II. WSPC.
Zedeh, L.A., 1989. Knowledge representation in fuzzy logic. IEEE Trans. Knowl. Data Eng. 1, 89–100. https://doi.org/10.1109/69.43406
Zheng, Y., Chen, Y., Li, Q., Xie, X., Ma, W.-Y., 2010. Understanding Transportation Modes Based on GPS Data for Web Applications. ACM Trans Web 4, 1:1–1:36. https://doi.org/10.1145/1658373.1658374
Zimowski, M., Tourangeau, R., Ghadialy, R., Pedlow, S., 1997. Nonresponse in household travel surveys (Prepared for Federal Highway Administration No. Report DOT-T-98-4). Chicago, Illinois.

Appendix 1. Synthesizing 25 selected papers on purpose imputation

Author	Data, study area, collection device	Trip end detection	Feature selection	Method type	Method description	Accuracy	Ground truth
I. Transportation Science							
1. Wolf et al. (2001)	156 trips; 13 participants; 3-day survey; in Atlanta (US) in 2000 In-vehicle GPS devices	RB	- GI: POI data and polygon-based land use - Activity: start time, duration	RB	Matching table for 10 categories (home, shop, go to work, pick-up/drop-off, change mode, recreation, personal business, eat, unknown)	79.5%	Paper diary
2. Wolf et al. (2004)	39 participants (28 fulltime workers and 11 retirees); 30 days in Borlange (Sweden) in 2002 In-vehicle GPS devices	RB	- GI: POI data and polygon-based land use - Activity: duration, time of day, day of week - Participant: home address, profession, working hours	PB	Estimation of probability based on distance to POI. 10 categories (home, leisure, long-term shopping, daily shopping, work, school, work related, private business, pick up/drop off)	NA	NA
3. Liao et al. (2007)	4 participants; 7 days per person; 40,000 points with 10,000 segments per person	PB	- GI: street map, POI data - Activity: stay duration, time of day, day of week	SML	Hierarchical Conditional Random Fields to recognize activities and places	85.2%-90.6%	Manually label
4. McGowen and McNally (2007)	170,000 activities of 17,000 households per day; 2-day survey, in California (US) in 2000-2001; Geo-coding location reported	NA	- GI: polygon-based land use - Activity: time of day, duration, start time, activity history (repeat duration, repeat location) - Participant: age, gender, driving license, employment status.	SML	Discriminant analysis Classification (regression) tree model 5 activities (school/work, discretionary, coupling, maintenance, egress through wait for vehicle)	73% 74%	Paper diary
5. Stopher et al. (2008)	21 participants in the Sydney Household Travel and 45 ones in pilot Travel Behavior Change Program in Canberra (Australia) Personal devices	RB	- GI: land use - Activity: duration - Participant: home, school, work, frequently visited stores	RB	Heuristic rules. 10 purposes (home-based {work, education, shopping, eating, personal/medical, recreational, pick-up/drop-off, other}, non-home-based {work-other, other-other})	Over 60%	Web-based diary
6. Bohte and Maat (2009)	1,104 participants; one week per person; 3 municipalities in the Netherlands in 2007 Personal devices	RB	- GI: POI and polygon-based data - Participant: home, work address	RB	Heuristic rule based on the closest distance to POI. 7 purposes (work, study, shop, social visit, recreation, home, other)	43%	Web-based diary
7. Deng and Ji (2010)	36 participants; 226 trips; 2.5-month data collection in Shanghai (China)	RB	- GI: polygon-based data - Participant: occupation, income, family structure, age - Activity: weekdays, weekend, time of day - Trip: speed, modes, distance, duration	SML	Decision tree (C5.0) with adaptive boosting 6 purposes (work, school, pick up/drop off, shopping/recreation, business visit, others)	87.6%	Web-based diary
8. Chen et al. (2010)	25 participants for one day and 24 participants for 5 weekdays; in New York (US) Personal devices	Clustering trip ends	- GI: polygon-based land use - Activity: time of day, duration, historical activity frequency	RB & PB	Multinomial logit for high-density areas Deterministic matching for low-density areas 4 purposes (work/school, personal business, shopping, recreation) for two groups (home-based, non-home-based)	67% – 78%	Paper diary

9. Lu et al. (2012)	3,188 trips; Twin Cities Metro Area, Minnesota (US) in 2008 In-vehicle devices	Engine on/off signal	- GI: polygon-based data - Participant: income, race, education level - Trip: start and end time of current, previous and next trip, duration	SML	Decision tree 10 purposes (home, work, shopping, daycare, dining, driving others, services, school, social/recreation, others)	73.4%	Web-based diary
10. Shen and Stopher (2013b)	4,133 trips; Greater Cincinnati region (USA) in 2009 Personal devices	NA	- GI: land use - Activity: duration, start time, end time, tour - Participant: home, school, work, frequently visited stores	RB	Heuristic rules. 5 purposes (home, work, education, shopping, other)	66.5%	Web-based diary
11. Oliveira et al. (2014)	1,352 participants (subsample of Atlanta Household Survey); 22,734 activities (10,512 non-home ones); in 2011 in Atlanta (US) Personal devices	NA	- GI: land use - Activity: duration - Participant: occupation, age, household's information, etc. - Trip: trip mode, mode of next trip, etc.	PB & SML	Two-level Nested Logit Model Decision tree 12 purposes (*mode transfer*, work, pick up, drop off, maintenance, work, work-related, attending class, shopping, eating, religious ones, entertainment, social visit)	60% 65%	Paper diary
12. Montini et al. (2014)	156 participants; 6,938 activities; in 2012 in Zurich (Switzerland) Personal devices every 1s	Clustering trip ends	- Activity: day of week, duration, start time. - Participant: education level, age, income, marital status, home and work address - Trip: mode before and after activity	SML	Random forest. 8 purposes (*mode transfer*, home, work, shopping, recreation, business, pick up/drop off, others)	84.4%	Web-based diary
13. Feng and Timmermans (2015)	329 participants; 10,545 activities except "return home" purpose; 2012-2013 in Rijnmond Region (the Netherlands) Personal devices every 3s	NA	- GI: POI data - Activity: start time, duration - Trip: mode (walk, bike, bus, car, taxi, tram, metro and train)	SML	Random forest Decision tree Bayesian belief network 11 purposes (paid work, daily shopping, help parents/children, non-daily shopping, recreation, social, voluntary work, service, leisure, pick up, study)	96.8% 69.8% 46.2%	Web-based diary
14. Xiao et al. (2016)	321 participants; 2,409 person days; 7,039 activities; 10/2013-06/2015 in Shanghai (China) Smartphones every 1s	NA	- GI: POI and polygon-based data - Participant: age, gender, income, occupation, child in household - Activity: duration, time of week, start time - Trip: mode (walk, bike, e-bike, bus, car)	SML	Artificial neural network with particle swarm optimization Artificial neural network with back propagation Support vector machine Bayesian network Multinomial logit 6 purposes (home, work/education, eat out, shopping, social visit, pick up/drop off)	96.5% 93.8% 87.1% 86.6% 80.2%	Telephone-based collection
15. Meng et al., (2017)	8,631 GPS trajectories in California Travel Household Survey	NA	- GI: POI data from Google places - Activity: time and duration, previous activity - Trip: modes - Social network: Twitter	SML	Bayesian network Random forest K-nearest neighbour Artificial neural network Support vector machine	87.8% 60.1% 38.5% 45.4% 33.7%	Available but not reported

16. Gong et al. (2018)	20 participants; 9,981 trips (5,512 in summer & 4,469 in winter); 12/2012-04/2013 in Hakodate (Japan) Smartphones every 30s	RB	- GI: POI data - Participant: home & work addresses, gender, age, household's information (size, vehicle number), driving frequency - Activity: time of day, day of time, time of week - Trip: duration, length - Weather: temperature, snow accumulation and precipitation	SML	Random forest 6 purposes (home, commute, meal, shopping, recreation, others)	Only for separate purposes	Manual labelling
17. Cui et al. (2018)	10,474 participants (42,431 households); Bay Area, California (US) in 2012-2013 Personal and in-vehicle devices	Comparison between GPS data and ground truth	- GI: POI data from Google places - Participant: age, occupation/employment - Trip: departure time, duration, distance - Social network: Twitter	SML	Bayesian neural network Support vector machine Artificial neural network K-nearest neighbor Random forest 6 purposes (eat out, personal, recreation, shopping, transportation, education)	90.5% 49.9% 48.2% 45.1% 61.9%	Web-based and mail-based diary
18. Yazdizadeh et al. (2019)	6,845 participants with validated data; 102,904 trips; 10-11/2016 in Montreal (Canada) Smartphones	RB	- GI: land use data - Participant: age, occupation, gender, home's neighbourhood value, home/work/education addresses - Activity: day of week, time of day - Trip: travel time, modes, origin and destination in comparison with Montreal Island - Social network: Foursquare	SML	Random forest 6 purposes (education, health, leisure, shopping, return home, work)	71.26%	Prompted recall by smartphone
19. Krause and Zhang, (2019)	260 participants, 36,000 trips, 70 days; 10/2011-02/2012 in 22 states of US On-board devices every minute	RB	- GI: POI data - Participants' characteristics - Trip: start-of-trip information	SML	Decision tree (J48) 7 purposes (home, school, shopping, social, work, driving, other)	NA	Online web-based diary
II. Human Geography							
20. Furletti et al. (2013)	28 participants; around 30,000 trips; in Flanders (Belgium) In-vehicle devices	NA	- GI: POI data, opening/working time	PB	Gravity model based on distance 6 purposes (services, food, daily shopping, shopping, education, leisure)	43%	Daily diaries
21. Reumers et al. (2013)	1,250 households; one-week survey; 290 activities (test set); in Flanders (Belgium) in 2006 – 2007 GPS-enabled Personal Digital Assistants (PDA) every 1s	NA	- Activity: start time, duration	SML	Decision tree. 6 purposes (home, work, bring-get, leisure, shopping, social)	75.9%	Paper diary
22. Gong et al. (2016)	6,600 taxis; one-week survey; 06/2019 in Shanghai (China) In-vehicle devices every 10s	Signal of whether passenger on-board	- GI: POI data, opening time - Activity: drop-off time	PB	Bayes' rules based on spatial and temporal constraints 9 purposes (in-home, work-related, transfer, dining, shopping, recreation, schooling, lodging, medical)	NA; use spatio-temporal pattern; trip length; trip direction	NA, use common sense

23. Usyukov (2017)	108 cyclists, 541 survey days (5 days per cyclist), 2011 in Waterloo (Canada) Low-cost personal devices every 5s	RB	- GI: land use data - Trip: start time, end time	RB & PB	Heuristic rule for home activities Discrete choice for non-home activities 3 purposes (home, work, other)	NA, proportion of activities	NA, use Transportation Tomorrow Survey (2006)
24. P. Wang et al. (2017)	188,363 trajectories with 13,455 pick-up and 17,926 drop-off points; in NYC (USA), 01-06/2015 In-vehicle devices	Signal of whether passenger on-board	- Trip: start time - Social network: Foursquare	PB	General probabilistic model 3 purposes (work-oriented, entertainment-oriented, nightlife-oriented)	NA, use spatio-temporal pattern	NA, use common sense
25. Chen et al. (2019)	110,000* of 13 million taxi trips; in Manhattan, NYC (USA) In-vehicle devices	Signal of whether passenger on-board	- Trip: time of day, day of week, duration - Social network: Foursquare	UML	Autoencoder and K-means 5 purposes (dining, recreation, work, homing, others)	NA, use spatio-temporal pattern	NA, use common sense

GI is “Geographic Information”; RB is “Rule-based”; PB is ”Probability-based” ; SML is “Supervised Machine Learning”; UML is “Unsupervised Machine Learning”; NA is “Not Available”; 100,000 for training Autoencoder and 10,000 for evaluating model.

Appendix 2: Questionnaire for background information collection

BẢNG CÂU HỎI VỀ THÔNG TIN
CỦA NGƯỜI THAM GIA KHẢO SÁT DỮ LIỆU ĐI LẠI BẰNG GPS
QUESTIONNAIRE FOR BACKGROUND INFORMATION
OF PARTICIPANTS IN THE GPS-BASED MOBILITY SURVEY

Số được cấp *(Given number):*……. (Đây là số sử dụng để nhập vào phần mềm (app) trong quá trình cài đặt ban đầu) *(This is the number which you will type in the process of installing the app)*

Họ và tên – Không bắt buộc *(Full name - non-mandatory)*:……..
Địa chỉ nhà: Số/Đường *(Number/Road):*……..……..……..……..……..…
(Home address) Phường/thôn xã *(Ward):*……..……..……...……..……
Quận/Huyện *(District):*……..……

1. Giới tính *(Gender):*

- ☐ A1. Nam *(Male)*
- ☐ A2. Nữ *(Female)*

2. Năm sinh *(Year of birth):*……..

3. Trong hộ gia đình đang chung sống, vị trí (cao nhất) của bạn hiện là *(In your household, you are):*

- ☐ A1. Ông/Bà *(Grandparent)*
- ☐ A2. Cha *(Father)*
- ☐ A3. Mẹ *(Mother)*
- ☐ A4. Con/ Cháu *(Child)*
- ☐ A5. Khác, làm rõ *(Other and specify)*……

4. Bạn/Ông/Bà có những loại giấy tờ nào sau đây *(What type of licenses/documents do you have)*:

- ☐ A1. Bằng lái xe ô tô *(Car driving license)*
- ☐ A2. Bằng lái xe máy *(Motorcycle driving license)*
- ☐ A3. Vé tháng xe bus *(Monthly public transport ticket)*

5. Nghề nghiệp chính của Bạn/Ông/Bà *(Main occupation)*:

- ☐ A1. Học sinh *(Pupil)*
- ☐ A2. Sinh viên *(Student)*
- ☐ A3. Viên chức/ Nhân viên văn phòng/ Giáo viên/ Bác sỹ *(Staff, Officer, Teacher, Doctor)*
- ☐ A4. Công nhân/ Thợ/ Nông dân *(Worker, Farmer)*
- ☐ A5. Đang tìm việc *(Job seeker)*
- ☐ A6. Về hưu *(Retired)*
- ☐ A7. Tự kinh doanh, làm việc tự do *(Freelance, Entrepreneur)*
- ☐ A8. Khác, làm rõ *(Other and specify)*……

6. Công việc của Bạn/Ông/Bà đòi hỏi phải thường xuyên phải rời khỏi trụ sở/văn phòng làm việc chính *(Do you have a dynamic job that requires you to frequently work and travel outside your main office)*:

- ☐ A1. Có *(Yes)*
- ☐ A2. Không *(No)*

7. Trình độ học vấn (cao nhất) của Bạn/Ông/Bà *(Highest educational level)*:

- ☐ A1. Cấp 2 - Trung học cơ sở *(Elementary school)*
- ☐ A2. Cấp 3 - Trung học phổ thông *(High school)*
- ☐ A3. Đại học *(Graduate qualification)*
- ☐ A4. Sau đại học *(Post-graduate qualification)*

8. Loại nhà bạn đang sử dụng *(Type of your housing):*

- ☐ A1. Nhà riêng *(Private house)*
- ☐ A2. Căn hộ, chung cư *(Apartment)*
- ☐ A3. Loại khác *(Other)*

9. Tổng thu nhập của gia đình Bạn/Ông/Bà (triệu/tháng) *(Household's total monthly income):*

- ☐ A1. Dưới 5.5 *(<250€)*
- ☐ A2. 5.5 – dưới 11 *(250€ - <500€)*
- ☐ A3. 11 – dưới 16.5 *(500€ - <750€)*
- ☐ A4. 16.5 – dưới 22 *(750€ - <1000€)*
- ☐ A5. 22 – dưới 27.5 *(1000€ - <1250€)*
- ☐ A6. 27.5 – dưới 33 *(1250€ - <1500€)*
- ☐ A7. 33 – dưới 38.5 *(1500€ - <1750€)*
- ☐ A8. 38.5 – dưới 44 *(1750€ - <2000€)*
- ☐ A9. 44 – dưới 49.5 *(2000€ - <2250€)*
- ☐ A10. 49.5 – dưới 55 *(2250€ - <2500€)*
- ☐ A11. 55 trở lên *(2500€ or over)*

10. Số lượng phương tiện dưới đây gia đình Bạn/Ông/Bà có *(Number of following vehicle types available in your household):*

A1. …… xe đạp *(Bicycle)*

A2. …… xe đạp điện *(E-bicycle)*

A3. …… xe máy *(Motorcycle)*

A4. …… xe con *(Car)*

11. Số lượng thành viên trong gia đình Bạn/Ông/Bà *(Number of members in your household):*

……

Trong đó có …… trẻ em dưới 15 tuổi *(Number of children under 15)*

Trân trọng cám ơn!

(Thank you for your participation)

Appendix 3: Results of testing sensitivity of bus detection to thresholds of speed and distance

In case the speed changes between 0.7m/s and 3.3m/s whilst the distance fixes of 100m

Distance (m)	Speed (m/s)	Correctly predicted bus segment number	Predicted bus segment number	Actual bus segment number	Precision	Recall	Fscore
100	0.7	61	77	97	79.22%	62.89%	70.11%
100	0.9	61	77	97	79.22%	62.89%	70.11%
100	1.1	61	77	97	79.22%	62.89%	70.11%
100	1.3	61	77	97	79.22%	62.89%	70.11%
100	1.5	61	77	97	79.22%	62.89%	70.11%
100	1.7	61	77	97	79.22%	62.89%	70.11%
100	1.9	61	77	97	79.22%	62.89%	70.11%
100	2.1	71	87	97	81.61%	73.20%	77.17%
100	2.3	83	102	97	81.37%	85.57%	83.42%
100	2.5	92	114	97	80.70%	94.85%	87.20%
100	2.7	94	128	97	73.44%	96.91%	83.56%
100	2.9	95	141	97	67.38%	97.94%	79.83%
100	3.1	95	158	97	60.13%	97.94%	74.51%
100	3.3	95	171	97	55.56%	97.94%	70.90%

In case the distance changes between 30m and 150m whilst the speed fixes of 2.5m/s

Distance (m)	Speed (m/s)	Correctly predicted bus segment number	Predicted bus segment number	Actual bus segment number	Precision	Recall	Fscore
30	2.5	13	16	97	81.25%	13.40%	23.01%
40	2.5	19	22	97	86.36%	19.59%	31.93%
50	2.5	25	31	97	80.65%	25.77%	39.06%
60	2.5	33	39	97	84.62%	34.02%	48.53%
70	2.5	49	58	97	84.48%	50.52%	63.23%
80	2.5	62	74	97	83.78%	63.92%	72.51%
90	2.5	79	98	97	80.61%	81.44%	81.03%
100	2.5	92	114	97	80.70%	94.85%	87.20%
110	2.5	92	133	97	69.17%	94.85%	80.00%
120	2.5	92	150	97	61.33%	94.85%	74.49%
130	2.5	93	176	97	52.84%	95.88%	68.13%
140	2.5	94	192	97	48.96%	96.91%	65.05%
150	2.5	94	208	97	45.19%	96.91%	61.64%

Appendix 4:
Prediction results of simple hierarchical process and all-in-one process

For case of all-in-one process based on Random Forest

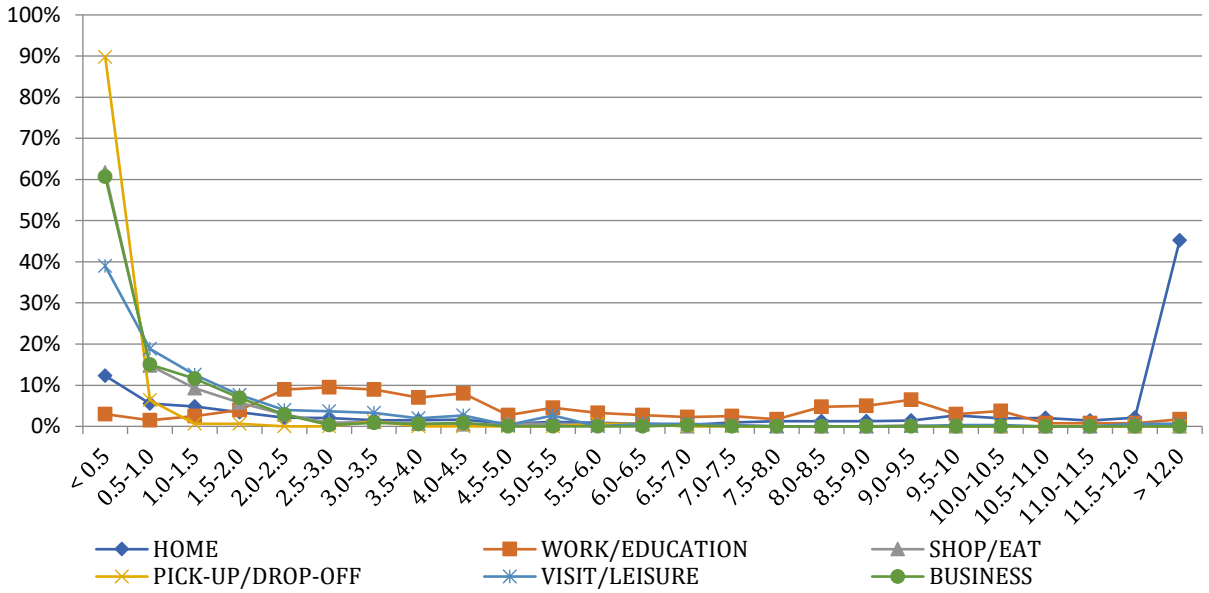
		Predicted						Recall	Fscore
		Walk	Bike	Bus	Motorcycle	Car	Total		
Reported	Walk	178	0	0	8	0	186	95.7%	93.4%
	Bike	2	4	0	19	1	26	15.4%	25.0%
	Bus	0	0	2	17	0	19	10.5%	16.0%
	Motorcycle	12	2	3	276	22	315	87.6%	79.9%
	Car	3	0	1	56	92	152	60.5%	68.9%
	Total	195	6	6	376	115	698		
Precision		*91.3%*	*66.7%*	*33.3%*	*73.4%*	*80.0%*		**Accuracy**	**79.1%**

For case of simple hierarchical process based on rules

		Predicted						Recall	Fscore
		Walk	Bike	Bus	Motorcycle	Car	Total		
Reported	Walk	632	109	2	15	0	758	83.4%	89.1%
	Bike	0	58	9	37	0	104	55.8%	27.2%
	Bus	0	4	56	35	2	97	57.7%	17.4%
	Motorcycle	22	140	368	668	47	1245	53.7%	61.8%
	Car	6	11	111	163	296	587	50.4%	63.5%
	Total	660	322	546	918	345	2791		
Precision		95.8%	18.0%	10.3%	72.8%	85.8%		**Accuracy**	**61.3%**

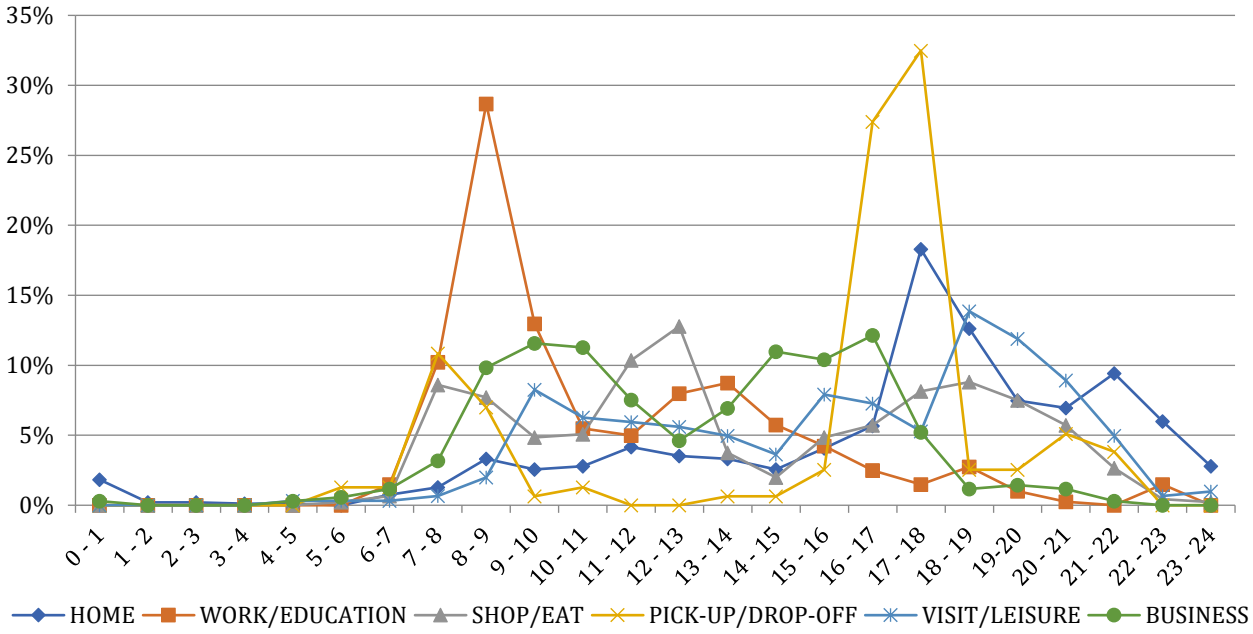
Appendix 5: Duration distribution of activities

Duration (hour)	Home		Work/ Education		Shopping/ Eating		Pick-up/ Drop-off		Visit/ Leisure		Business	
	No	*Percent*	*No*	*Percent*	*No*	*Percent*	*No*	*Percent*	*No*	*Percent*	*No*	*Percent*
< 0.5	115	12.30%	12	2.99%	280	61.67%	141	89.81%	118	38.94%	210	60.69%
0.5-1.0	52	5.56%	6	1.50%	67	14.76%	10	6.37%	57	18.81%	52	15.03%
1.0-1.5	45	4.81%	10	2.49%	42	9.25%	1	0.64%	38	12.54%	40	11.56%
1.5-2.0	32	3.42%	16	3.99%	26	5.73%	1	0.64%	23	7.59%	24	6.94%
2.0-2.5	20	2.14%	36	8.98%	13	2.86%	0	0.00%	12	3.96%	10	2.89%
2.5-3.0	19	2.03%	38	9.48%	4	0.88%	0	0.00%	11	3.63%	1	0.29%
3.0-3.5	14	1.50%	36	8.98%	5	1.10%	2	1.27%	10	3.30%	3	0.87%
3.5-4.0	14	1.50%	28	6.98%	3	0.66%	0	0.00%	6	1.98%	2	0.58%
4.0-4.5	15	1.60%	32	7.98%	2	0.44%	0	0.00%	8	2.64%	3	0.87%
4.5-5.0	6	0.64%	11	2.74%	3	0.66%	0	0.00%	1	0.33%	0	0.00%
5.0-5.5	10	1.07%	18	4.49%	3	0.66%	0	0.00%	8	2.64%	0	0.00%
5.5-6.0	8	0.86%	13	3.24%	2	0.44%	1	0.64%	0	0.00%	0	0.00%
6.0-6.5	6	0.64%	11	2.74%	2	0.44%	1	0.64%	2	0.66%	0	0.00%
6.5-7.0	3	0.32%	9	2.24%	0	0.00%	0	0.00%	2	0.66%	1	0.29%
7.0-7.5	9	0.96%	10	2.49%	1	0.22%	0	0.00%	1	0.33%	0	0.00%
7.5-8.0	12	1.28%	7	1.75%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
8.0-8.5	12	1.28%	19	4.74%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
8.5-9.0	12	1.28%	20	4.99%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
9.0-9.5	13	1.39%	26	6.48%	1	0.22%	0	0.00%	0	0.00%	0	0.00%
9.5-10	25	2.67%	12	2.99%	0	0.00%	0	0.00%	1	0.33%	0	0.00%
10.0-10.5	18	1.93%	15	3.74%	0	0.00%	0	0.00%	1	0.33%	0	0.00%
10.5-11.0	19	2.03%	3	0.75%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
11.0-11.5	13	1.39%	3	0.75%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
11.5-12.0	20	2.14%	3	0.75%	0	0.00%	0	0.00%	2	0.66%	0	0.00%
> 12.0	423	45.24%	7	1.75%	0	0.00%	0	0.00%	2	0.66%	0	0.00%
Total	**935**	**100%**	**401**	**100%**	**454**	**100%**	**157**	**100%**	**303**	**100%**	**346**	**100%**



Appendix 6: Start time distribution of activities

Start time (hour)	Home		Work/ Education		Shopping/ Eating		Pick-up/ Drop-off		Visit/ Leisure		Business	
	No	*Percent*	*No*	*Percent*	*No*	*Percent*	*No*	*Percent*	*No*	*Percent*	*No*	*Percent*
0 - 1	17	1.82%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	1	0.29%
1 - 2	2	0.21%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
2 - 3	2	0.21%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
3 - 4	1	0.11%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%
4 - 5	1	0.11%	0	0.00%	0	0.00%	0	0.00%	1	0.33%	1	0.29%
5 - 6	0	0.00%	0	0.00%	1	0.22%	2	1.27%	1	0.33%	2	0.58%
6 -7	7	0.75%	6	1.50%	3	0.66%	2	1.27%	1	0.33%	4	1.16%
7 - 8	12	1.28%	41	10.22%	39	8.59%	17	10.83%	2	0.66%	11	3.18%
8 - 9	31	3.32%	115	28.68%	35	7.71%	11	7.01%	6	1.98%	34	9.83%
9 - 10	24	2.57%	52	12.97%	22	4.85%	1	0.64%	25	8.25%	40	11.56%
10 - 11	26	2.78%	22	5.49%	23	5.07%	2	1.27%	19	6.27%	39	11.27%
11 - 12	39	4.17%	20	4.99%	47	10.35%	0	0.00%	18	5.94%	26	7.51%
12 - 13	33	3.53%	32	7.98%	58	12.78%	0	0.00%	17	5.61%	16	4.62%
13 - 14	31	3.32%	35	8.73%	17	3.74%	1	0.64%	15	4.95%	24	6.94%
14 - 15	24	2.57%	23	5.74%	9	1.98%	1	0.64%	11	3.63%	38	10.98%
15 - 16	38	4.06%	17	4.24%	22	4.85%	4	2.55%	24	7.92%	36	10.40%
16 - 17	53	5.67%	10	2.49%	26	5.73%	43	27.39%	22	7.26%	42	12.14%
17 - 18	171	18.29%	6	1.50%	37	8.15%	51	32.48%	16	5.28%	18	5.20%
18 - 19	118	12.62%	11	2.74%	40	8.81%	4	2.55%	42	13.86%	4	1.16%
19-20	70	7.49%	4	1.00%	34	7.49%	4	2.55%	36	11.88%	5	1.45%
20 - 21	65	6.95%	1	0.25%	26	5.73%	8	5.10%	27	8.91%	4	1.16%
21 - 22	88	9.41%	0	0.00%	12	2.64%	6	3.82%	15	4.95%	1	0.29%
22 - 23	56	5.99%	6	1.50%	2	0.44%	0	0.00%	2	0.66%	0	0.00%
23 - 24	26	2.78%	0	0.00%	1	0.22%	0	0.00%	3	0.99%	0	0.00%
Total	**935**	**100%**	**401**	**100%**	**454**	**100%**	**157**	**100%**	**303**	**100%**	**346**	**100%**



Appendix 7:
Results of purpose inference by five models using different features

Results of model 1: Using time-related features

		Predicted						Recall	F^{class}
		Home	*Work/ Education*	*Shopping/ Eating*	*Pick-up/ Drop-off*	*Visit/ Leisure*	*Business*		
Reported	*Home*	168	20	21	6	7	12	**71.8%**	**74.7%**
	Work/ Education	12	77	2	0	0	9	**77.0%**	**73.3%**
	Shopping/ Eating	16	1	55	5	14	23	**48.2%**	**45.1%**
	Pick-up/ Drop-off	1	0	9	25	1	3	**64.1%**	**62.5%**
	Visit/ Leisure	17	9	14	2	21	13	**27.6%**	**33.9%**
	Business	2	3	29	3	5	44	**51.2%**	**46.3%**
Precision		**77. 8%**	**70.00%**	**42.3%**	**61.0%**	**43.8%**	**42.3%**	*Accuracy* **60.1%**	F^{model} **56.0%**

Results of model 2: Using time- and user-related features

		Predicted						Recall	F^{class}
		Home	*Work/ Education*	*Shopping/ Eating*	*Pick-up/ Drop-off*	*Visit/ Leisure*	*Business*		
Reported	*Home*	170	20	16	5	10	13	**72.6%**	**74.1%**
	Work/ Education	17	75	3	0	1	4	**75.0%**	**71.8%**
	Shopping/ Eating	16	2	56	11	12	17	**49.1%**	**49.6%**
	Pick-up/ Drop-off	1	0	10	24	1	3	**61.5%**	**55.2%**
	Visit/ Leisure	17	9	11	2	26	11	**34.2%**	**40.3%**
	Business	4	3	16	6	3	54	**62.8%**	**57.4%**
Precision		**75.6%**	**68.8%**	**50.0%**	**50.0%**	**49.1%**	**52.9%**	*Accuracy* **62.4%**	F^{model} **58.1%**

Results of model 3: Using time- and user- and predicted mode-related features

		Predicted						Recall	F^{class}
		Home	*Work/ Education*	*Shopping/ Eating*	*Pick-up/ Drop-off*	*Visit/ Leisure*	*Business*		
Reported	*Home*	168	20	17	4	9	16	**71.8%**	**73.8%**
	Work/ Education	15	75	4	0	0	6	**75.0%**	**72.1%**
	Shopping/ Eating	14	1	57	8	13	21	**50.0%**	**51.4%**
	Pick-up/ Drop-off	2	0	8	26	1	2	**66.7%**	**65.0%**
	Visit/ Leisure	18	9	12	2	23	12	**30.3%**	**36.5%**
	Business	4	3	10	1	4	64	**74.4%**	**61.8%**
Precision		**76.0%**	**69.4%**	**52.8%**	**63.4%**	**46.0%**	**52.9%**	*Accuracy* **63.6%**	F^{model} **60.1%**

Results of model 4: Using time- and user- and actual mode-related features

		Predicted						Recall	F^{class}
		Home	*Work/ Education*	*Shopping/ Eating*	*Pick-up/ Drop-off*	*Visit/ Leisure*	*Business*		
Reported	*Home*	166	21	20	5	8	14	**70.9%**	**73.9%**
	Work/ Education	16	73	5	0	0	6	**73.0%**	**70.5%**
	Shopping/ Eating	14	1	58	5	13	23	**50.9%**	**51.6%**
	Pick-up/ Drop-off	1	0	7	27	1	3	**69.2%**	**69.2%**
	Visit/ Leisure	15	9	12	1	27	12	**35.5%**	**41.9%**
	Business	3	3	9	1	4	66	**76.7%**	**62.9%**
Precision		**77.2%**	**68.2%**	**52.3%**	**69.2%**	**50.9%**	**53.2%**	*Accuracy* **64.3%**	F^{model} **61.7%**

Results of model 5: Using time- and user- and actual mode- and home-related features

		Predicted						Recall	F^{class}
		Home	*Work/ Education*	*Shopping/ Eating*	*Pick-up/ Drop-off*	*Visit/ Leisure*	*Business*		
Reported	*Home*	220	5	5	0	2	2	**94.0%**	**92.1%**
	Work/ Education	5	81	6	0	2	6	**80.0%**	**80.4%**
	Shopping/ Eating	9	2	58	7	15	23	**58.8%**	**59.6%**
	Pick-up/ Drop-off	6	0	7	21	2	3	**56.4%**	**62.9%**
	Visit/ Leisure	2	11	17	2	32	12	**47.4%**	**51.4%**
	Business	4	3	10	2	5	62	**72.1%**	**66.7%**
Precision		**90.2%**	**80.8%**	**60.4%**	**71.0%**	**56.3%**	**62.0%**	*Accuracy* **75.0%**	F^{model} **68.8%**