# ABSTRACT


| | |
|---|---|
| Title of Thesis: | IMPUTING SOCIAL DEMOGRAPHIC INFORMATION BASED ON PASSIVELY COLLECTED LOCATION DATA AND MACHINE LEARNING METHODS |
| | Yixuan Pan, Master of Science, 2018 |
| Thesis Directed By: | Professor Lei Zhang, Department of Civil and Environmental Engineering |

Multiple types of passively collected location data (PCLD) have emerged during the past 20 years. Its capability in travel demand analysis has also been studied and revealed. Unlike the traditional surveys whose sample is designed efficiently and carefully, PCLD features a non-probabilistic sample of dramatically larger size. However, PCLD barely contains any ground truth for both the human subjects involved and the movements they produce. The imputation for such missing information has been evaluated for years, including origin and destination, travel mode, trip purpose, etc. This research intends to advance the utilization of PCLD by imputing social demographic information, which can help to create a panorama for the large volume of travel behaviors observed and to further develop a rational weighting procedure for PCLD. The Conditional Inference Tree model has been employed to address the problems because of its abilities to avoid biased variable selection and overfitting.

# IMPUTING SOCIAL DEMOGRAPHIC INFORMATION BASED ON PASSIVELY COLLECTED LOCATION DATA AND MACHINE LEARNING METHODS

by

Yixuan Pan

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2018

Advisory Committee:
Professor Lei Zhang, Chair
Professor Paul Schonfeld
Professor Vanessa Frias-Martinez

# Dedication

To my beloved parents Fenping Li and Tongbin Pan, who always provide me with comfort and support.

# Acknowledgements

The research was partially funded by National Science Foundation (Award #1649189: "RAPID: Transit Network Disruption, Service Reliability, and Travel Behavior") and the National Transportation Center at University of Maryland. Opinions herein do not necessarily represent the views of the research sponsors. The author is responsible for the statements in the thesis.

First, I would like to express my sincere gratitude to my advisor, Dr. Lei Zhang, for his continuous support and guidance during the past three years. His rich experience and insightful comments have always helped to lead me out of my "dilemma zone".

I would also thank Dr. Paul Schonfeld and Dr. Vanessa Frias-Martinez for their dedication to my committee. I am extremely grateful for Dr. Schonfeld being willing to serve on my committee at the last minute and I am highly appreciative of Dr. Frias-Martinez's expertise in passively collected location data. Without their valuable suggestions, I would not be able to finish the thesis.

I owe my special thanks to my team members: Cory Krause and Liang Tang. Cory did an excellent job on the sample design and data collection for the 2011/2012 GPS travel survey. Liang is as wonderful as a data master. He provided me with generous help in the trip processing part. His knowledge about machine learning also shed some light on my problems.

I also want to thank the project group members who helped to conduct the SafeTrack survey: the project PI, Dr. Shanjiang Zhu, and the members from George Mason

University: Zhuo Yang and Hamza Masud. In addition, I would like to thank Dr. Chenfeng Xiong for his leadership in the project and my group members for their help in distributing the questionnaires and flyers: Bo Peng, Arash Asadabadi, Jun Zhao, Hang Yang, and Weiyi Zhou. Thanks to Arefeh Nasri for introducing the Smart Location Database to me.

I feel thankful to have many kind and sincere friends, who always make me relaxed: Yaou Zhang, Minha Lee, Zheng Zhu, Yao Cheng, Liu Xu, Han Dong, Mengpin Ge, Di Yang, and Yumu Zhang.

Last but not least, I have my deepest gratitude to my father, Tongbin Pan, and my mother, Fenping Li. They relieve any anxiety I may have with the promise of home.

It is impossible to remember all, and I apologize to those I've left out. Thank you all!

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| ANN | Artificial Neural Networks |
| AUC | Area under the Curve |
| CBG | Census Block Group |
| CDR | Call Detail Record |
| ETD | Electronic Travel Diary |
| FHWA | Federal Highway Administration |
| GPS | Global Positioning System |
| NHTS | National Household Travel Survey |
| OD | Origin and Destination |
| PCLD | Passively Collected Location Data |
| POI | Point of Interest |
| SVM | Support Vector Machine |
| SLD | Smart Location Database |
| VI | Variable Importance |

# Chapter 1: Introduction

As the travel demand analysis keeps advancing for the past several decades, data always play an important role and serve as the foundation of nearly all the studies. From count data derived by roadway sensors to behavior data collected by complicated travel surveys, data benefit the researchers by providing support for transportation planning, infrastructure construction, traffic management and so on. Along with the development of Global Positioning System (GPS) and handheld devices, passively collected location data (PCLD) emerged at the end of 20$^{th}$ century. From then on, the potential of PCLD has been deliberately excavated by scholars in the transportation sector. However, the anonymousness of PCLD still remains unsolved, which has obstructed the exhaustive exploitation of PCLD. The thesis seeks to address the problem by developing an imputation method for social demographic information. The expected benefits will be uncovering the representativeness of PCLD sample and further extending the application of PCLD.

## *1.1. Background*

Travel behaviors, such as trip origin and destination, travel mode, and trip purpose, have been studied as a major topic in the transportation field for decades. One of the most popular and prevalent data sources is the traditional household travel survey, who has an efficiently designed area sample. The National Household Travel Survey (NHTS) is conducted by Federal Highway Administration (FHWA) every five to eight years and includes more than 120,000 households across the nation [1].

Meanwhile, the local agencies also tend to organize the regional household travel survey to recruit more sample units for the specific area, which is usually around ten thousand households [2]. The household travel survey is mainly fulfilled through telephone interviews in order to document the travel diaries of the entire household on one typical weekday. It has its advantages in collecting all-around information on the interviewed household and providing reliable details on the reported trips. On the other hand, there exist problems like underreported trips, expensive survey costs, short-term travel diary, etc.

As the drawbacks of the traditional survey methods kept unsolved, an emerging data source has drawn researchers' attention in the past years - passively collected location data (PCLD). PCLD gets its name from the origination that it is not generated by people's subjective report but through a positioning device along with the traveler. The device can be as ordinary as a cell phone with location services or as professional as an in-vehicle GPS device. The locations of the device along with the timestamps will be recorded despite the fact that different devices can produce location data with different accuracy and frequency.

PCLD has also experienced its own evolution, from intentionally recruiting a sample of limited size to employing a large portion of population who own a device with location services. The prevailing PCLD nowadays features an enormous and non-probabilistic sample, such as call detail records (CDRs), cell phone GPS data, and social media location-based services. In spite of the attractiveness, the extremely large samples of PCLD are usually anonymous in order to protect the privacy of the

users, which only allows a rough weighting procedure. Barely with any ground truth information about the users and their trips, the primary step to utilize PCLD is to impute trip information from the timestamped positions. A variety of studies have attempted to detect the trip origin and destination, travel mode, trip purpose or activity type. Although the trip information underneath can be inferred at a relatively high accuracy, the representativeness of PCLD sample remains unclear and becomes a limit to exploit PCLD, which inspires the research topic of this thesis.

### *1.2. Objectives*

The objective of this study is searching for a method to impute the sensitive individual-level social demographics based on PCLD, including gender, age, education, and household income. To fulfill the objective, several tasks will be accomplished: 1) evaluating the state-of-the-art methods and algorithms for similar problem specification; 2) for demographic classification, exploring what are the significant attributes depicting travel behaviors; 3) examining the model performance through case studies using two typical examples for the prevailing PCLD; 4) discussing the applications of social demographic imputation.

The evaluation of the methods will be done based on relatively small datasets with full knowledge. In order to demonstrate the feasibility of applying such methods to the enormous and anonymous PCLD in reality, the theoretical strength and the model adaptability will be both considered in addition to the prediction accuracy.

## *1.3. Contributions*

This is the first study that imputes sensitive individual-level social demographics including income level based on PCLD. This is also the first one utilizing a multimodal trajectory dataset supported by multiple positioning systems for demographic imputation.

### 1.3.1. Data-wise

The thesis gives a comprehensive literature review on the evolution and utilization of PCLD. The two examples of PCLD involved are an in-vehicle GPS dataset and a smartphone location dataset, both of which are the trending types of PCLD nowadays. Furthermore, the multimodal trajectory dataset is studied for demographic imputation for the first time. The thesis will demonstrate the challenges in processing PCLD of various data qualities, especially related to trajectories of rather low quality. The data quality of the trajectories here is discussed in two dimensions: the recording frequency and the coordinate accuracy.

Other than PCLD, the only additional data source needed is the Smart Location Database (SLD) released by the United States Environmental Protection Agency (EPA) [3]. SLD contains nearly a hundred attributes summarizing land use, demographics, transportation accessibility, etc. at census block group (CBG) level nationwide. Hence the findings and experience of this study can be easily tested and evaluated in other geographic settings.

#### 1.3.2. Methodology-wise

In this study, a new tree-based model—conditional inference tree (CIT)—is examined and applied to impute social demographics. The model has been selected based on its capability of preventing variable selection bias and overfitting problems.

A thorough analysis on the feature set construction is delivered. Four aspects are taken into account including intuitive travel behaviors, home/work geographic characteristics, frequency of visiting different POIs, and frequency of trips with imputed purposes. This is the first study incorporating the imputed features of travel behaviors into the demographic prediction. The prediction strength of the features above is examined and compared.

#### 1.3.3. Application-wise

The study aims at further exploitation of PCLD in practice. The social demographics are widely and continuously employed as the basis of sampling and weighting. As the drawback of PCLD being the unclear representativeness, the study attempts to develop a method to impute the social demographics of PCLD sample. It advances the utilization of PCLD while avoids violating the privacy.

### *1.4. Outline*

The remainder of the thesis is organized as follows.

Chapter 2 provides a comprehensive literature review concerning the evolution of PCLD, the investigation of PCLD, and the state-of-the-art imputation methods to utilize PCLD. Chapter 3 introduces the two PCLD datasets used in this study, which

are representative of the prevailing PCLD, and the Smart Location Database (SLD) as a supplement data source. Chapter 4 demonstrates the procedure of processing PCLD and the mechanism of CIT and CIT-based random forests. The feature set construction is also covered within the chapter. The model results are further illustrated and compared in Chapter 5. Finally, a review of the research and future works are concluded in Chapter 6.

# Chapter 2: Literature Review

The chapter will provide a comprehensive literature review regarding various aspects of PCLD. It is composed of three parts: the introduction and evolution of PCLD, the previous efforts on extracting information from PCLD, and the prevalent methods for imputing missing information from PCLD.

## *2.1. Passively Collected Location Data*

In transportation field, PCLD is usually obtained through GPS travel survey. The first passively collected location dataset was created at the end of last century, known as "Lexington Area Travel Data". One hundred households were included after pre-solicitation efforts. The survey comprised an in-vehicle GPS device and a post-usage interview. They concluded that it was successful to install the GPS system in the household vehicle to collect raw GPS data as well as to allow manual input of travel information [4]. A handful of experiments also proved the feasibility of collecting travel data via GPS devices, either a handheld electronic travel diary (ETD) with GPS or a passive in-vehicle GPS system, to complement traditional household travel surveys [5-11]. The early studies emphasized the advantages of the GPS survey in collecting the misreported or underreported trips from traditional surveys and documenting more detailed travel activities. Meanwhile, there arose several concerns. For ETD with GPS, users may not carry the device when they consider it a burden. The passive in-vehicle GPS system is only able to capture driving trips and lacks a user interface to validate the trip information.

More researches have been done since the practicability of GPS travel survey was demonstrated. Shen and Stopher (2014) performed a comprehensive review of GPS travel survey and GPS data-processing methods. It is found that GPS travel survey has been applied for mutiple purposes [12]. Schönfelder et al. (2002) conducted a long-term GPS survey for transport safety purposes, which addressed the speeding problem specifically [13]. Bohte et al. (2007) and Bohte & Maat (2009) utilized a GPS travel survey to look into the relationship between residential self-selection and travel behavior [14, 15]. Pasquier et al. (2008) measured the effects of outdoor advertising [16]. Papinski et al. (2009) explored the travelers' decision in route choice through a person-based GPS survey [17]. Stopher et al. (2009) and Stopher et al. (2013) monitored and evaluated voluntary travel behavior change employing the GPS survey [18,19]. Other studies were dedicated to complement or even replace the traditional household travel survey [20-26].

Since mobile phone—and later smartphone—gained their popularity, investigation into the individual-level mobility pattern has become more practical. The great value of various emerging data sources has been revealed too, including call detail record (CDR), cell phone GPS data, social media location-based services, etc. Call detail record (CDR) provides details on calls and messages, such as timestamp, duration, and location(s) of routing cell tower(s) [27]. Gonzalez et al. (2008) combined two sets of CDRs to explore the individual mobility pattern. One is composed of six-month records for 100,000 randomly selected anonymous individuals and another complementary dataset captured the location of 206 mobile phone users every two hours for one week [28]. Further studies on human mobility have been conducted

based on similar datasets [29-32]. CDR is also applied to other research topics such as social network, residential location, socioeconomic level, etc. [33-35]. Despite the large volume of data, CDR is limited to its spatial resolution determined by the density of cell towers but requires less advanced phones and should cause less concern about the user privacy.

GPS-enabled mobile phone is a more convenient and less expensive replacement of the handheld ETD with GPS. The influence of the mobile phone location services on intelligent transportation system was discussed by Zhao, 2000. Then it is proved feasible to utilize a GPS-enabled cell phone to monitor locations and movements rather than a dedicated in-vehicle GPS system [37-42]. Cottrill et al. (2013) shared their experience in designing a smartphone-based mobility survey, which provided a better user interface than GPS-based travel survey [43]. Since GPS offers much more precise locations, the access to individual-level mobile GPS trajectories is highly restricted. There are several private sector companies who generated aggregated level of location data to reveal travel demand, such as INRIX, StreetLight Data, AirSage, etc. [44, 45].

Social media location data is more complicated and comprehensive since the spatial information could be implied in the posted text or the uploaded picture other than being directly recorded. At first, it mainly helps to enhance the contents of geographic and spatial data. Flanagin and Metzger (2008) included the photo-sharing site, Flickr, in their discussion about volunteered geographic information (VGI) [46]. De Choudhury et al. (2010) tried to automatically generate the travel itineraries for

popular touristic cities based on the photo streams uploaded to Flickr. They explored where and when the travelers were by mining the large amount of photos with timestamps shared by them [47]. Sui and Goodchild (2011) developed and complemented their original argument, 'GIS as media' [48], with the new opinion that '(Social) Media as GIS'. They illustrated the location-based social networking sites become more like GIS as they provided users' locations with timestamps [49]. Naaman (2011) dig into the four aspects of geographic information that can be derived from social awareness streams (SAS) data, including districts, landmarks and attractions, paths (and Itineraries), and activities. Twitter, Facebook, the photo-sharing site Flickr, and the Foursquare location and presence-sharing service are all counted as SAS platforms [50]. Zhong et al. (2015) combined the location check-ins from Sina Weibo (China's Twitter) and the Points of Interest (POIs) of Sina Weibo and Dianping (a review website similar to Yelp) to realize user profiling [52]. Riederer et al. (2015) collected two-year public photo metadata from Instagram. They revealed the potential of social media location data in two ways: first, they demonstrated that the human mobility patterns drawn from photo-sharing networks are comparable with those from CDRs; after that, they proved that an individual's ethnicity could be predicted solely based on the location data [53]. In addition, there are more studies utilizing such data to inspire new location-based services [54], predict the next location to visit [55], link users across domains [56], identify user's home location [57], propose possible activity companion [58], etc.

PCLD is widely applied beyond the transportation field. Troped et al. (2008) employed the GPS and accelerometer data to predict the physical activity mode, such

as walking, running, biking, or driving an automobile [59]. Gilbert and Karahalios (2009) utilized social media data to measure and predict the tie strength between social media friends [60]. Soto et al. (2011) tried to predict the socioeconomic levels of a population based on the aggregated CDRs. De Montjoye et al. [61]. (2013) tracked a long term of human mobility traces and concluded that they are highly unique, which draws discussion on the privacy protection of individuals [62]. Zhang et al. (2015) investigated the characteristics of mobile network behavior based on two types of telecommunication data, user-oriented and network-oriented [63].

### *2.2. Trip Information Imputation*

Along with the development of PCLD, a lot of attempts have been made to derive the travel information from the raw data. Gong et al. (2014) conducted a literature review on the methodologies of deriving personal trips from GPS data [64]. Four processing procedures are discussed including data error recognition, trip identification, travel mode detection, and trip purpose inference. The potential of utilizing GPS trace data for travel behavior analysis was evaluated by Schönfelder et al. (2002) [13]. They tried to post-process the data to identify the drivers, trip ends, stops, trip purposes, and the potential to construct all-mode activity patterns using driving GPS records. Chung and Shalaby (2005) developed a map-matching algorithm to identify the roadway links traveled with the GPS data collected by GPS tracers and a GIS database. Built upon that, a rule-based model is constructed to detect the travel mode configuration including predefined multimodal patterns [65]. An enhanced framework was later proposed [66] that first applied a rule-based model to segment trips by mode transfer point (MTP) and then used a fuzzy logic-based algorithm to

identify mode within trip segments. The following research on trip identification and mode detection by Schuessler and Axhausen (2009) employed a fuzzy logic approach to detect the mode followed by a reasonability check. They also highlighted the model's capability of dealing with a large sample and getting rid of manual intervention [67]. Gonzalez et al. (2010) developed a smartphone app TRAC-IT, in which a neural network algorithm for mode detection was embedded. The so-called multi-layer perceptron took speed, acceleration, estimated horizontal accuracy, and more as input variables [68]. Zhang et al. (2011) proposed a multi-stage algorithm: the three mode classes (walk, bike, motorized vehicles) are identified in the first stage by speed, acceleration, etc.; and in the second stage, the detailed modes under motorized vehicles are identified using Support Vector Machines (SVMs) method [69]. Gong et al. (2012) constructed a GIS algorithm to impute the travel mode from the enormous amount of GPS data in New York City, a complex urban environment, where the urban canyon effects and the multimodal transportation network need more attention [70]. Nitsche et al. (2014) mainly utilized the acceleration data collected by smartphone to automatically reconstruct the trips. They also employed a Discrete Hidden Markov Model (DHMM) to compute the travel modes [71].

In addition to travel mode, Wolf et al. (2004) conducted a proof-of-concept study. They integrated GPS trace details, associated survey details, and external POI data for the activity purpose imputation method and opened the discussion about trip-end identification in a GPS processing system [72]. Stopher et al. (2008) constructed a rule-based method to identify trip purpose based on the parcel-level GIS data and extra information collected, such as home/work/school location, and the two most

frequently used grocery stores or supermarkets [73]. Bohte and Matt (2009) inferred trip purpose using GIS information other than home and work locations. The travel mode was determined considering both speed and transit route [15]. Elango and Guensler (2010) tried to identify trip purposes (home, work, maintenance, discretionary, and multipurpose) based on the home/work locations and the closet POI to trip ends [74]. Huang et al. (2010) developed an algorithm for activity identification incorporating spatial temporal POIs' attractiveness (STPA). STPA not only addressed the static attractiveness of business but also added a dynamic factor to demonstrate the variation due to the time of day, in which the business's related activity usually happens [75]. Liu et al. (2013) imputed activity purposes based on mobile phone call locations and a set of machine learning algorithms [76]. Shen and Stopher (2013) further included tour type identification in their study [77]. Kim et al. (2014) developed a learning model to impute the activity associated with the given stop using data collected by a smartphone-based travel survey [78]. Oliveira et al. (2014) compared nested multinomial logit and decision tree model in terms of performance. They first categorized household members into eight person types and then incorporated GPS travel data to impute trip purposes [79]. Ermagun et al. (2017) utilized Google Place to realize real-time trip purpose prediction. They also found that random forest outperforms nested logit models [80]. More studies discussed the performance of machine learning methods in trip purpose imputation [81-84].

Regarding the relatively new topic, imputing social demographic information based on PCLD, the literature is more limited. Lu and Pas (1999) demonstrated the relationship between social demographics, activity participation, and travel behavior

through a structure equation model. Although the paper focused on the direct and indirect effects of social demographics on travel behaviors, such as the number of trips per day, it inspired the possibility of studying the problem in a reversed way, which is to infer travelers' social demographics based on their travel behaviors [85]. Altshuler et al. (2012) first tried to include some indirect location features (numbers of different cell tower IDs and different Wi-Fi network names) to impute individual attributes like ethnicity, whether a student, and whether a US-native [86].

Auld et al. (2015) defined the problem in a specific scenario whether demographic characteristics of travelers could be derived from travel behaviors. Their method can be divided into two parts: person type clustering based on the similarity of their travel patterns and demographics modeling under each person type, including education, age, gender, license, and household type (defined by household size, number of vehicles, and presence of child). They used various models and algorithms to impute different attributes: partial decision tree classification algorithm (PART) for person type and license possession, nested logit for education, ordinal logit for age categories, binary logit for gender, and C4.5 for household type. They achieved similar prediction accuracy between the training data and test data from two surveys but their model is restricted to several assumptions, such as the GPS trace data needs to cover at least one full day of travel and the home/work/school locations need to be available [87]. Such assumptions may not be fulfilled in some prevailing data sources. For example, the location data gathered by mobile phones or CDR is of lower quality than that via GPS devices. Thus, processed data can miss some trips.

Zhong et al. (2015) did a similar job for a larger amount of users and their location check-ins through social network. The main demographic characteristics considered are gender, age, and education background. Since the location check-ins are not continuous, the feature set was composed of POI and temporal information. They also compared several methods for each response variable type, including logistic regression, SVM, neuron networks, etc. [52]. Riederer et al. (2015) aimed to infer the demographics (ethnicity and gender) from people's location data collected by Instagram. They utilized a simple Bayesian inference method and compared the model performances with or without auxiliary data (Census data and surrounding venue data from Foursquare) [53]. Roy and Pebesma (2017) inferred gender first and then age groups under each gender type based on anonymized mobile phone GPS trajectories. For gender imputation, they chose a supervised learning approach of Linear Discriminant Analysis (LDA) and for age groups, a decision-tree based classification approach. They constructed the feature set with trip-based information and POI data as well as the frequently visited places they discovered [88].

### *2.3. Imputation Methods*

It can be concluded that the missing information of PCLD are usually treated as nominal variables. Therefore, the imputation of such responses is modeled as the nonlinear classification problem. Feng and Timmermans (2016) summarized and compared the algorithms applied to mode detection, including naive Bayesian, Bayesian network, logistic regression, multilayer perceptron, support vector machine, decision table, and C4.5. Those methods are also dominant in the imputation of trip or activity purposes [89].

The naive Bayesian method is mainly based on the Bayes' rule, where all the predictors are assumed to be independent from each other. The Bayesian network relaxes the assumption and considers the joint probability of an attribute with its parent attributes. But the joint probability distribution could hardly be employed when the dimensions of predictors and its possible values exceed two. To simplify the risk model, the conditionally independent assumption is often made in real-world applications [90].

Logistic regression in general is a regression model with the dependent variable to be categorical. It has been extensively used to model the discrete choice problem in transportation sector. The family includes several common model specifications, such as binary logit, multinomial logit and ordinal logit regarding the type of response variable, or nested logit in order to capture additional relationships between predictors [91]. To allow the variation of coefficients among decision makers, mixed logit with random coefficients was introduced [92]. However, it is sometimes hard for logistic regression to capture the nonlinear and complicated influence of independent variables in the real world. Also, it is hard to accommodate data with small sample size but large feature set.

One intuitive way to handle the nonlinear problems is to employ machine learning methods, which has been evolved for almost 60 years. Some typical examples of machine learning algorithms are artificial neural networks (ANNs) [93], decision trees [94], and support vector machine (SVM) [95]. They were developed to address

the classification problem and become superior in their own way. For example, ANNs frequently outperform on huge and complex problems [96].

The Conditional Inference Tree (CIT) is selected to address the demographics imputation problem due to its eligibility to capture the influences of predictors from a small training dataset and to be interpreted in an intuitive way. It was first proposed by Hothorn et al. in 2006 [97] and later extended to the ensemble method as random forests [98]. The highlight of the CIT is its ability to handle the variable selection bias and overfitting problem that usually exist in decision trees. The method will be described thoroughly in Section 4.2.

# Chapter 3: Data

The chapter introduces the two PCLD datasets used for the study. The social demographic information of the data providers is also collected, including gender, age group, education level, and household income level.

## *3.1. 2011-2012 In-vehicle GPS Travel Survey*

The in-vehicle GPS travel survey was conducted between October 2011 and February 2012 in Maryland. A dedicated in-vehicle GPS device was installed and kept recording the vehicle's location every one minute if any movement was detected. To obtain the ground truth for the survey subjects and the trips captured, both an initial participation form and a recall survey were designed to collect the social demographic information and one-day travel diary. More details about the survey design could be found in [99].

From the 230 initial survey participants, 163 subjects (69 females and 93 males) are selected who have provided at least one of the four social demographic attributes and continuous GPS traces of more than 30 days. Overall, there are more senior people recruited in the survey as shown in Figure 3-1. Although the sample is selected via a stratified random sampling method, there remains a bias towards people with high education level and household income level (Figure 3-2, 3-3).

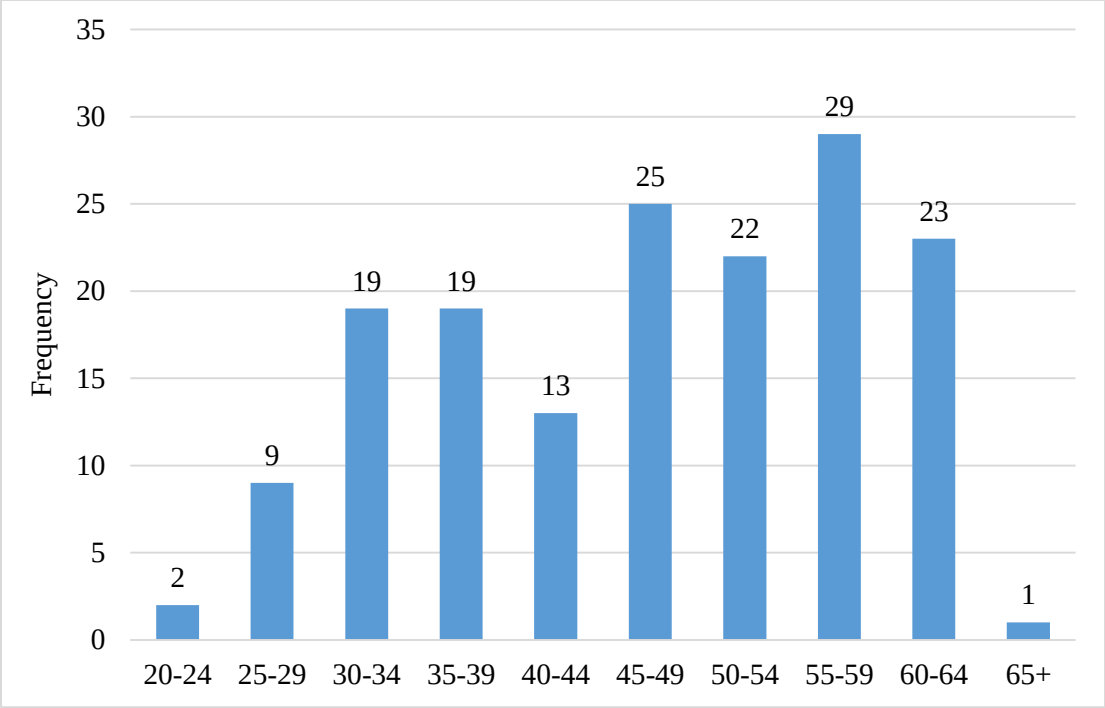


Figure 3-1. In-vehicle GPS Survey: Age Group Distribution

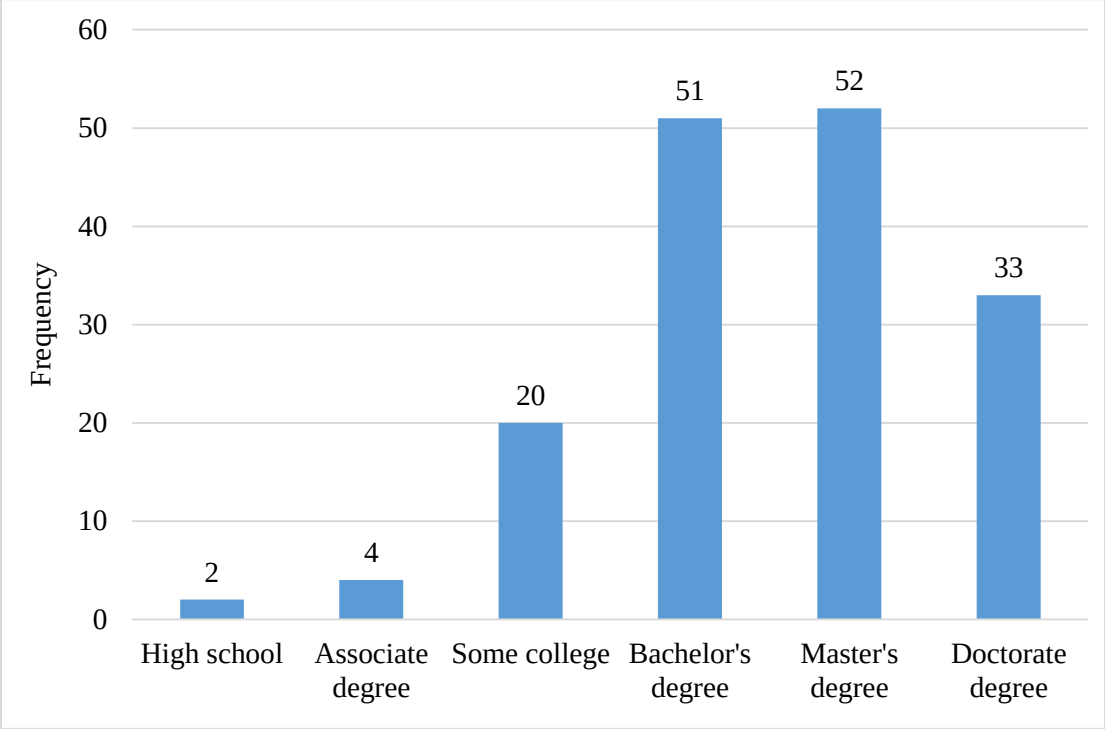


Figure 3-2. In-vehicle GPS Survey: Education Level Distribution

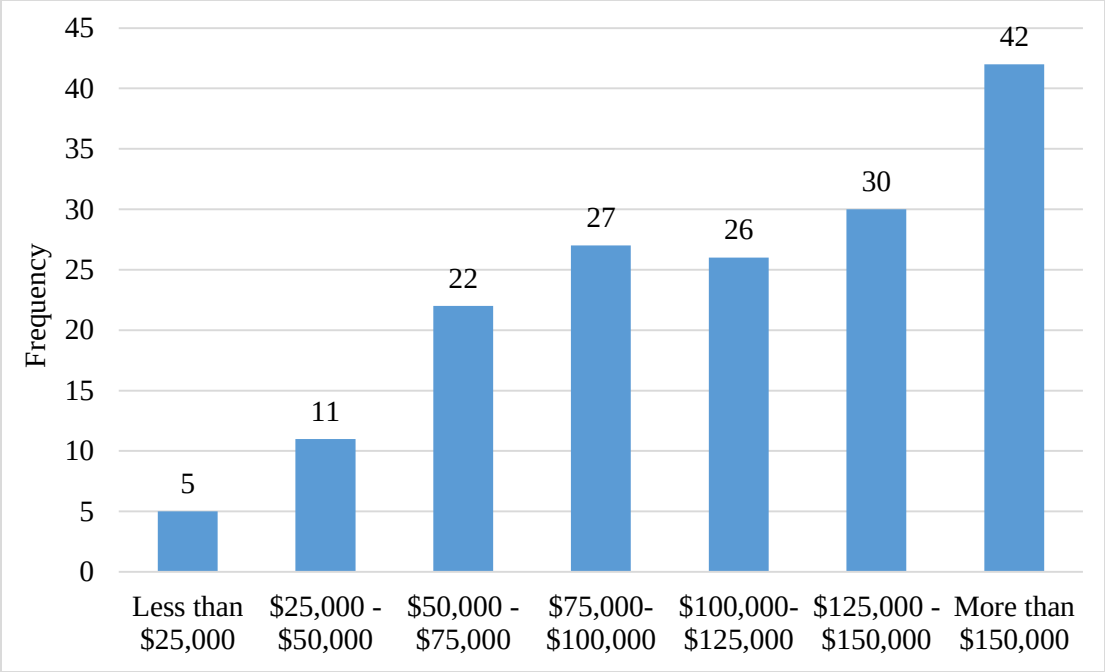


Figure 3-3. In-vehicle GPS Survey: Household Income Level Distribution

Through a previous study [100], the raw data have been processed into trips based on the trip end identification rules, including spatial movements less than 200m, temporal duration greater than 5min, and any point speed recorded less than 5m/s. In addition, the trip purpose was imputed using random forests and HERE POI dataset. The overall prediction accuracy is above 80%. The imputation for home and work trips are more accurate with all the home trips and nearly 90% of work trips correctly predicted. The imputed mode will also be considered in the demographics prediction other than the travel behavior characteristics, land use, and POI information.

### *3.2. 2017 SafeTrack Smartphone Travel Survey*

SafeTrack is an accelerated metro work plan which performs maintenance during a relatively short period in order to improve safety and reliability of the Washington Metrorail system. Since June 2016, 16 surges have been finished in different lines.

The main impact of one surge is either continuous single track or line segment shutdown. To study the influence of reduced service on metro riders' travel behaviors, the National Transportation Center (NTC) at University of Maryland conducted a long-term survey in collaboration with George Mason University.

For the first 11 surges, the travel behavior data were collected through paper-based and web-based questionnaires. Later, the survey team employed a smartphone application (shortened as app), TravelHelper, developed by NTC research team. The installation invitation was sent to previous survey respondents. More flyers were also distributed in the affected metro stations during the following surges until the end of SafeTrack. After that, an online recall survey was emailed to the app users. The questions are separately designed for the users enrolled from previous surges and from the flyer distribution. In general, the social demographic information for both groups of user are gathered.

#### 3.2.1. Questionnaire-based Survey

Eventually, there are 128 app users (71 females and 57 males) who have provided at least one of the four social demographic attributes and stayed active for more than 3 days. Since the app users were originally enrolled at the metro stations, there is no control of social demographics over the selected sample. Moreover, the sample units tend to have higher education level and income level (Figure 3-5, 3-6).

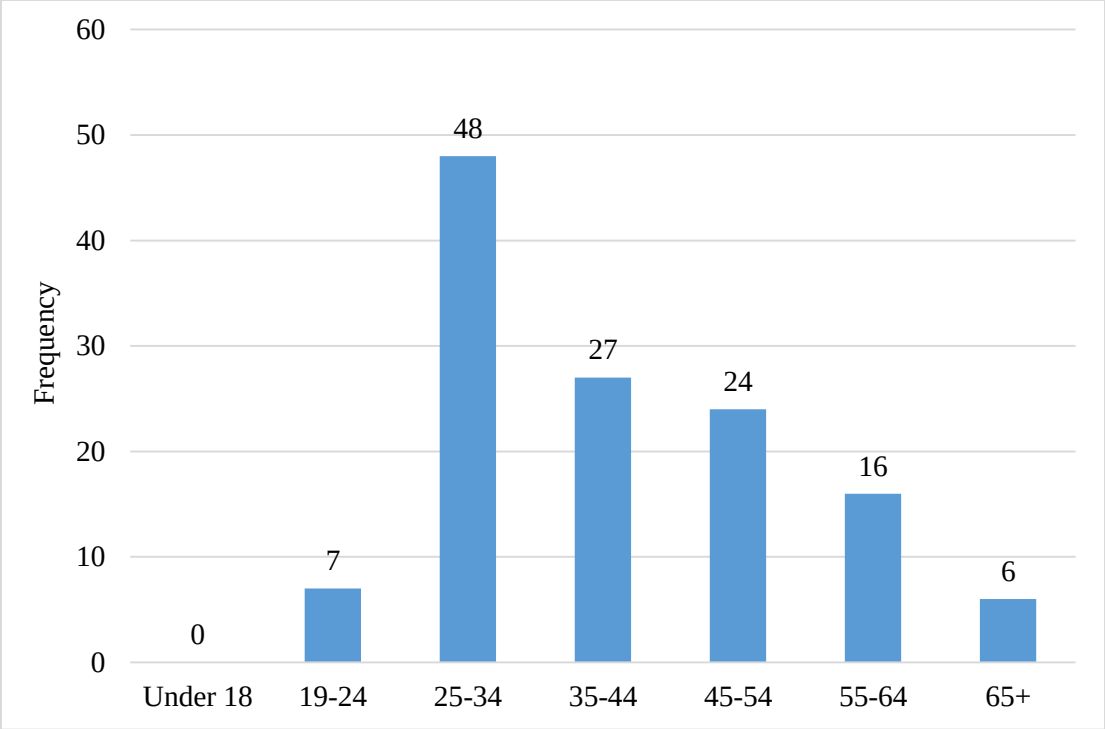


Figure 3-4. Smartphone-based Survey: Age Group Distribution

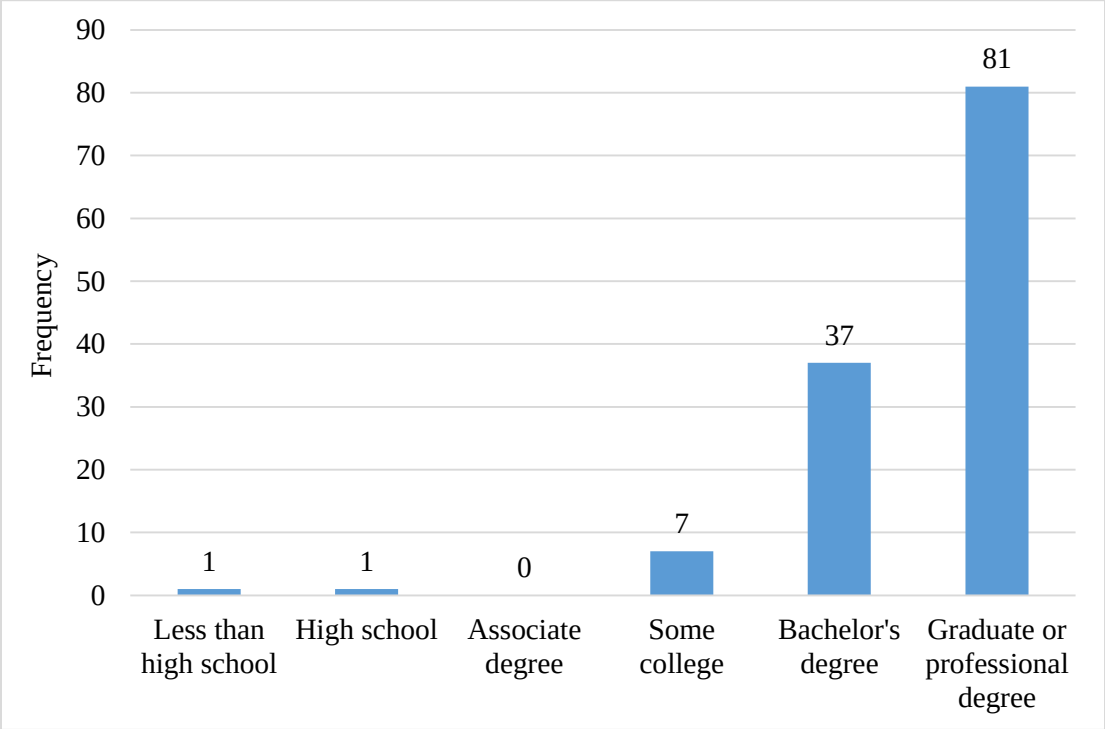


Figure 3-5. Smartphone-based Survey: Education Level Distribution

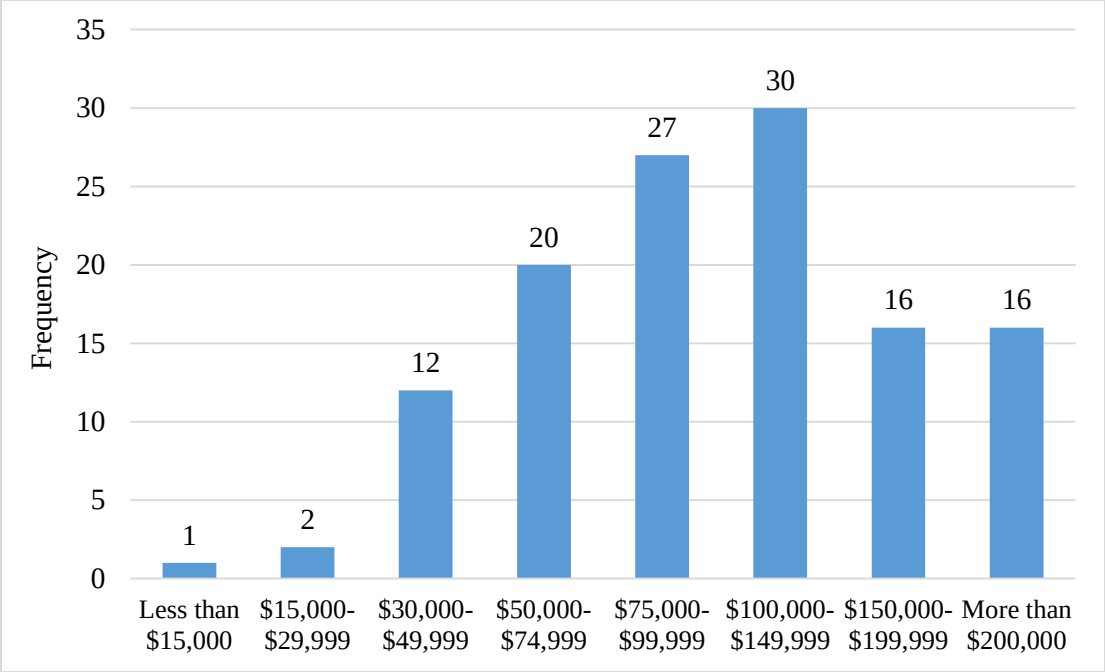


Figure 3-6. Smartphone-based Survey: Household Income Level Distribution

### 3.2.2. App-based Survey

Each app user was assigned a unique identification character string (mobile ID) once registered. There are two versions of TravelHelper developed for iOS and Android with similar framework. The app users could voluntarily record their trips specifying the travel mode and trip purpose. When they clicked on the start or end button, the app automatically recorded the timestamps and locations. Since the app users were not required to report the actual address, the trip ends could be missing if the app failed to locate the smartphone via any of the three positioning methods (GPS, Wi-Fi, and cell towers).

The app is also enabled to track the users in the background as long as it is not completely shut down. But due to the characteristics of the two operating systems, the reporting frequency and location accuracy are varied (Table 3-1). For the Android

version, the functions to monitor location are defined by NTC team so the parameters are clearly stated. For the iOS version, the functions are predefined by Apple Developer and the parameters within are unknown. The frequency and accuracy are summarized based on the observed records. Because of the multisource positioning scheme, the speed information is missing when the record is not generated through smartphone-embedded GPS. More details will be introduced in Chapter 4.

Table 3-1. Location Data Quality for iOS and Android

| Operating System | Frequency | | | Location Accuracy |
|---|---|---|---|---|
| | Scenario | Regular interval | Fastest interval | |
| Android | High battery, moving | 30s | 10s | High |
| | High battery, static | 10min | 5min | Balanced |
| | Low battery, moving | 5min | 1min | Balanced |
| | High battery, static | 30min | 10min | Balanced |
| iOS | Static | Until significant location change is detected | | 2m - more than 100km |
| | Moving | 1-1137s | | |

### *3.3. Smart Location Database*

The Smart Location Database (SLD) is a public domain data product provided by the U.S. EPA Smart Growth Program. The SLD summarizes demographic, employment, and built environment variables for every Census block group (CBG) defined by 2010 Census for the entire U.S. Several attributes are considered for the study. The SLD integrates different aspects of area characteristics at a fine geographic resolution. Nevertheless, there exists inconsistency between the population and the number of workers that some block groups have more workers (based on home location) than population. Thus, the data in such block groups are considered missing. Most

attributes from the SLD are straightforward but the one to depict land use diversity is slightly complicated and calculated as follows:

$$\text{D2a\_EpHHm} = -\frac{\sum_{i=1}^{n}(P_i \times \ln P_i)}{\ln n \times Ac\_Unpr}$$

where $P_i$ is the proportion of the employees or the housing units in land use type $i$ found in a block group and $n$ is the number of types ($n = 4$), which is commercial/industrial/institutional, retail, recreational, and residential [101].

Table 3-2. Selected Attributes from the SLD

| Attribute | Description | Data Source |
|---|---|---|
| Ac_Unpr | Total land area in acres that is not protected from development (i.e., not a park or conservation area) | Census, Navteq parks, PAD-US |
| D1a | Gross residential density (HU/acre) on unprotected land | SLD |
| D1b | Gross population density (people/acre) on unprotected land | " |
| D1c | Gross employment density (jobs/acre) on unprotected land | " |
| D2a_EpHHm | Employment and household entropy | " |
| D3a | Total road network density | NAVSTREETS |
| D4b050 | Proportion of CBG employment within ½ mile of fixed-guideway transit stop | TOD Database 2012, SLD |
| TotPop | Population, 2010 | 2010 decennial Census |
| P_WrkAge | Percent of population that is working aged, 2010 | 2010 decennial Census |
| Workers | Number of workers in CBG (home location), 2010 | Census LEHD, 2010 |
| R_LowWageWk | Number of workers earning $1250/week or less (home location), 2010 | " |
| R_MedWageWk | Number of workers more than $1250/week but less than $3333/week (home location), 2010 | " |
| R_HiWageWk | Number of workers earning $3333/week or less (home location), 2010 | " |
| TotEmp | Total employment, 2010 | " |
| E_LowWageWk | Number of workers earning $1250/week or less (work location), 2010 | " |
| E_MedWageWk | Number of workers more than $1250/week but less than $3333/week (work location), 2010 | " |
| E_HiWageWk | Number of workers earning $3333/week or less (work location), 2010 | " |

# Chapter 4: Methodology

To fulfill the objectives of the study, the methodology for processing raw data of PCLD is developed in Section 4.1. The mechanism of CIT and CIT-based random forests is then introduced in Section 4.2. Section 4.3 describes the feature set construction for training the machine learning methods. Four sets of features are included and examined.

## *4.1. PCLD Processing*

The PCLD collected from the in-vehicle GPS devices and the smartphone app have the same structure of raw data, which is composed of device/account ID, latitude, longitude, speed, accuracy, and timestamps. But the reporting frequency and location accuracy differentiate between the data sources and thus the processing procedures are also different.

### 4.1.1. In-vehicle GPS Data

The raw data generated by in-vehicle GPS device have relatively high frequency and accuracy so the processing method is developed by detecting the stops through continuous location points. As demonstrated in Figure 4-1, the trip end or activity location is identified as a set of successive points barely move. The criteria include that any distance between the first point and the rest points is less than 200m, the duration from the first point to the last one is no shorter than 5min, and any point speed detected in the set is no greater than 5m/s. If a set of successive points met all the criteria, then the set is considered as a trip end and the centroid of it is calculated

as the actual activity location. In addition, round trips without a middle stop longer than 5min are deleted since the trip purpose will be hard to infer. More details could be found in [100].

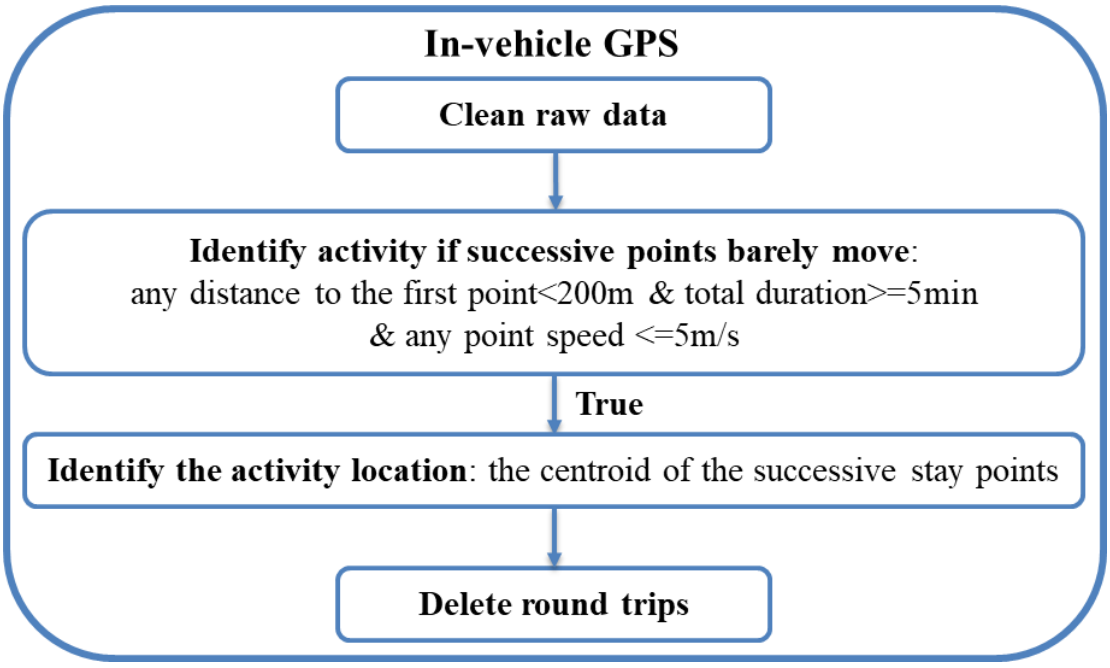


Figure 4-1. Processing Procedure for In-vehicle GPS Raw Data

#### 4.1.2. Smartphone Location Data from iOS Version

The tracking strategy and data quality of the iOS version app is extremely different from that of in-vehicle GPS device. Instead of recording the locations with fixed intervals, the iOS app only starts to track the users when a significant location change is detected. Meanwhile, the app is enabled to position the smartphone through multiple ways, such as Wi-Fi and cell towers, other than GPS. As a result, the app could capture several locations within a short period (e.g. 1min) and try to record more than one coordinates even within 1sec if the accuracy is low. The data acquired will be uploaded to the cloud server for storage when the smartphone has access to

the Internet. The storage method sometimes causes wrong temporal order of raw data. In such cases, the first step is to sort the location records by timestamp and to keep only one record with the highest accuracy within one minute to avoid ambiguity.

Another issue is that once the app starts tracking the location, it occasionally needs to cool down before terminating the function. Accordingly, the location records should be deleted after the user stopped the trip in reality. Such points are considered to be recorded after the actual trip ends: 1) the moving distance is less than 200m; 2) the average speed between points is less than 1m/s; 3) the duration between points is less than 20min. The average speed is used instead of point speed because the speed of most records is unavailable.

The following step is to identify the trip ends by recording intervals. Based on the characteristics of the raw data, 20min is selected as the threshold to prevent separating a single trip. Consequently, many activities of short duration will be ignored and later treated by post-checking (Figure 4-2).

The post-checking is designed to clean the misreported trips due to low accuracy, adjust the coordinates of trip ends, and break the round trips. It should be mentioned that the round trip imputed from the iOS app is different from that from in-vehicle GPS data. For example, the round trip inferred from the iOS app may last for nearly one hour with only four points captured. During the round trip, the user may have finished a grocery shopping, which should be taken into account.

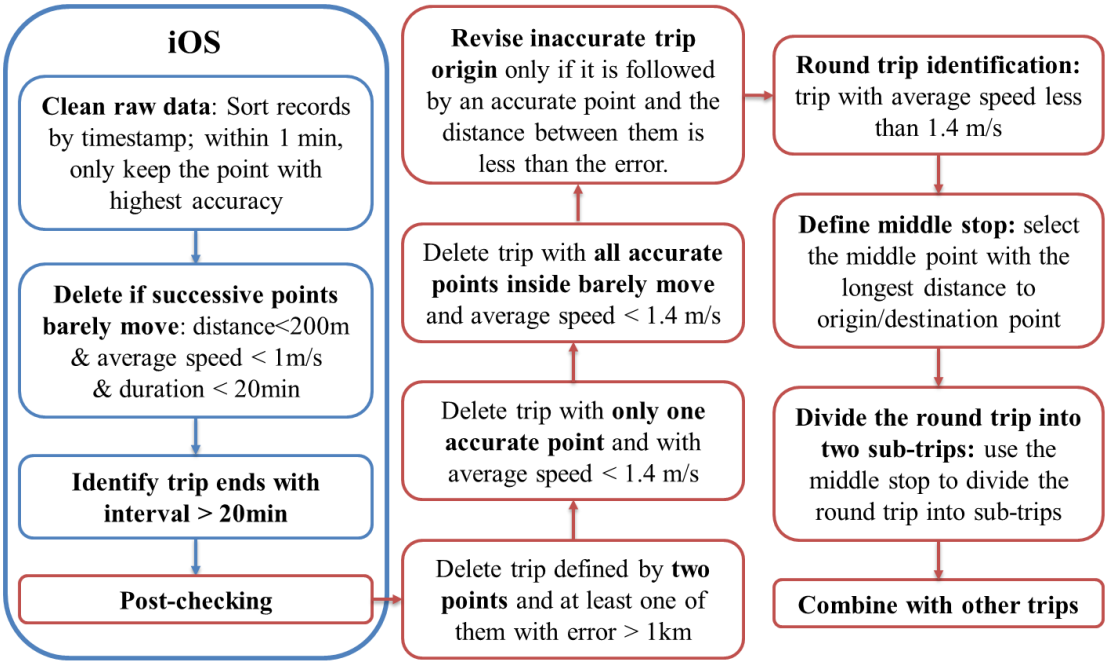


Figure 4-2. Processing Procedure for Smartphone Location Data of iOS Version

After the primary processing, some trips may be inferred incorrectly since the app mistakenly locates the smartphone. A trip is highly possible to be inaccurate if it is defined by only two raw data points and at least one of them is reported with accuracy worse than 1km. Likewise, the trips will be deleted with only one accurate point and the average speed is less than walking speed (1.4m/s). It is also observed that multiple accurate points exist within one trip but they are very close to each other. For instance, the user's home location and another location more than 1km away from home were recorded alternately every hour from late night till early morning. The home location is reported with higher accuracy (less than 50m) while the other point with very low accuracy (more than 1km). To address the issue, the trips are deleted if the average speed is less than 1.4m/s and all the accurate points inside barely move.

In order to locate the trip start more precisely, the start point of a trip is adjusted if: 1) the accuracy of the start point is worse than 200m; 2) the following (second) point is accurate (less than 200m); 3) the distance between the two points is no longer than the error of the first inaccurate point.

Then the round trip is identified as the average speed less than 1.4m/s. As mentioned before, the trip start is usually not recorded immediately so the distance between trip ends is not set as the criterion for round trip recognition. The middle stop is determined to be the point which is farthest from either trip end. The round trip is later divided by the middle stop into two sub-trips.

#### 4.1.3. Smartphone Location Data from Android Version

The location data collected by the Android version app is very similar to that by in-vehicle GPS device except for the lower frequency and accuracy. So the processing procedure is almost the same as described in Section 4.1.1 except for two steps. Due to the relatively low frequency and the further integration with iOS dataset, the trip ends are identified with total duration equal to or greater than 20min. Additionally, some short trips, usually as walking trips, are deleted if the distance between trip ends is less than 200m and the travel time is less than 5min (Figure 4-3).

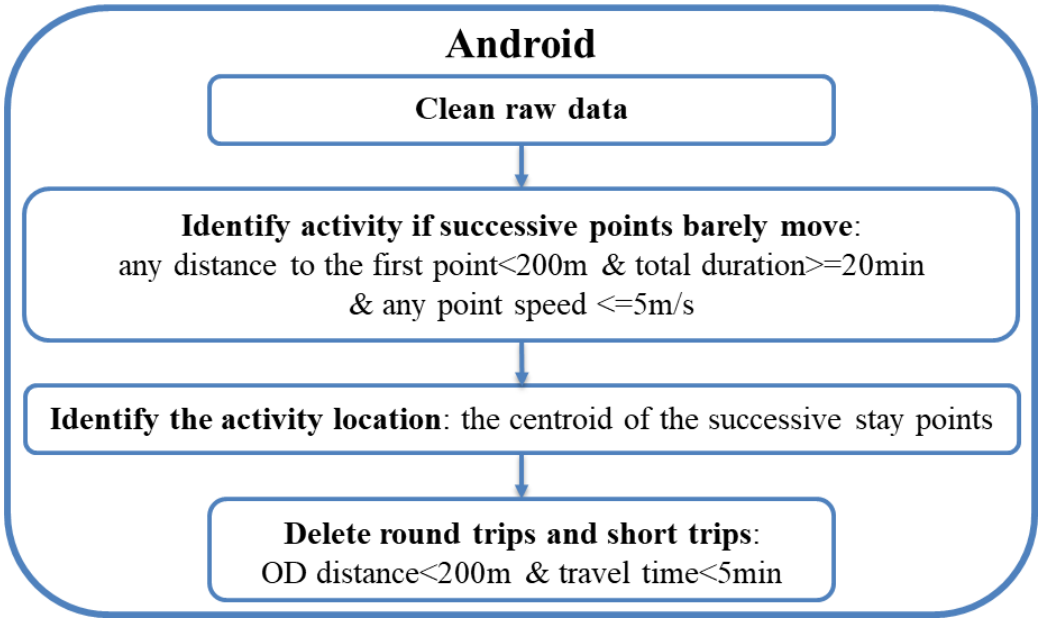


Figure 4-3. Processing Procedure for Smartphone Location Data of Android Version

#### 4.1.4. Summary

Based on the pervious processing strategies, the number of trips identified for each user is summarized in Figure 4-4. It can be observed that the number of imputed trips for each GPS device user tends to follow the normal distribution. On the other hand, the iOS users are skewed to having fewer imputed trips and the Android users have highly varied number of imputed trips.

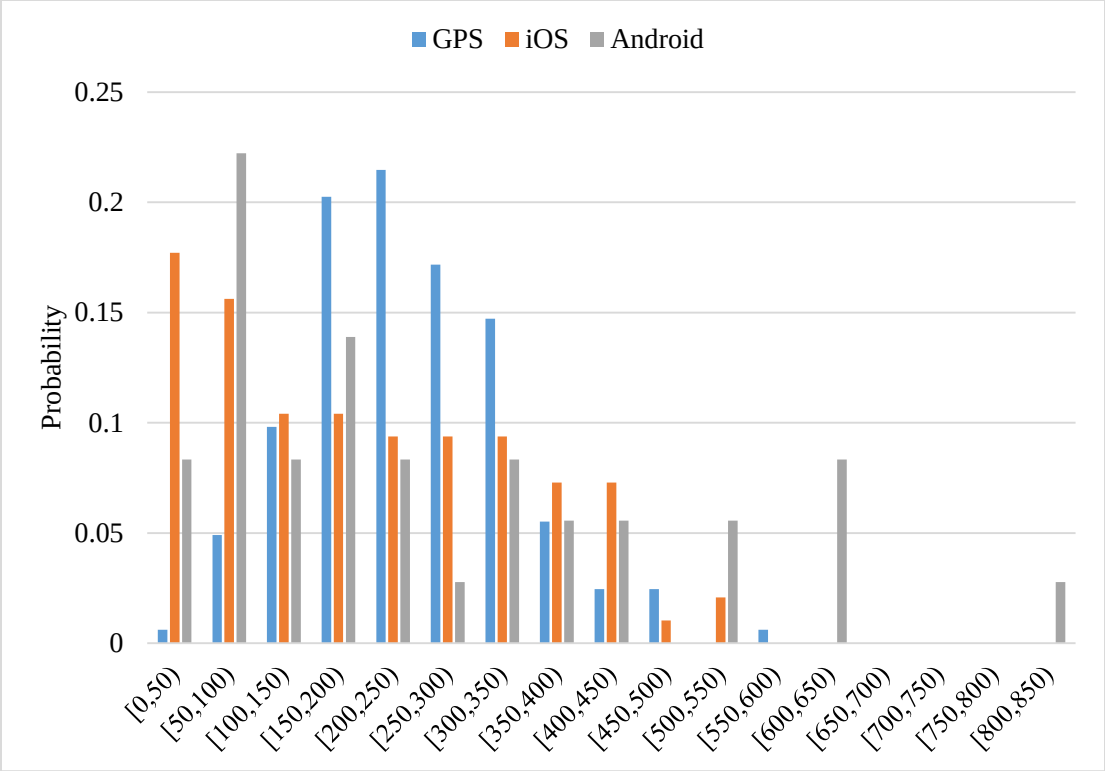


Figure 4-4. Processing Procedure for Smartphone Location Data of Android Version

## *4.2. Conditional Inference Trees*

### 4.2.1. Overview

The Conditional Inference Tree (CIT) is a recursive partitioning framework with tree-structured regression models and conditional inference procedures embedded. As shown in Figure 4-5, the main contributions of CITs are: 1) avoiding the overfitting problem by checking the global null hypothesis of independence between all the covariates and the response(s); 2) avoiding the variable selection bias by first selecting the covariate with the strongest correlation to the response(s) in each iteration. The unified framework is demonstrated in [98] and the following sections will introduce the specification applied to the imputation problem.

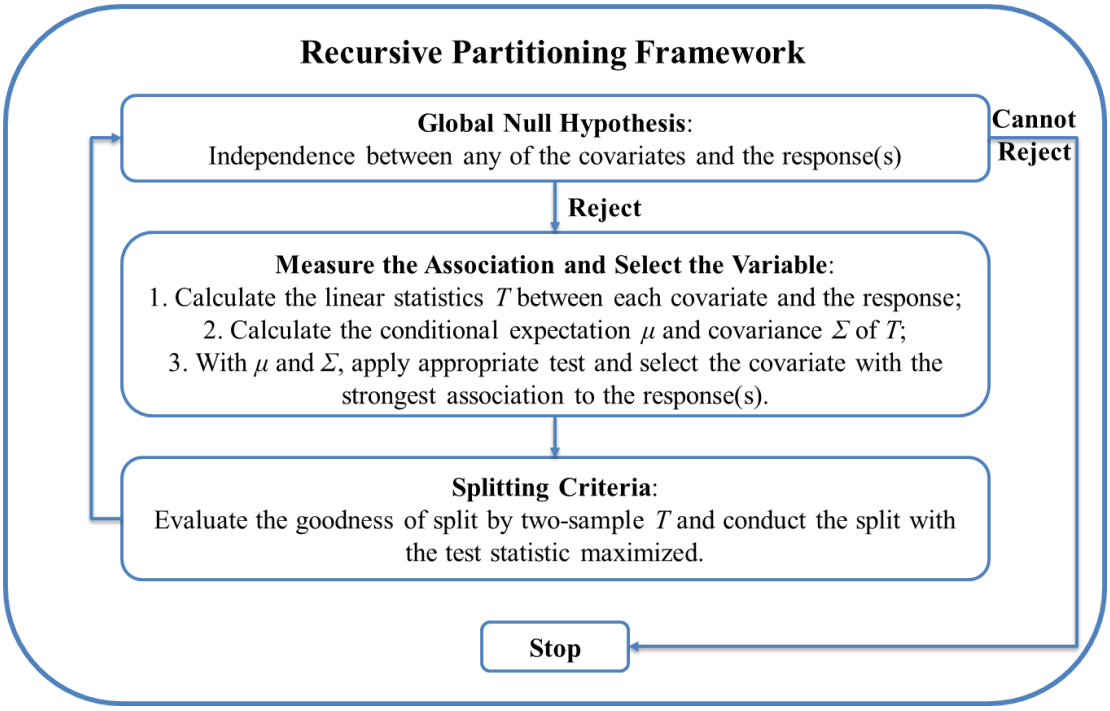


Figure 4-5. The Framework of Conditional Inference Tree

4.2.2. Variable Selection and Stopping Criterion

The specification of the model is given below with the univariate formulation, which is applied to this study. First, the learning sample is defined as:

$$\mathcal{L}_n = \{(Y_i, X_{1i}, \ldots, X_{mi}), i = 1, \ldots, n\} \quad (1)$$

where $n$ is the sample size, $m$ is the number of covariates, $Y_i$ is the response in the $i$th observation, and $X_{ji}$ is the $j$th covariate in the $i$th observation, which can be missing. The nonnegative integer valued case weights $w = (\omega_1, \ldots, \omega_n)$ is assigned to the learning sample.

The global null hypothesis is composed of the $m$ partial hypotheses: $H_0 = \bigcap_{j=1}^{m} H_0^j$ with $H_0^j: D(Y|X_j) = D(Y)$ $(j = 1, \ldots, m)$, where $D(Y|X_j)$ is the conditional

distribution of the response $Y$ given the covariate $X_j$. The case weights $\omega_i$ are either zero or one for simplification. The association between $Y$ and $X_j$ is then measured by:

$$T_j(\mathcal{L}_n, w) = \sum_{i=1}^{n} \omega_i g_j(X_{ji}) h\big(Y_i, (Y_1, \dots, Y_n)\big) \qquad (2)$$

where $g_j$ is a nonrandom transformation of the covariate $X_j$ and $h$ is the influence function. For the nominal response with $J$ levels, the influence functions can be defined as $h\big(Y_i, (Y_1, \dots, Y_n)\big) = e_J(Y_i)$. If the response is ordinal, the influence function is the same and a score vector can be added to the linear statistics. For numeric covariates, the transformation can be defined as $g_{ji}(x) = x$.

The formulations for the conditional expectation $\mu_j$ and covariance $\Sigma_j$ of $T_j(\mathcal{L}_n, w)$ were originally derived by Strasser and Weber (1999) and can be found in Hothorn et al. (2006). Once the conditional expectation and covariance are ready, the linear statistic can be standardized and the test statistics $c$ is used to examine if the significance level of the association is below or at level α (typically set to 5%). Since all the responses studied are nominal variables, the quadratic form of the test statistic is applied for efficiency: $c_{quad}(\boldsymbol{t}, \mu, \Sigma) = (\boldsymbol{t} - \mu)\Sigma^{+}(\boldsymbol{t} - \mu)^{T}$, where $\Sigma^{+}$ is the Moore-Penrose inverse of Σ.

#### 4.2.3. Splitting Criteria

After the covariate $X_{j^*}$ is selected to perform the binary split, the goodness of a split is evaluated by the special case of the linear statistics which only has two samples. For

all possible subsets $A$, the two-sample statistic is formulated as (3) and the split $A^*$ is selected with the test statistics maximized (4):

$$T_{j^*}^{A}(\mathcal{L}_n, w) = \sum_{i=1}^{n} \omega_i I\big(X_{j^*i} \in A\big) h\big(Y_i, (Y_1, \ldots, Y_n)\big) \tag{3}$$

$$A^* = \underset{A}{\operatorname{argmax}}\, c(\boldsymbol{t}_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A}) \tag{4}$$

#### 4.2.4. Random Forests

The concept of random forests is first introduced by Breiman in 2001 [102] based on his earlier work about bagging predictors [103]. The basic idea is to construct a forest with multiple tree predictors and further generate an aggregated predictor. It has been extensively studied and examined in practice. Following the unified framework of the CIT, random forests composed of CITs are also examined [104]. The method produces an aggregated predictor in the following way: for each bootstrap sample drawn from the original sample, an CIT is constructed for the classification task. Instead of utilizing the whole variable set, only a small subset is randomly selected for each CIT. Eventually, the response is predicted as an average or majority vote from all the trees within the forest.

Random forests cannot be interpreted and visualized in an intuitive way as the CIT. The importance of each variable needs to be measured based on all the trees within the forest. The naïve method is to count the number of trees where one variable has been used. Another method is to evaluate the performance improvement made by each variable, such as Gini importance. A more advanced method is called permutation accuracy importance [102], which measures the variable importance (VI)

by breaking its original association with the response(s) and evaluating how much the prediction accuracy has decrease. The basic formulation to calculate the VI of $X_j$ in tree $t$ is demonstrated as:

$$VI^{(t)}(X_j) = \frac{\sum_{i \in B^{(t)}} I(\gamma_i = \hat{\gamma}_i^{(t)})}{|B^{(t)}|} - \frac{\sum_{i \in B^{(t)}} I(\gamma_i = \hat{\gamma}_{i,p}^{(t)})}{|B^{(t)}|} \quad (5)$$

where $B^{(t)}$ is the out-of-bag sample for tree $t$, $\hat{\gamma}_i^{(t)}$ is the predicted class for observation $i$ before permutation and $\hat{\gamma}_{i,p}^{(t)}$ is the predicted class after permutation of $X_j$. Then the overall VI is measured as:

$$VI(X_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(X_j)}{ntree} \quad (6)$$

A conditional permutation scheme was also introduced, which permutes the values of $X_j$ conditionally on the variables $Z$ who have empirical correlation with $X_j$ [105]. The term $\hat{\gamma}_{i,p}^{(t)}$ in the basic formulation will be changed to $\hat{\gamma}_{i,p|Z}^{(t)}$.

To further develop a robust way considering class imbalance, an AUC-based permutation variable importance measure was proposed [106]. The measure is based on the area under the curve (AUC) instead of prediction accuracy. For a binary response variable $Y$, AUC is the probability that a randomly selected observation from class $Y = 1$ receives higher scores for class $Y = 1$ than a randomly selected observation from class $Y = 0$. Since AUC is measured for binary response variable and there are missing values in the feature set, the basic formulation of the permutation accuracy importance is employed in the following contents.

#### 4.2.5. Implementation

The CIT has been implemented in the R package “party” and “partykit”. The command “ctree” is used to derive the classifier of a single CIT and “cforest” for the classifier of random forests made of CITs. For additional control over the CIT and CIT-based random forests, the function “ctree_control” and “cforest_unbiased” are employed. To compute the variable importance, the function “varimp” is also utilized.

### *4.3. Feature Set Construction*

As the general idea of the study is to impute social demographic information from PCLD, the feature set first contains attributes indicating the travelers’ travel behaviors. Some common examples are daily trip rate, departure time, travel distance, etc. On top of that, the home location and work location can be inferred from PCLD so the area characteristics of the two places are also considered.

#### 4.3.1. Travel Behavior Statistics

Table 4-1 gives a summary of all the attributes regarding travel behaviors. As the app may be shut down sometime during the survey, the definition of active day is employed. An active day is counted only if there is at least one trip captured. The average trip rate is the average number of daily trips generated in the user’s active days. Two more trip rates are calculated separately for the active weekdays and the active weekends. The ratio between the average weekend trip rate and the weekday trip rate is included as well since the following attributes regarding departure time are measured by proportion.

Table 4-1. Attributes for Travel Behavior Characteristics

| Category | Attribute | Description |
|---|---|---|
| Trip Rate | Avg_Trip | Average daily trip rate |
| | Avg_Trip_WDay | Average weekday daily trip rate |
| | Avg_Trip_WEnd | Average weekend daily trip rate |
| | WDay_WEnd_trip_rate | The ratio of trip rate on weekends over that on weekdays |
| Departure Time | WDay/WEnd_am_prob | % of trips starting at AM peak on weekdays/weekends |
| | WDay/WEnd_md_prob | % of trips starting at midday on weekdays/weekends |
| | WDay/WEnd_pm_prob | % of trips starting at PM peak on weekdays/weekends |
| | WDay/WEnd_nt_prob | % of trips starting at night on weekdays/weekends |
| | WDay/WEnd_am_var | % of trips with the dominant OD pair within the weekday/weekend AM peak |
| | WDay/WEnd_md_var | % of trips with the dominant OD pair within the weekday/weekend midday |
| | WDay/WEnd_pm_var | % of trips with the dominant OD pair within the weekday/weekend PM peak |
| | WDay/WEnd_nt_var | % of trips with the dominant OD pair within the weekday/weekend night |
| | WDay/WEnd_am_dist | Distance of the dominant OD pair within the weekday/weekend AM peak |
| | WDay/WEnd_md_dist | Distance of the dominant OD pair within the weekday/weekend midday |
| | WDay/WEnd_pm_dist | Distance of the dominant OD pair within the weekday/weekend PM peak |
| | WDay/WEnd_nt_dist | Distance of the dominant OD pair within the weekday/weekend night |
| Travel Time | TT_Q0 | Minimum travel time |
| | TT_Q5/25/50/75/95 | 5th/25th/50th/75th/95th percentile of travel time |
| | TT_Q1 | Maximum travel time |
| OD Distance | TD_Q0 | Minimum OD distance |
| | TD_Q5/25/5/75/95 | 5th/25th/50th/75th/95th percentile of OD distance |
| | TD_Q1 | Maximum OD distance |
| Maximum Speed Recorded | Max_Spd_Q0 | Minimum of the maximum speed recorded |
| | Max_Spd_Q5/25/50/75/95 | 5th/25th/50th/75th/95th percentile of the maximum speed recorded |
| | Max_Spd_Q1 | Maximum of the maximum speed recorded |

The second part is the characteristics of departure time. One day is divided into four segments: AM peak (6 - 10 a.m.), midday (10 a.m. - 3 p.m.), PM peak (3 - 7 p.m.), night (7 p.m. - 6 a.m.). The difference between weekday and weekend should also be included, which leads to eight time periods in total. Besides the percentage of trips starting at different time periods (e.g., WDay_am_prob), the most frequent or

dominant OD pair within each time period is extracted to evaluate the variation and feature of the user's travel pattern. The dominant OD pair is defined as follows:

1) Summarize the frequency of the OD pairs ($p \in P$) in each time period, where OD is considered at the block group level. Sort the OD pairs by frequency.

2) From the most frequent OD pair $p_1 \in P$, compare the average coordinates of trip ends for $p_1$ with those for the remaining OD pairs ($p_i, i > 1$). If the distance between OD pair $p_j$ and $p_1$ is less than 1km, then $p_j$ is combined with $p_1$ to become $p^* \in Q$, where $Q$ represents the set of the adjusted OD pairs. Otherwise, $p_1$ is removed from $P$ to $Q$.

3) Repeat 2) until there is no OD pair in $P$. Sort the OD pairs in $Q$ by frequency. The first OD pair in $Q$ is defined as the dominant OD pair.

The variation of travel pattern is measured by the percentage of the trips following the dominant OD pair within each time period (e.g., WDay_am_var). The less varied the user's travel pattern is, the higher the value will be. The distances of the dominant OD pairs are included since they probably represent the daily commuting distance or the typical grocery shopping distance of the user.

Other attributes concerning the distribution of travel time, OD distance, and the maximum speed recorded are involved. As mentioned in Section 3.2.2, the speed information is missing for many records but the maximum speed recorded can also be interpretative. The records with speed larger than 120 mph are removed since they are observed to be part of air trips and thus the trips are always incomplete due to signal

issue during the flight. For all the three values, the 5th, 25th, 50th, 75th, and 95th percentile are taken into consideration in addition to the minimum and maximum.

#### 4.3.2. Geographic Information

In this study, the comprehensive land use data is considered for two CBGs where the user's home and work locations belong. Only a small portion of people reported the home location or recorded trips with home/work as the purpose, so the home and work CBGs are inferred based on the frequency and departure time as follows:

- The two most frequent OD pairs are denoted as $p_i(t_i, o_i, d_i)$ $(i = 1, 2)$, where $t_i$ represents the time period when the OD pair is travelled, $o_i$ is the origin CBG of the OD pair and $d_i$ is the destination CBG. The five most visited CBGs are counted as $d\_CBG_j, j = 1, \dots, 5$.

- If the two most frequent OD pairs are generated in weekday AM or PM peak and one of the two most visited CBGs is either the origin of the AM peak OD pair or the destination of the PM peak OD pair, then the CBG is labelled as the home CBG:

  $K_1 = \{o \in p(t = WDay\_am), d \in p(t = WDay\_pm)\}$

  if $K_1 \neq \emptyset$

  for $j$=1 to 2

  if $d\_CBG_j \in K_1$

  $home\_CBG = d\_CBG_j,\ j_{home} = j$

- If the home CBG is not found, then check if one of the two most visited CBGs is either the origin or destination of the OD pair during weekday night:

  if $K_1 = \emptyset$

  $K_2 = \{o \in p(t = WDay\_nt), d \in p(t = WDay\_nt)\}$

  if $K_2 \neq \emptyset$

  for $j$=1 to 2

  if $d\_CBG_j \in K_2$

  $home\_CBG = d\_CBG_j, j_{home} = j$

- If the two most frequent OD pairs are generated in weekday AM or PM peak and one of the five most visited CBGs is either the destination of the AM peak OD pair or the origin of the PM peak OD pair, then the CBG is labelled as the work CBG:

  $K_3 = \{d \in p(t = WDay\_am), o \in p(t = WDay\_pm)\}$

  if $K_3 \neq \emptyset$

  for $j$=1 to 5

  if $d\_CBG_j \in K_3$

  $work\_CBG = d\_CBG_j, j_{work} = j$

- If the home CBG is still not found, the most visited CBG which is not labelled as the work CBG will be labelled as the home CBG:

  if $j_{home} = null$

  $j_{home} = \min j \in \{1 \leq j \leq 5, j \neq j_{work}\}$

- If the work CBG is not found, the most visited CBG which is not labelled as the home CBG will be labelled as the work CBG:

  if $j_{work} = null$

  $$j_{work} = \min j \in \{1 \leq j \leq 5, j \neq j_{home}\}$$

- The home CBG and work CBG are found:

  $$home\_CBG = d\_CBG_{j_{home}}$$

  $$work\_CBG = d\_CBG_{j_{work}}$$

Once the home and work CBGs are defined, the area characteristics will be introduced from the SLD as summarized in Table 4-2.

Table 4-2. Attributes for Geographic Information

| Location | Attribute | Description |
|---|---|---|
| Both home and work CBGs | Ac_Unpr | Total land area in acres that is not protected from development (i.e., not a park or conservation area) |
| | D1a | Gross residential density (HU/acre) on unprotected land |
| | D1b | Gross population density (people/acre) on unprotected land |
| | D1c | Gross employment density (jobs/acre) on unprotected land |
| | D2a_EpHHm | Employment and household entropy |
| | D3a | Total road network density |
| | D4b050 | Proportion of CBG employment within ½ mile of fixed-guideway transit stop |
| Home CBG | home_P_WrkAge | % of population that is working aged |
| | home_P_WORKERS | % of workers in CBG |
| | home_P_LowWage | % of workers earning $1250/week or less |
| | home_P_MedWage | % of workers earning more than $1250/week but less than $3333/week |
| | home_P_HiWage | % of workers earning $3333/week or less |
| Work CBG | work_P_LowWage | % of workers earning $1250/week or less |
| | work_P_MedWage | % of workers earning more than $1250/week but less than $3333/week |
| | work_P_HiWage | % of workers earning $3333/week or less |

The area and density information are directly borrowed from Table 3-2 and other demographic attributes are computed based on the SLD. Some examples are:

$$home\_P\_Workers = Workers/TotPop \times 100\%,$$

$$home\_P\_HiWage = R\_HiWageWk/Workers \times 100\%.$$

$$work\_P\_HiWage = E\_HiWageWk/TotEmp \times 100\%.$$

#### 4.3.3. POIs and Imputed Trip Purpose

The model containing the aforementioned features only is named “Naïve Model”. In addition, the previous study [100] employed random forests to infer the trip purpose and reached more than 80% of accuracy for the in-vehicle GPS survey. Though it was concluded that POI information does not play an important role in trip purpose prediction, the frequency of visiting places with various POI categories is considered for the demographic imputation based on the in-vehicle GPS dataset (Table 4-3). The difference between “daily_poi_near” and “daily_poi_gene” is whether the POI information is extracted based on the nearest place for trip ends or based on the dominant category with a buffer of 250m. The model including features about POI information is named “POI Model”.

The imputed trip purpose is also added to the feature set for the in-vehicle GPS dataset to evaluate its contribution. The attributes related to imputed trip purpose are summarized in Table 4-4, where more details about work/shop/social trips are considered.

Table 4-3. Attributes for POI Information

| Attribute | Description |
|---|---|
| daily_poi_near/gene_1 | # of daily trips visiting places about automotive |
| daily_poi_near/gene_2 | # of daily trips visiting community service centers |
| daily_poi_near/gene_3 | # of daily trips visiting restaurants |
| daily_poi_near/gene_4 | # of daily trips visiting travel destinations |
| daily_poi_near/gene_5 | # of daily trips visiting transportation hubs |
| daily_poi_near/gene_6 | # of daily trips visiting miscellaneous places |
| daily_poi_near/gene_7 | # of daily trips visiting shopping places |
| daily_poi_near/gene_8 | # of daily trips visiting education institutions |
| daily_poi_near/gene_9 | # of daily trips visiting places about entertainment |
| daily_poi_near/gene_10 | # of daily trips visiting medical places |
| daily_poi_near/gene_11 | # of daily trips visiting business facilities |
| daily_poi_near/gene_13 | # of daily trips visiting border crossing |
| daily_poi_near/gene_14 | # of daily trips visiting parks or recreational places |
| daily_poi_near/gene_15 | # of daily trips visiting parking places |
| daily_poi_near/gene_16 | # of daily trips visiting financial institutions |
| daily_poi_gene_99 | # of daily trips visiting places where multiple categories of POIs |

Table 4-4. Attributes for Imputed Trip Purpose

| Attribute | Description |
|---|---|
| home_prob | % of home trips |
| work_prob | % of work trips |
| shop_prob | % of shopping trips |
| soci_prob | % of social/recreational trips |
| pick_prob | % of pick-up/drop-off trips |
| othe_prob | % of other trips |
| work_WDay_am/md/pm/nt | % of work trips at AM peak/midday/PM peak/night on weekdays |
| work_WEnd_am/md/pm/nt | % of work trips at AM peak/midday/PM peak/night on weekends |
| shop_WDay_am/md/pm/nt | % of shopping trips at AM peak/midday/PM peak/night on weekdays |
| shop_WEnd_am/md/pm/nt | % of shopping trips at AM peak/midday/PM peak/night on weekends |
| soci_WDay_am/md/pm/nt | % of social/recreational trips at AM peak/midday/PM peak/night on weekdays |
| soci_WEnd_am/md/pm/nt | % of social/recreational trips at AM peak/midday/PM peak/night on weekends |

# Chapter 5: Imputation Results

Based on the classifiers (CIT and CIT-based random forests) and feature sets introduced in Chapter 4, the imputation results are demonstrated and discussed within this chapter. Section 5.1 compares the prediction accuracy for the in-vehicle GPS dataset based on the four sets of features and evaluates the variable importance in each case. Section 5.2 looks into the results for the smartphone location dataset, which includes multimodal traveling data. Section 5.3 summarizes the findings from both examinations.

## *5.1. Imputation Results for the In-vehicle GPS Dataset*

To study and compare the goodness of imputation based on different feature sets, four models are specified for the in-vehicle GPS dataset in Table 5-1. The focus of the comparison will be evaluating the prediction strength of POI information and imputed purpose, which could provide some suggestions for feature set construction in future research.

Table 5-1. Model Specification

| | Naïve Model | POI Model | Purpose Model | Full Model |
|---|---|---|---|---|
| Travel Behavior Statistics | ✓ | ✓ | ✓ | ✓ |
| Geographic Information | ✓ | ✓ | ✓ | ✓ |
| POI Information | | ✓ | | ✓ |
| Imputed Trip Purpose | | | ✓ | ✓ |

#### 5.1.1. Gender

The imputation results for the test datasets are listed in Table 5-2 categorized by the four model specifications. The 7-fold cross-validation is employed to evaluate the model performance. To make the results comparable, the random seeds to generate the bootstrap samples are fixed.

In the table, "Recall" is the fraction of relevant instances that have been retrieved over the total amount of relevant instances (i.e., the proportion of correctly imputed instances) and "Precision" is the fraction of relevant instances among the retrieved instances. "F1" is the F1 score, which is the harmonic mean of precision and recall ($F_1 = 2 \times \frac{precision \times recall}{precision + recall}$). "Overall" is the proportion of correctly imputed instances for both groups. "CIT" represents the single CIT classifier with the significance level α for the variable association test. "Random Forests" represents the CIT-based random forests and the number of trees is by default set as 500.

The classifier with the best and balanced performance is marked in bold for each model. The criteria include that the overall accuracy increase should be greater than the summation of the F1 score increase and the F1 scores should all be greater than 10%. For example, the overall accuracy of the random forest classifier (58.38%) is larger than that of the CIT with 10% significance level in the naïve model, but the summation of the F1 score increase (-16.27%+7.30%=-8.97%) is smaller than the accuracy increase (1.54%). As a result, the CIT with 10% significance level is marked as the best model.

Within each model, the two CIT classifiers have the better and more balanced prediction strength while the CIT-based random forest shows a strong tendency to predict the "Male" group correctly whose sample size is 30% larger. Among the four models, the naïve model has the highest accuracy. It indicates that POI information and imputed purpose do not benefit the imputation of gender much. Nevertheless, the travel behavior statistics and the geographic information of the imputed home and work locations have provided considerable evidence for gender prediction.

Table 5-2. Imputation Accuracy for Gender

| | Recall | | Precision | | F1 | | Overall |
|---|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male | |
| Naïve Model | | | | | | | |
| CIT (α=0.05) | 0.2783 | 0.7054 | 0.4285 | 0.5656 | 0.3374 | 0.6278 | 0.5135 |
| **CIT (α =0.10)** | **0.4837** | **0.6222** | **0.4937** | **0.6230** | **0.4887** | **0.6226** | **0.5684** |
| Random Forests | 0.2190 | 0.8496 | 0.6048 | 0.5889 | 0.3215 | 0.6956 | 0.5838 |
| POI Model | | | | | | | |
| CIT (α =0.05) | 0.4733 | 0.5366 | 0.4254 | 0.5833 | 0.4480 | 0.5590 | 0.5127 |
| **CIT (α =0.10)** | **0.5241** | **0.5078** | **0.4484** | **0.5781** | **0.4833** | **0.5407** | **0.5249** |
| Random Forests | 0.1594 | 0.7921 | 0.3337 | 0.5548 | 0.2157 | 0.6525 | 0.5227 |
| Purpose Model | | | | | | | |
| **CIT (α =0.05)** | **0.4676** | **0.5891** | **0.4802** | **0.5924** | **0.4738** | **0.5908** | **0.5365** |
| CIT (α =0.10) | 0.5247 | 0.5212 | 0.4465 | 0.5942 | 0.4824 | 0.5553 | 0.5240 |
| Random Forests | 0.1883 | 0.8247 | 0.5286 | 0.5704 | 0.2777 | 0.6744 | 0.5525 |
| Full Model | | | | | | | |
| **CIT (α =0.05)** | **0.4354** | **0.6351** | **0.4883** | **0.5975** | **0.4603** | **0.6157** | **0.5489** |
| CIT (α =0.10) | 0.4782 | 0.5841 | 0.4645 | 0.5958 | 0.4712 | 0.5899 | 0.5365 |
| Random Forests | 0.2417 | 0.8293 | 0.5803 | 0.5889 | 0.3413 | 0.6887 | 0.5779 |

The CIT with 10% significance level in the naïve model is visualized in Figure 5-1. It can be noted that the instances all belong to class "Male" when the 5 percentiles of OD distances are longer than 0.4km and the OD distance of the dominant OD pair during weekday PM peak is larger than 10.5km. It may indicate that males intend to

have longer commuting distance. Almost all the instances belong to class “Female” if the 5 percentiles of OD distances are shorter than 0.4km and the transit accessibility of home is higher. It can be summarized that females tend to live in denser block groups and the areas with larger proportion of working-aged people.

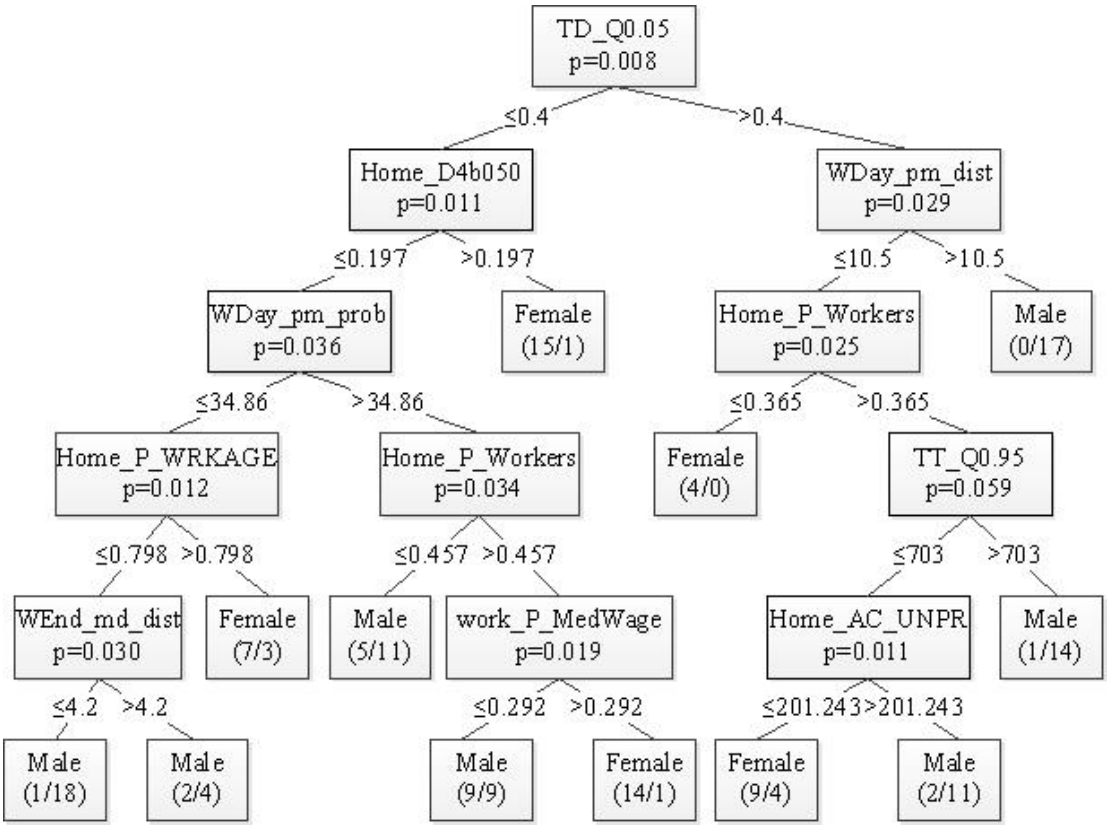


Figure 5-1. Gender: The CIT with 10% Significance Level in the Naïve Model

### 5.1.2. Age Group

The model has been tested for age of three groups and two groups. The cut points are selected to be 35 and 65 years old. The assumptions are that people under 35 are thought to be young and travel more diversely, people aged between 35 and 65 are more mature and may follow less varied travel patterns, and people over 65 are often retired and may travel less than the other two groups. However, there is only one

sample unit aged over 65 years old from the in-vehicle GPS survey, which is not sufficient for training the model. As a result, the last two groups are combined and the age is finally categorized as “Under 35” and “35+”.

In Table 5-3, random forests have better performance in most cases but the bias towards “35+” class (whose sample size is four times larger) is still significant. Though the accuracy of the naïve model is the highest, the POI information and the imputed purpose have significantly increased the prediction strength for the “Under 35” class. It indicates that the attributes regarding POIs help to impute the age group a lot. The full model does not outperform the POI model and purpose model, which may result from the strong association between POIs and trip purposes.

Table 5-3. Imputation Accuracy for Age Group without Weight Adjustment

| | Recall | | Precision | | F1 | | Overall |
|---|---|---|---|---|---|---|---|
| | Under 35 | 35+ | Under 35 | 35+ | Under 35 | 35+ | |
| Naïve Model | | | | | | | |
| CIT (α=0.05) | 0.1524 | 0.8260 | 0.1537 | 0.8127 | 0.1531 | 0.8193 | 0.7149 |
| CIT (α =0.10) | 0.1524 | 0.8181 | 0.1512 | 0.8103 | 0.1518 | 0.8142 | 0.7087 |
| **Random Forests** | **0.0929** | **0.9211** | **0.1083** | **0.8200** | **0.1000** | **0.8676** | **0.7648** |
| POI Model | | | | | | | |
| CIT (α=0.05) | 0.1524 | 0.7531 | 0.1463 | 0.7948 | 0.1493 | 0.7734 | 0.6536 |
| CIT (α =0.10) | 0.1524 | 0.7531 | 0.1463 | 0.7948 | 0.1493 | 0.7734 | 0.6536 |
| **Random Forests** | **0.3405** | **0.8662** | **0.3503** | **0.8383** | **0.3453** | **0.8520** | **0.7522** |
| Purpose Model | | | | | | | |
| **CIT (α=0.05)** | **0.3524** | **0.7816** | **0.3129** | **0.8445** | **0.3315** | **0.8118** | **0.7157** |
| **CIT (α =0.10)** | **0.3524** | **0.7816** | **0.3129** | **0.8445** | **0.3315** | **0.8118** | **0.7157** |
| Random Forests | 0.0714 | 0.8927 | 0.1429 | 0.8109 | 0.0952 | 0.8498 | 0.7465 |
| Full Model | | | | | | | |
| CIT (α=0.05) | 0.3238 | 0.7904 | 0.3097 | 0.8394 | 0.3166 | 0.8141 | 0.7157 |
| CIT (α =0.10) | 0.3238 | 0.7904 | 0.3097 | 0.8394 | 0.3166 | 0.8141 | 0.7157 |
| **Random Forests** | **0.3071** | **0.8484** | **0.2863** | **0.8194** | **0.2963** | **0.8336** | **0.7271** |

The default case weight is one for all instances. Since the number of instances in the "35+" group is more than double that in the "Under 35" group, an integer weight is applied to each instance in the "Under 35" group based on the number of each class in each training dataset within the cross-validation. The weighting strategy is as follows. For class $c_m$ $(m = 1, \dots, M)$, there is class $c_{m^*}$ with the largest number of instances $n_{m^*}$. The weight for each instance in class $c_{m^*}$ is set as one ($\omega_{i \in c_{m^*}} = 1$) and those for the instances in other classes ($\omega_{i \in c_m}, m \neq m^*$) are set as the integer part of $\frac{n_{m^*}}{n_m}$.

As shown in Table 5-4, the weight adjustment has helped to decrease the error rate of imputing the "Under 35" group when the CIT classifier is applied. However, it does not benefit the random forest classifier.

Table 5-4. Imputation Accuracy for Age Group with Weight Adjustment

| | Recall | | Precision | | F1 | | Overall |
|---|---|---|---|---|---|---|---|
| | Under 35 | 35+ | Under 35 | 35+ | Under 35 | 35+ | |
| Naïve Model | | | | | | | |
| **CIT (α=0.05)** | **0.4167** | **0.8365** | **0.3143** | **0.8555** | **0.3583** | **0.8459** | **0.7462** |
| CIT (α =0.10) | 0.3929 | 0.8365 | 0.3000 | 0.8490 | 0.3402 | 0.8427 | 0.7400 |
| Random Forests | 0.3143 | 0.6606 | 0.1935 | 0.8203 | 0.2395 | 0.7318 | 0.6052 |
| POI Model | | | | | | | |
| **CIT (α=0.05)** | **0.4881** | **0.7372** | **0.2706** | **0.8515** | **0.3482** | **0.7902** | **0.6787** |
| **CIT (α =0.10)** | **0.4881** | **0.7372** | **0.2706** | **0.8515** | **0.3482** | **0.7902** | **0.6787** |
| Random Forests | 0.2643 | 0.6637 | 0.1832 | 0.8091 | 0.2164 | 0.7292 | 0.6038 |
| Purpose Model | | | | | | | |
| **CIT (α=0.05)** | **0.4810** | **0.7517** | **0.2532** | **0.8631** | **0.3318** | **0.8036** | **0.6989** |
| **CIT (α =0.10)** | **0.4810** | **0.7517** | **0.2532** | **0.8631** | **0.3318** | **0.8036** | **0.6989** |
| Random Forests | 0.3405 | 0.6543 | 0.2058 | 0.8168 | 0.2565 | 0.7266 | 0.5991 |
| Full Model | | | | | | | |
| CIT (α=0.05) | 0.3952 | 0.7226 | 0.1798 | 0.8395 | 0.2472 | 0.7767 | 0.6557 |
| CIT (α =0.10) | 0.3952 | 0.7226 | 0.1798 | 0.8395 | 0.2472 | 0.7767 | 0.6557 |
| **Random Forests** | **0.5762** | **0.6475** | **0.2768** | **0.8513** | **0.3739** | **0.7356** | **0.6238** |

Looking into the CIT predictor with 5% significance level and weight adjustment in the naïve model (Figure 5-2), it can be summarized that the younger people (under 35) tend to take longer trips. They tend to live in the areas with higher proportion of workers earning median wages but smaller proportion of workers earning high wages. The areas with smaller block group size and higher density of road networks are also preferred by the "under 35" group. In contrast, the "35+" group tends to take shorter trips, e.g. on weekend midday. They usually live in the areas with more high income workers and larger block group size. They are probably insensitive to the road network density. Overall, they may prefer to live in suburban areas.

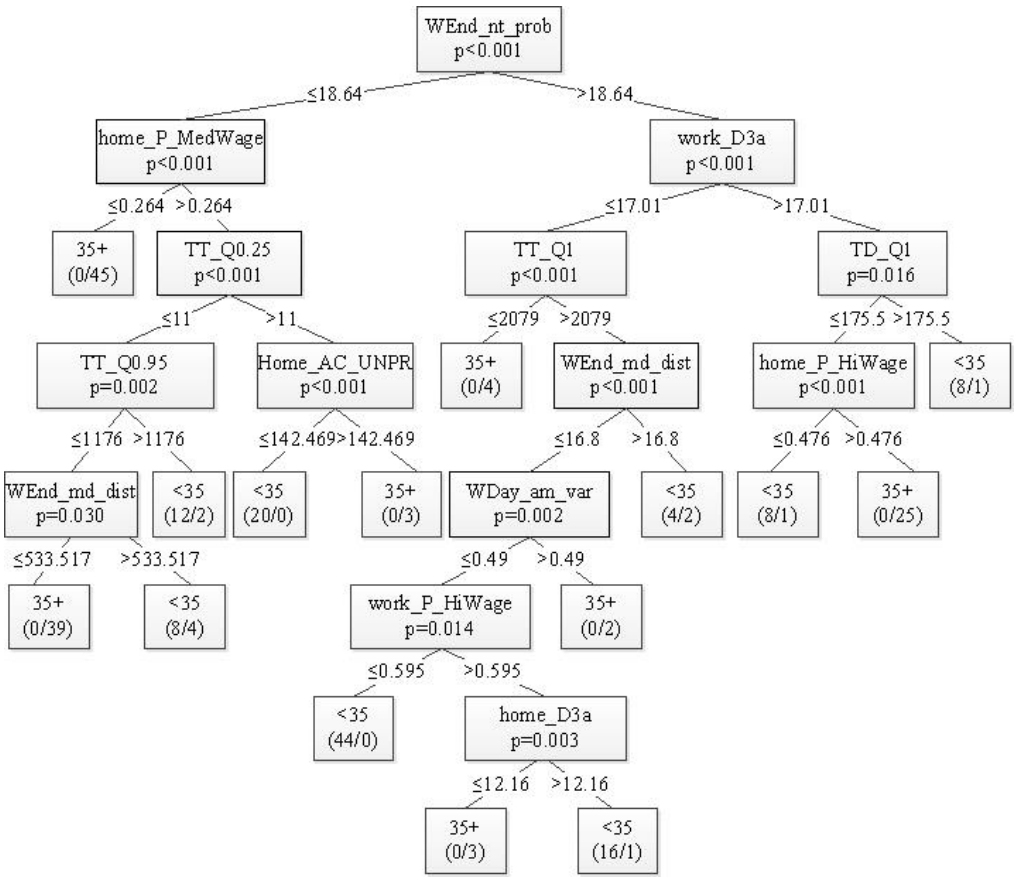


Figure 5-2. Age: The CIT with 5% Significance Level and Weight Adjustment in the Naïve Model

5.1.3. Education Level

The education level is originally surveyed with six categories (high school, associate degree, some college, bachelor's degree, master's degree, and doctoral degree) but the first two categories only covered two and four subjects. It is later regrouped into three levels: less than bachelor's degree (LB), bachelor's degree (B), and graduate degree (G). The imbalance also exists among the three classes so the model with weight adjustment is examined.

Table 5-5. Imputation Accuracy for Education Level without Weight Adjustment

| | Recall | | | Precision | | | F1 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | LB | B | G | LB | B | G | LB | B | G | |
| Naïve Model | | | | | | | | | | |
| **CIT α=0.05** | **0.2476** | **0.2905** | **0.7246** | **0.2262** | **0.3470** | **0.5962** | **0.2364** | **0.3162** | **0.6541** | **0.4956** |
| CIT α =0.10 | 0.3190 | 0.3569 | 0.6075 | 0.1706 | 0.4063 | 0.6020 | 0.2224 | 0.3800 | 0.6047 | 0.4641 |
| Random Forests | 0.1000 | 0.3563 | 0.5319 | 0.1476 | 0.2700 | 0.5852 | 0.1192 | 0.3072 | 0.5572 | 0.3938 |
| POI Model | | | | | | | | | | |
| **CIT α=0.05** | **0.5000** | **0.3804** | **0.4466** | **0.3241** | **0.3663** | **0.5573** | **0.3933** | **0.3732** | **0.4959** | **0.4335** |
| CIT α =0.10 | 0.5000 | 0.3661 | 0.4336 | 0.2935 | 0.3663 | 0.5471 | 0.3699 | 0.3662 | 0.4838 | 0.4211 |
| Random Forests | 0.2143 | 0.4408 | 0.5198 | 0.1420 | 0.3541 | 0.5397 | 0.1708 | 0.3927 | 0.5295 | 0.4322 |
| Purpose Model | | | | | | | | | | |
| **CIT α=0.05** | **0.2238** | **0.3978** | **0.4890** | **0.2349** | **0.3462** | **0.5247** | **0.2292** | **0.3702** | **0.5062** | **0.4022** |
| CIT α =0.10 | 0.2810 | 0.4754 | 0.4061 | 0.2111 | 0.3855 | 0.5407 | 0.2411 | 0.4258 | 0.4639 | 0.3960 |
| Random Forests | 0.1190 | 0.2329 | 0.5921 | 0.0397 | 0.1746 | 0.5231 | 0.0595 | 0.1996 | 0.5554 | 0.3832 |
| Full Model | | | | | | | | | | |
| **CIT α=0.05** | **0.5000** | **0.4034** | **0.4850** | **0.3024** | **0.4512** | **0.5652** | **0.3769** | **0.4260** | **0.5221** | **0.4581** |
| **CIT α =0.10** | **0.5000** | **0.4605** | **0.4630** | **0.3024** | **0.4274** | **0.5845** | **0.3769** | **0.4433** | **0.5167** | **0.4581** |
| Random Forests | 0.3238 | 0.2793 | 0.4539 | 0.3429 | 0.2015 | 0.4895 | 0.3331 | 0.2341 | 0.4710 | 0.3521 |

In Table 5-5, it can be observed that the CIT classifier with 5% significance level performs better for each model specification. The POI information and imputed purpose do not improve the prediction accuracy significantly but they to some extent help correctly impute the "LB" class whose sample size is the smallest.

The imputation results with weight adjustment have been listed in Table 5-6. The weight adjustment does not increase the prediction accuracy though in some models, such as the naïve model and the POI model, it helps to identify the "LB" class more.

Table 5-6. Imputation Accuracy for Education Level with Weight Adjustment

| | Recall | | | Precision | | | F1 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | LB | B | G | LB | B | G | LB | B | G | |
| Naïve Model | | | | | | | | | | |
| CIT α=0.05 | 0.3905 | 0.2338 | 0.5483 | 0.2204 | 0.3141 | 0.5249 | 0.2818 | 0.2680 | 0.5363 | 0.4062 |
| CIT α =0.10 | 0.3905 | 0.2338 | 0.5223 | 0.2204 | 0.3141 | 0.5117 | 0.2818 | 0.2680 | 0.5170 | 0.3938 |
| **Random Forests** | **0.4429** | **0.1692** | **0.5776** | **0.2549** | **0.3750** | **0.5721** | **0.3236** | **0.2331** | **0.5748** | **0.4319** |
| POI Model | | | | | | | | | | |
| CIT α=0.05 | 0.5095 | 0.2931 | 0.4940 | 0.3699 | 0.3253 | 0.5000 | 0.4286 | 0.3083 | 0.4970 | 0.4127 |
| **CIT α =0.10** | **0.5095** | **0.3645** | **0.4736** | **0.3699** | **0.3396** | **0.5136** | **0.4286** | **0.3516** | **0.4928** | **0.4127** |
| Random Forests | 0.5952 | 0.2297 | 0.4835 | 0.2420 | 0.5048 | 0.5472 | 0.3441 | 0.3157 | 0.5134 | 0.4067 |
| Purpose Model | | | | | | | | | | |
| **CIT α=0.05** | **0.3000** | **0.2517** | **0.4477** | **0.1888** | **0.2341** | **0.5039** | **0.2318** | **0.2426** | **0.4742** | **0.3575** |
| **CIT α =0.10** | **0.3000** | **0.2517** | **0.4477** | **0.1888** | **0.2341** | **0.5039** | **0.2318** | **0.2426** | **0.4742** | **0.3575** |
| Random Forests | 0.5238 | 0.0143 | 0.5665 | 0.2196 | 0.0286 | 0.5125 | 0.3095 | 0.0190 | 0.5381 | 0.3833 |
| Full Model | | | | | | | | | | |
| CIT α=0.05 | 0.3143 | 0.2471 | 0.5059 | 0.2279 | 0.2501 | 0.5247 | 0.2642 | 0.2486 | 0.5151 | 0.3889 |
| CIT α =0.10 | 0.3143 | 0.3185 | 0.4855 | 0.2279 | 0.2653 | 0.5383 | 0.2642 | 0.2894 | 0.5105 | 0.3889 |
| **Random Forests** | **0.5381** | **0.2549** | **0.5719** | **0.3214** | **0.3889** | **0.6359** | **0.4025** | **0.3079** | **0.6022** | **0.4559** |

The CIT with 5% significance level and without weight adjustment in the naïve model is visualized in Figure 5-3. Taking the terminal node with the most instances imputed correctly for class "G" as an example, the people with graduate or professional degree tend to drive carefully and take shorter trips at PM peak on weekends. They also tend to live in the areas with lower roadway density. For the "LB" group, they tend to drive faster and take longer trips at PM peak on weekends. The "B" group is more similar to the "G" group — they prefer to work in the areas with higher land use diversity and take shorter trips on weekend PM peak.

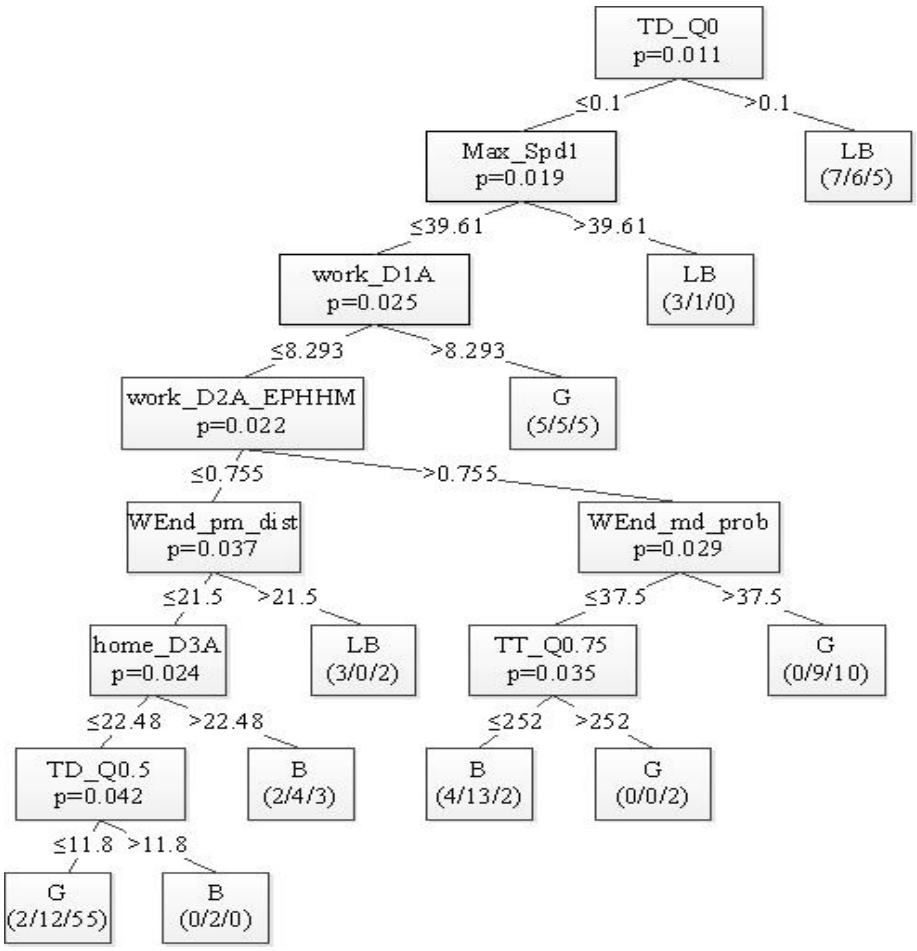


Figure 5-3. Education: The CIT with 5% Significance Level and without Weight Adjustment in the Naïve Model

5.1.4. Household Income Level

The household income level is originally surveyed with seven categories as described in Section 3.1 and later regrouped into three levels: less than $50,000 (low), $50,000-$150,000 (middle), $150,000+ (high). The levels are defined according to the low/median/high weekly wages in the SLD. Since the 7-fold cross-validation may result in many cases of zero instance for a certain group in the test dataset, the 3-fold cross-validation is applied instead. The imputation results without weight adjustment are listed in Table 5-7.

Table 5-7. Imputation Accuracy for Income Level without Weight Adjustment

| | Recall | | | Precision | | | F1 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Low | Mid | High | Low | Mid | High | Low | Mid | High | |
| Naïve Model | | | | | | | | | | |
| **CIT α=0.05** | **0.1333** | **0.6880** | **0.1171** | **0.1818** | **0.6156** | **0.1454** | **0.1538** | **0.6498** | **0.1297** | **0.4906** |
| CIT α =0.10 | 0.1333 | 0.6678 | 0.1367 | 0.1818 | 0.6111 | 0.1556 | 0.1538 | 0.6382 | 0.1455 | 0.4845 |
| Random Forests | 0.0000 | 0.7589 | 0.2343 | - | 0.6372 | 0.2611 | - | 0.6927 | 0.2470 | 0.5525 |
| POI Model | | | | | | | | | | |
| **CIT α=0.05** | **0.2444** | **0.5963** | **0.1409** | **0.1185** | **0.6235** | **0.1576** | **0.1596** | **0.6096** | **0.1488** | **0.4484** |
| CIT α =0.10 | 0.2444 | 0.5788 | 0.1409 | 0.1051 | 0.6177 | 0.1576 | 0.1470 | 0.5976 | 0.1488 | 0.4362 |
| Random Forests | 0.0000 | 0.8565 | 0.1320 | - | 0.6498 | 0.1548 | - | 0.7390 | 0.1425 | 0.5822 |
| Purpose Model | | | | | | | | | | |
| CIT α=0.05 | 0.1778 | 0.7146 | 0.1213 | 0.1018 | 0.6217 | 0.2571 | 0.1295 | 0.6649 | 0.1648 | 0.5045 |
| CIT α =0.10 | 0.1778 | 0.6742 | 0.1998 | 0.1018 | 0.6328 | 0.3095 | 0.1295 | 0.6528 | 0.2428 | 0.5045 |
| **Random Forests** | **0.1333** | **0.6951** | **0.2884** | **0.1250** | **0.6567** | **0.2990** | **0.1290** | **0.6754** | **0.2936** | **0.5344** |
| Full Model | | | | | | | | | | |
| CIT α=0.05 | 0.1778 | 0.6924 | 0.0975 | 0.1111 | 0.6406 | 0.1870 | 0.1368 | 0.6655 | 0.1282 | 0.4856 |
| **CIT α =0.10** | **0.1778** | **0.6520** | **0.1760** | **0.1111** | **0.6538** | **0.2394** | **0.1368** | **0.6529** | **0.2029** | **0.4856** |
| Random Forests | 0.0000 | 0.7570 | 0.1974 | 0.0000 | 0.6237 | 0.2922 | - | 0.6839 | 0.2356 | 0.5337 |

Among the four models, the random forest classifier in the purpose model has the highest prediction accuracy. But the random forest classifier in either the naïve model or the purpose model fails to identify any instance from the low income group. It is an extreme case of the previous observed bias as the sample size of the main income class "Mid" is six times larger than that of the low income class.

The results with weight adjustment are listed in Table 5-8. Although the weight adjustment benefits the imputation of the low and high income group, it has not improved the overall accuracy.

Table 5-8. Imputation Accuracy for Income Level with Weight Adjustment

| | Recall | | | Precision | | | F1 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Low | Mid | High | Low | Mid | High | Low | Mid | High | |
| Naïve Model | | | | | | | | | | |
| **CIT α=0.05** | **0.2444** | **0.6523** | **0.2469** | **0.2063** | **0.6255** | **0.2698** | **0.2237** | **0.6386** | **0.2578** | **0.5027** |
| **CIT α =0.10** | **0.2444** | **0.6523** | **0.2469** | **0.2063** | **0.6255** | **0.2698** | **0.2237** | **0.6386** | **0.2578** | **0.5027** |
| Random Forests | 0.1778 | 0.4263 | 0.3948 | 0.0812 | 0.6404 | 0.2894 | 0.1115 | 0.5119 | 0.3340 | 0.3981 |
| POI Model | | | | | | | | | | |
| **CIT α=0.05** | **0.3000** | **0.5407** | **0.2035** | **0.1935** | **0.6190** | **0.1301** | **0.2353** | **0.5772** | **0.1587** | **0.4237** |
| **CIT α =0.10** | **0.3000** | **0.5407** | **0.2035** | **0.1935** | **0.6190** | **0.1301** | **0.2353** | **0.5772** | **0.1587** | **0.4237** |
| Random Forests | 0.5952 | 0.2297 | 0.4835 | 0.2420 | 0.5048 | 0.5472 | 0.3441 | 0.3157 | 0.5134 | 0.4067 |
| Purpose Model | | | | | | | | | | |
| **CIT α=0.05** | **0.3889** | **0.5794** | **0.2212** | **0.1958** | **0.6548** | **0.2683** | **0.2605** | **0.6148** | **0.2425** | **0.4668** |
| **CIT α =0.10** | **0.3889** | **0.5794** | **0.2212** | **0.1958** | **0.6548** | **0.2683** | **0.2605** | **0.6148** | **0.2425** | **0.4668** |
| Random Forests | 0.1222 | 0.4310 | 0.3705 | 0.0463 | 0.6583 | 0.3050 | 0.0672 | 0.5209 | 0.3346 | 0.3803 |
| Full Model | | | | | | | | | | |
| CIT α=0.05 | 0.2556 | 0.6206 | 0.1236 | 0.1741 | 0.6199 | 0.1702 | 0.2071 | 0.6202 | 0.1432 | 0.4553 |
| CIT α =0.10 | 0.2556 | 0.6206 | 0.1236 | 0.1741 | 0.6199 | 0.1702 | 0.2071 | 0.6202 | 0.1432 | 0.4553 |
| **Random Forests** | **0.3667** | **0.4893** | **0.4472** | **0.1714** | **0.6826** | **0.3426** | **0.2336** | **0.5700** | **0.3880** | **0.4791** |

The importance for the first 15 variables in random forest classifier without weight adjustment in the purpose model is listed as Figure 5-4. The employment density in the home location shows the strongest prediction power. The proportion of people at working age and the proportion of people earning low wages both help to classify the income level. Among the attributes related to travel behaviors, the social/recreational trips on weekends has the highest importance following by the variation of weekday trips, the first quantile of travel time, OD distance statistics, etc.

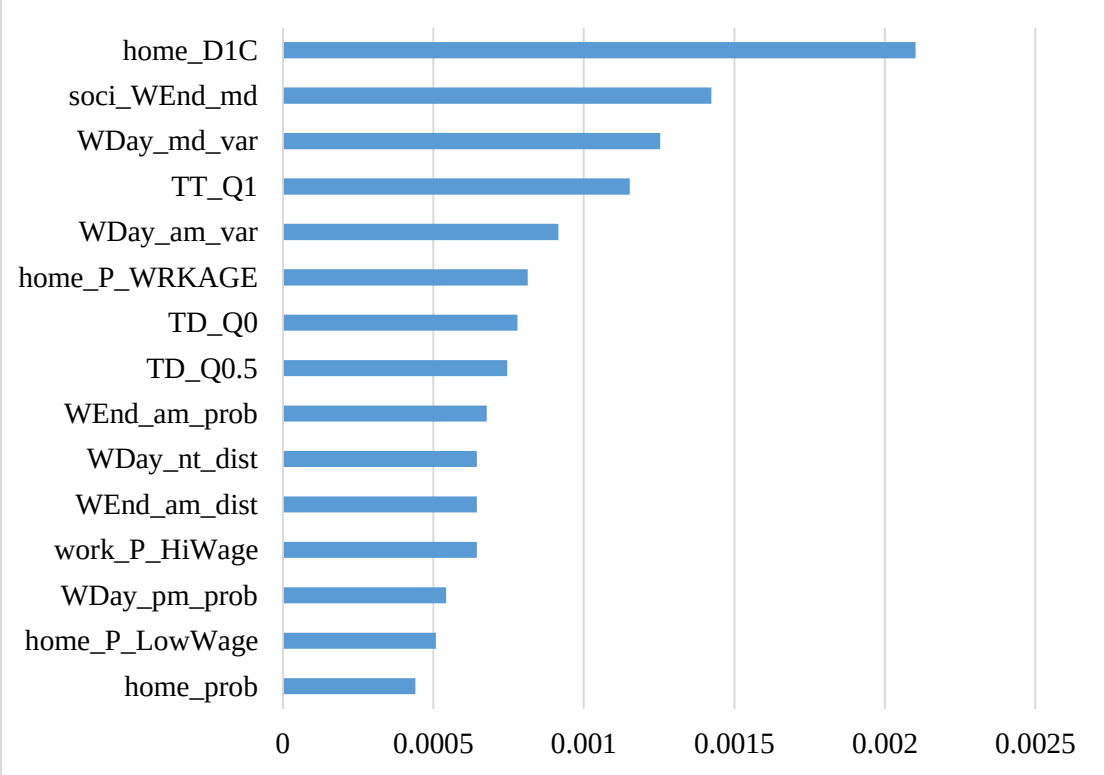


Figure 5-4. Income: Variable Importance without Weight Adjustment in the Purpose Model

## *5.2. Imputation Results for the Smartphone Location Dataset*

### 5.2.1. Gender

The imputation results for the test datasets in the 7-fold cross-validation are listed in Table 5-9. In general, the CIT with 10% significance level slightly outperforms the other two classifiers. It can be observed that CITs have more balanced accuracy for both groups while random forests seem to have higher accuracy for "Female" group, which has more instances.

Table 5-9. Imputation Accuracy of Naïve Model for Gender

| | Recall | | Precision | | F1 | | Overall |
|---|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male | |
| CIT (α=0.05) | 0.5006 | 0.5321 | 0.5747 | 0.4620 | 0.5351 | 0.4945 | 0.5206 |
| **CIT (α=0.10)** | **0.5328** | **0.5495** | **0.5814** | **0.5103** | **0.5560** | **0.5292** | **0.5274** |
| Random Forests | 0.7571 | 0.2986 | 0.5885 | 0.3763 | 0.6622 | 0.3330 | 0.5270 |

The CIT with 10% significance level is visualized in Figure 5-5. It can be summarized that females tend to take fewer trips on weekend night. Considering the attributes about geographic information, the work location of female has higher employment density and higher proportion of workers earning median wages. Males seem to have a travel routine at night on weekends.

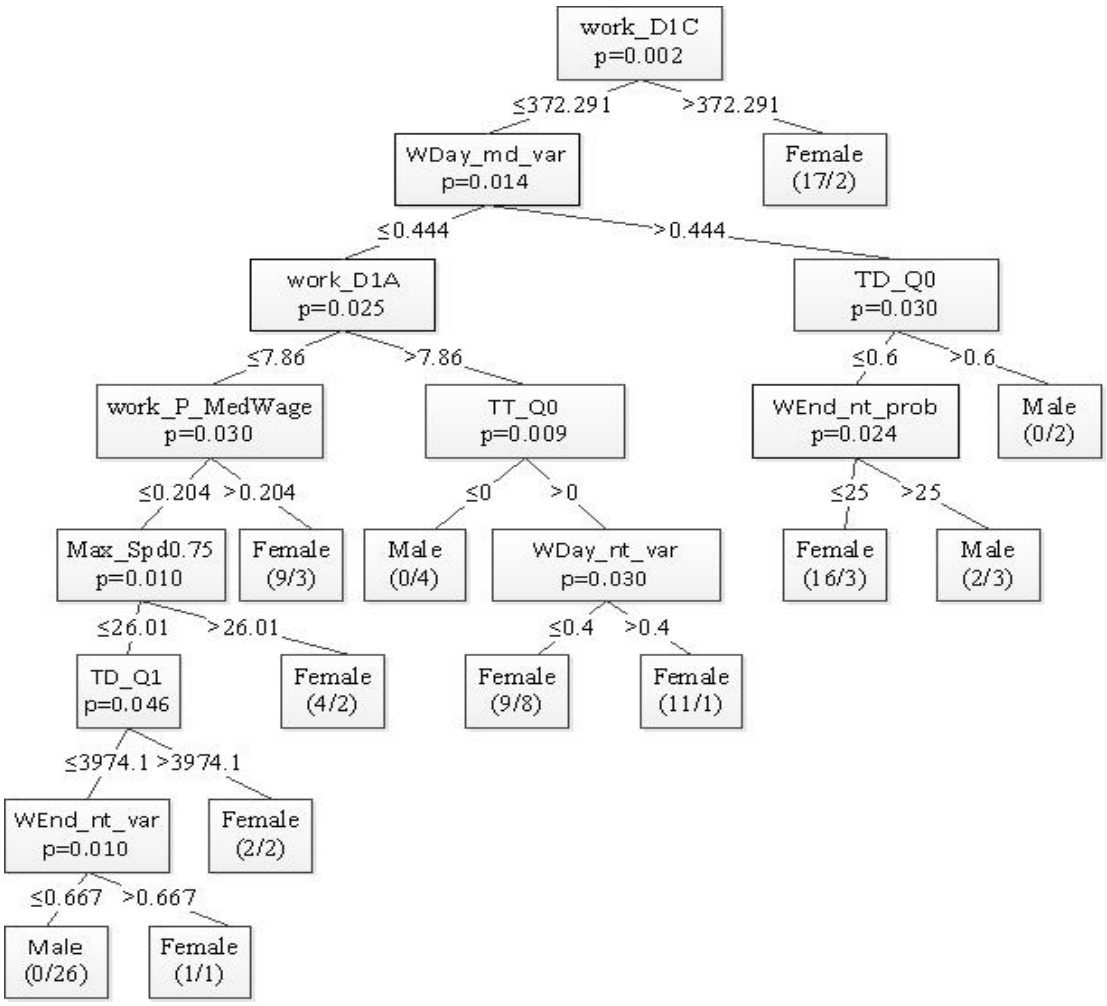


Figure 5-5. Gender: The CIT with 10% Significance Level

### 5.2.2. Age Group

The model has been tested for age of three groups and two groups. The cut points are also selected to be 35 and 65 years old. However, there is only six sample units aged over 65 years old in the smartphone location survey. As a result, the last two groups are combined and the age is finally categorized as “Millennials” (M) and “Non-millennials” (N) considering the survey time. For age imputation, the performance of random forests is slightly better than the other two and the CIT with $\alpha$=0.05 ranks second (Table 5-10).

Table 5-10. Imputation Accuracy of Naïve Model for Age Group

| | Recall | | Precision | | F1 | | Overall |
|---|---|---|---|---|---|---|---|
| | M | N | M | N | M | N | |
| CIT (α=0.05) | 0.5612 | 0.4984 | 0.4438 | 0.6164 | 0.4956 | 0.5512 | 0.5214 |
| CIT (α=0.10) | 0.5612 | 0.4841 | 0.4373 | 0.6096 | 0.4915 | 0.5397 | 0.5135 |
| **Random Forests** | **0.4757** | **0.5747** | **0.4996** | **0.6009** | **0.4873** | **0.5875** | **0.5373** |

The variable importance for the random forest classifier is ranked in Figure 5-6. For age group classification, the geographic information about home locations may play an important role, such as the road network density, the residential density, the transit accessibility, and the proportion of workers. It leads to inferences similar to Section 5.1.2 that younger people tend to live in the areas where it is convenient for them to commute. In addition, the OD distance on weekday PM peak and the variation of travel patterns at midday on weekdays should have considerable prediction strength.

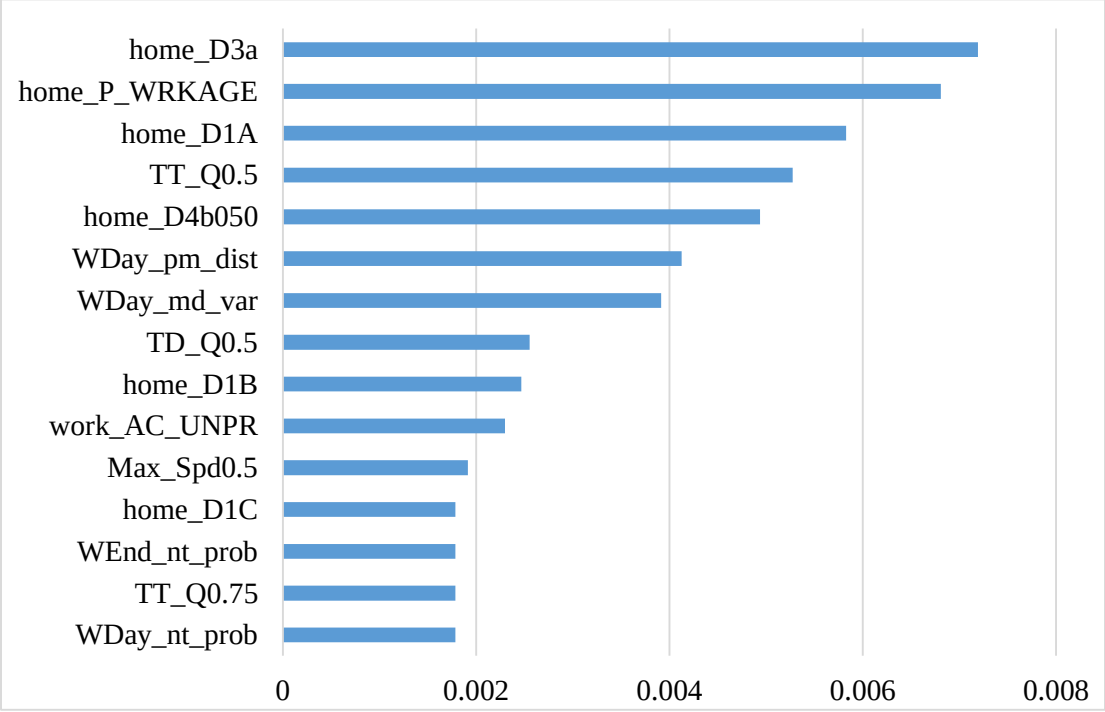


Figure 5-6. Age: Variable Importance without Weight Adjustment

5.2.3. Education Level

The education level is originally surveyed with six categories (less than high school, high school graduate, associate degree, some college, bachelor's degree, and graduate or professional degree) but later regrouped into three levels: less than bachelor's degree (LB), bachelor's degree (B), and graduate degree (G). Since the "LB" group only has nine observations, 3-fold cross-validation is employed and the classifier with weight adjustment is also evaluated (Table 5-11).

Table 5-11. Imputation Accuracy of Naïve Model for Education Level

| | Recall | | | Precision | | | F1 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | LB | B | G | LB | B | G | LB | B | G | |
| Without Weight Adjustment | | | | | | | | | | |
| CIT (α=0.05) | 0.0000 | 0.1000 | 0.8303 | 0.0000 | 0.1799 | 0.6569 | - | 0.1285 | 0.7335 | 0.5591 |
| CIT (α=0.10) | 0.0000 | 0.2778 | 0.6225 | 0.0000 | 0.2278 | 0.6551 | - | 0.2503 | 0.6384 | 0.4882 |
| **Random Forests** | **0.0000** | **0.0889** | **0.8984** | **-** | **0.2714** | **0.6290** | **-** | **0.1339** | **0.7400** | **0.5901** |
| With Weight Adjustment | | | | | | | | | | |
| **CIT (α=0.05)** | **0.3333** | **0.2778** | **0.6300** | **0.2407** | **0.3205** | **0.6218** | **0.2796** | **0.2976** | **0.6259** | **0.5029** |
| **CIT (α=0.10)** | **0.3333** | **0.3111** | **0.6192** | **0.2407** | **0.3325** | **0.6273** | **0.2796** | **0.3214** | **0.6232** | **0.5029** |
| Random Forests | 0.2500 | 0.3667 | 0.5338 | 0.0733 | 0.4111 | 0.6858 | 0.1133 | 0.3876 | 0.6003 | 0.4649 |

In Table 5-11, it can be observed that the classifiers without weight adjustment fail to categorize any instance into the "LB" group even though they generally reach higher accuracy. Among the three classifiers with weight adjustment, the overall performances of the two CIT classifiers are similar. On the other hand, the random forests classifier shows weaker imputation strength and mistakenly classifies the

instances of the “B” and “G” groups into the “LB” group according to the extreme small value of precision for the “LB” group.

The CIT with 5% significance level and weight adjustment is visualized in Figure 5-7. According to the number of instances within each class, the case weight is set as 8 for group “LB”, 2 for group “B”, and 1 for group “G”. The case weight should be considered when reading the information of terminal nodes.

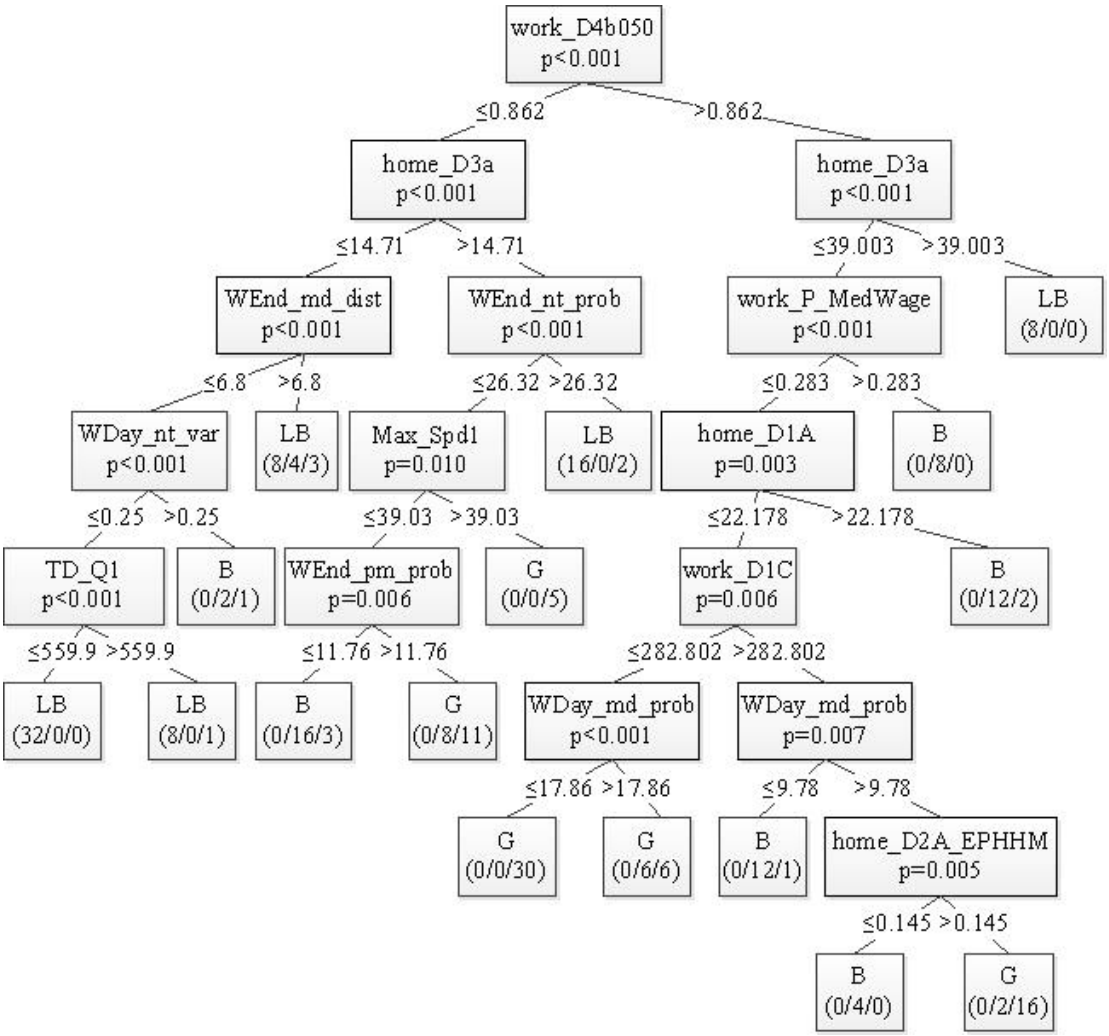


Figure 5-7. Education: The CIT with 5% Significance Level and Weight Adjustment

Due to the complexity of the tree, one terminal node for each class is illustrated where the most relevant instances are imputed correctly. For group “LB”, they either work in a CBG with higher transit accessibility and live in the areas with higher density of road network or work in the areas with lower transit accessibility and live in the areas with lower roadway density. Considering the travel behaviors, they have more variation of travel patterns at night on weekdays and they travel more at night on weekends. For group “B”, they prefer living in the areas with higher road network density and they travel less at night on weekends. For group “G”, they usually live in the areas with lower road network density and residential density, which may be suburban areas. Their work locations have lower proportion of people earning median wages and lower employment density. Travel behaviors have not contributed significantly in distinguishing people with bachelor’s degree or graduate degree, except for the proportion of trips starting at PM peak on weekends.

#### 5.2.4. Household Income Level

The household income level is originally surveyed with seven categories as described in Section 3.2.1 and later regrouped into three levels: less than $50,000 (low), $50,000-$150,000 (middle), $150,000+ (high). 3-fold cross-validation is employed to evaluate the accuracy.

In Table 5-12, the CIT classifier with 5% significance level has the best overall accuracy but has a very low recall value for high-income group. By examining the confusion matrix for one single model, it is found that a large portion of the high-income instances have been misclassified as the middle-income group. A possible

explanation is that the travel behaviors and the residential selection for the two groups may be alike.

Table 5-12. Imputation Accuracy of Naïve Model for Income Level

| | Recall | | | Precision | | | F1 | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Low | Mid | High | Low | Mid | High | Low | Mid | High | |
| Without Weight Adjustment | | | | | | | | | | |
| CIT (α=0.05) | 0.3167 | 0.7682 | 0.0714 | 0.3750 | 0.6318 | 0.1556 | 0.3434 | 0.6934 | 0.0979 | 0.5254 |
| **CIT (α=0.10)** | **0.3167** | **0.6541** | **0.1558** | **0.3194** | **0.6218** | **0.2037** | **0.3180** | **0.6375** | **0.1766** | **0.4849** |
| Random Forests | 0.0667 | 0.7377 | 0.0000 | 0.0714 | 0.5723 | 0.0000 | 0.0690 | 0.6445 | - | 0.4516 |
| With Weight Adjustment | | | | | | | | | | |
| CIT (α=0.05) | 0.1222 | 0.3937 | 0.3290 | 0.1667 | 0.5746 | 0.1882 | 0.1410 | 0.4673 | 0.2394 | 0.3619 |
| CIT (α=0.10) | 0.1222 | 0.3937 | 0.3290 | 0.1667 | 0.5746 | 0.1882 | 0.1410 | 0.4673 | 0.2394 | 0.3619 |
| **Random Forests** | **0.2500** | **0.4598** | **0.1970** | **0.0995** | **0.5791** | **0.1815** | **0.1424** | **0.5126** | **0.1889** | **0.3775** |

The CIT with 10% significance level and without weight adjustment is visualized in Figure 5-8. It can be noticed that high income group travel less at AM peak but more at midday on weekends. They also live in the areas with more workers earning high wages but lower density of employment. Similar as people with higher education, their work locations tend to have lower employment density too. The low income group travel less at midday and have smaller variation of travel patterns at AM peak on weekends. Their home locations have higher proportion of workers. The middle income group seems to take more trips than the low income group and have larger variation of travel patterns on weekend AM peak.

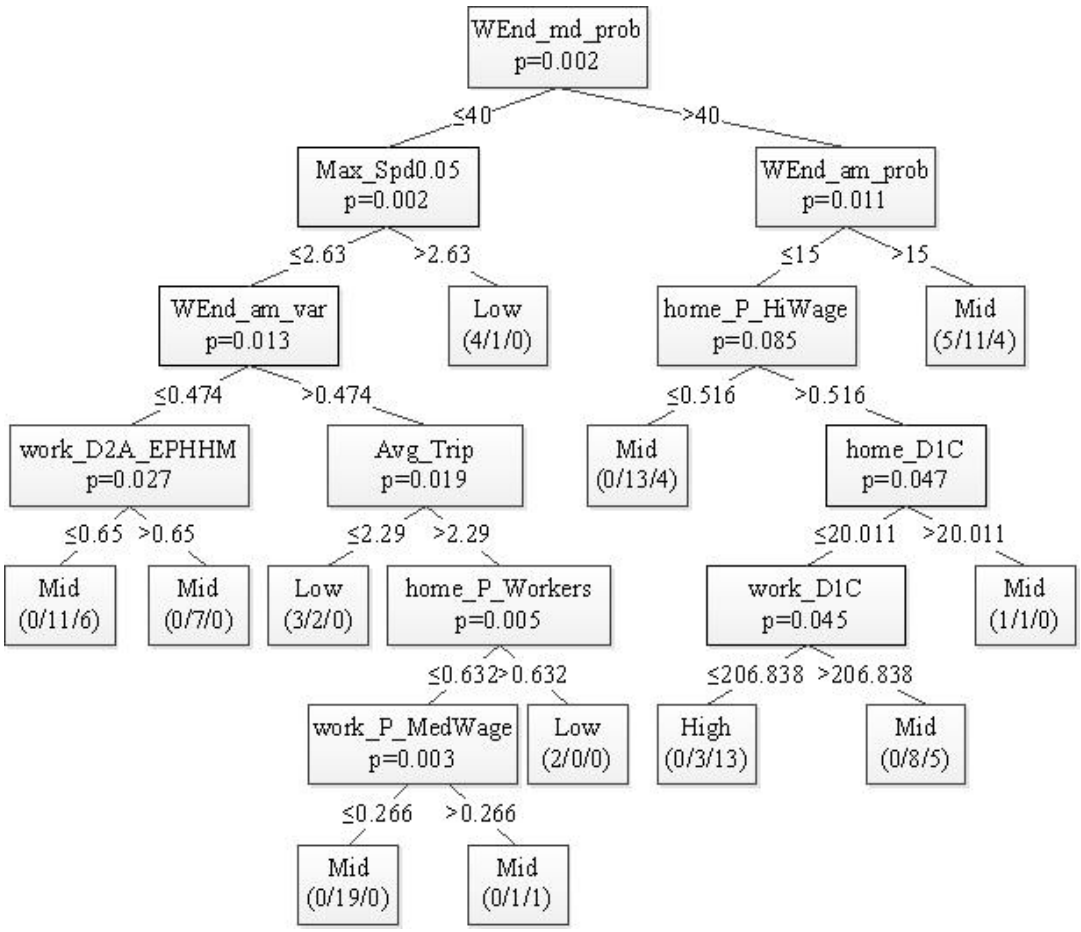


Figure 5-8. Income: The CIT with 10% Significance Level and without Weight Adjustment

*5.3. Discussion*

This section will summarize the detailed illustrations for each demographic attribute based on the two datasets. For gender imputation, the general rules are that people are more likely to be males if they travel longer, have smaller variation of travel patterns on weekend night, and live in the areas with fewer people of working age and larger CBG size. On the other hand, people who live in the areas with higher transit accessibility, higher proportion of workers, and smaller block group size are identified as females. Also, a person is recognized as female if she tends to follow a

travel routine on weekdays, travel less on weekend night, as well as work in the areas with higher employment density and higher proportion of workers earning median wages.

Those inferences are identical to the common sense. Females usually pay more attention to convenience, which can explain why the areas with more urban features are preferred. Their less activity at night on weekends may be due to the safety concerns. When it comes to the travel behaviors, males may take the lead role in long-distance driving trips.

Based on the results of age group imputation, the significant travel behavior variables include trip frequency on weekend night and travel distance. The younger people under 35 years old have higher probability to take longer trips. People are also classified into the "under 35" group if they live in the areas with more workers earning median wages and with higher density of road network. It is reasonable that people under 35 are at their early stage of career so they make the residential selection considering the convenience to commute. In contrast, people over 35 tend to live in the areas with larger proportion of high income workers but lower density of road network. It may be caused by their preferences in residential areas with better environment.

People with bachelor's degrees (B) and graduate degrees (G) share many similar characteristics. Compared to people whose education level is less than bachelor's degree (LB), they are identified as working in the areas with higher land use diversity. Besides that, the "LB" class either tends to live in the areas with lower roadway

density and work in the areas with lower transit accessibility or in the opposite way. There still exists difference between the “B” and “G” group. For instance, people with bachelor’s degree live in the areas with higher residential density while people with graduate degree in the areas with lower roadway density and lower residential density. The “G” group works in the areas with lower employment density and smaller proportion of middle income workers. It may indicate that people with higher education level prefer to live in suburban or rural areas.

The high income group shows propensity for residential selection similar to the high education group. The travel behaviors on weekends are generally key to income group classification. On weekends, the low income group travels less at midday and have smaller variation of travel patterns at AM peak period while the middle and high income groups make fewer trips at AM peak but more during midday. Overall, the middle income group also generate more trips than the low income group. The prediction power of travel behaviors on weekends may indicate that people in higher income groups have more social or recreational activities.

Though the two datasets were created with a five-year gap, there are some similar inferences drawn from the imputation results. For example, the age group classification is sensitive to road network density for both datasets. The inferences generated in both datasets also supplement each other and create more comprehensive implications regarding each demographic attribute.

In general, the POI information and imputed purpose do not improve the overall model performance significantly. Nevertheless, they help to identify the minor group correctly and thus should be still valuable for imputation.

# Chapter 6: Conclusion

## *6.1. Summary of Research*

Following the introduction on the background and objective of the research, a thorough literature review has been delivered in Chapter 2. The emergence, evolution, and three types of PCLD have been covered. The derivative studies are later summarized, including trip identification, travel mode detection, trip purpose inference, and social demographic imputation based on PCLD. The common methods for imputing missing information of PCLD are compared and evaluated too.

In Chapter 3, two datasets are introduced: one is an in-vehicle GPS dataset and another is a smartphone-based location dataset. While in-vehicle GPS devices provide more precise and accurate data, smartphones are able to capture trips of other modes, such as transit, bicycle, walking, etc. The Smart Location Database (SLD) is employed for its all-around feature set, fine geographic resolution, and wide coverage.

Chapter 4 develops different frameworks for processing raw PCLD considering the recording frequency and location accuracy. It is followed by the demonstration of the selected machine learning methods: conditional inference tree (CIT) and CIT-based random forests. Multiple types of feature sets are constructed for training the model. Since the feature selection is embedded in the CIT classifier, the feature sets comprise all the variables that may have an influence.

In Chapter 5, the imputation results for each demographic attribute and for each dataset are discussed in detail. The rules generated from CIT are visualized and the variables with higher importance are listed based on the random forest classifier. In Section 5.3, the inferences about each demographic attribute are analyzed across the datasets and no contradiction has been found.

### *6.2. Future Research*

Built upon the progress made by the thesis, several directions of future research can be explored. The remaining part of the section will provide some primary ideas and discussions on data quality, feature set enrichment, alternative imputation methods, sample recruitment and real-world application.

The datasets examined in the thesis have detailed trajectories of sample units with few trips missing. However, most large datasets of PCLD have even lower frequency of data records and lower precision for locations recorded. An analysis could be done to evaluate the feasibility of demographic imputation and the prediction strength based on different levels of knowledge owned by PCLD.

For the multimodal PCLD dataset, additional features can be considered about mode selection and the interaction terms of mode, departure time, travel distance, etc. There are also some attributes on the response side that are easier to identify, such as car ownership and household composition. They may be imputed and later serve as better intermediate variables than trip purpose.

In this study, CIT and CIT-base random forest are applied to impute the social demographics. They provide interpretable results and rules but may be inferior to other machine learning methods considering the imputation accuracy. It would be valuable to conduct a comprehensive comparison among the alternative imputation methods and summarize the advantages and disadvantages for different models.

Experience from the thesis also shed some light on the sample recruitment. The two datasets included were not originally designed for the social demographic imputation so the class imbalance problem has resulted in some limits. To develop an algorithm ready for practice, a sample is needed with balanced and comprehensive social demographic groups. Furthermore, both the sample design and the social demographic categorization should consider the application scenarios.

Beyond the research topics on the imputation process, the application of the imputed social demographics is also appealing. With imputation model established from the relatively small datasets with ground truth, the social demographics of large real-world anonymous PCLD datasets can be derived. They will then serve as the input to weighting the non-probabilistic sample of PCLD and be applied to estimate the travel behaviors for population. In addition, the imputation results can help in other fields, such as personalized location-based services and mobile advertising, which can bring more ease and convenience to daily life.

# Bibliography


[1] U.S. Department of Transportation, Federal Highway Administration, 2009 National Household Travel Survey. Retrieved from: http://nhts.ornl.gov.

[2] Metropolitan Washington Council of Governments, National Capital Region Transportation Planning Board, 2007/2008 Household Travel Survey. Retrieved from: http://www1.mwcog.org/transportation/activities/hts.

[3] U.S. Environmental Protection Agency, Office of Policy, Office of Sustainable Communities. Smart Location Database. Retrieved from: https://catalog.data.gov/dataset/smart-location-database-download.

[4] Batelle (1997). Global Positioning Systems for personal travel surveys: Lexington area travel data collection test. Final Report, Office of Highway Policy Information and Office of Technology Applications, Federal Highway Administration, Battelle Transport Division, Columbus.

[5] Yalamanchili, L., Pendyala, R., Prabaharan, N., & Chakravarthy, P. (1999). Analysis of Global Positioning System-based data collection methods for capturing multistop trip-chaining behavior. Transportation Research Record: Journal of the Transportation Research Board, (1660), 58-65.

[6] Draijer, G., Kalfs, N., & Perdok, J. (2000). Global positioning system as data collection method for travel research. Transportation Research Record: Journal of the Transportation Research Board, (1719), 147-153.

[7] Wolf, J. L. (2000). Using GPS data loggers to replace travel diaries in the collection of travel data. Doctoral dissertation, School of Civil and Environmental Engineering, Georgia Institute of Technology.

[8] Wolf, J., Guensler, R., Frank, L., & Ogle, J. (2000, July). The use of electronic travel diaries and vehicle instrumentation packages in the year 2000 Atlanta Regional Household Travel Survey: Test results, package configurations, and deployment plans. In 9th International Association of Travel Behaviour Research Conference.

[9] Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. Transportation Research Record: Journal of the Transportation Research Board, (1768), 125-134.

[10] Pearson, D. (2001) Global Positioning System (GPS) and travel surveys: Results from the 1997 Austin Household Survey. Paper presented at the Eighth Conference on the Application of Transportation Planning Methods, Corpus Christi, Texas, April 2001.

[11] Doherty, S. T., Noël, N., Gosselin, M. L., Sirois, C., & Ueno, M. (2001). Moving beyond observed outcomes: integrating global positioning systems and interactive computer-based travel behavior surveys (No. E-C026).3

[12] Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. Transport Reviews, 34(3), 316-334.

[13] Schönfelder, S., Axhausen, K. W., Antille, N., & Bierlaire, M. (2002). Exploring the potentials of automatically collected GPS data for travel behaviour analysis. Arbeitsberichte Verkehrs-und Raumplanung, 124.

[14] Bohte, W., Maat, K., & van Wee, B. (2007). Residential self-selection, the effect of travel-related attitudes and lifestyle orientation on residential location choice; evidence from the Netherlands. In 11th World Conference on Transport Research, Berkeley.

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

[16] Pasquier, M., Hofmann, U., Mende, F. H., May, M., Hecker, D., & Körner, C. (2008, May). Modelling and prospects of the audience measurement for outdoor advertising based on data collection using GPS devices (electronic passive measurement system). In Proceedings of the 8th International Conference on Survey Methods in Transport.

[17] Papinski, D., Scott, D. M., & Doherty, S. T. (2009). Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation research part F: traffic psychology and behaviour, 12(4), 347-358.

[18] Stopher, P., Clifford, E., Swann, N., & Zhang, Y. (2009). Evaluating voluntary travel behaviour change: Suggested guidelines and case studies. Transport Policy, 16(6), 315-324.

[19] Stopher, P. R., Moutou, C. J., & Liu, W. (2013). Sustainability of voluntary travel behaviour change initiatives: a 5-year study.

[20] Itsubo, S., & Hato, E. (2006). Effectiveness of household travel survey using GPS-equipped cell phones and Web diary: Comparative study with paper-based travel survey (No. 06-0701).

[21] Marchal, P., Roux, S., Yuan, S., Hubert, J. P., Armoogum, J., Madre, J. L., & Lee-Gosselin, M. E. H. (2008, May). A study of non-response in the GPS sub-sample of the French national travel survey 2007–08. In Proceedings of the 8th International Conference on Survey Methods in Transport, France (pp. 25-31).

[22] Krygsman, S. C., & Nel, J. H. (2009). The use of global positioning devices in travel surveys-a developing country application. SATC 2009.

[23] Stopher, P., & Wargelin, L. (2010, July). Conducting a household travel survey with GPS: Reports on a pilot study. In 12th World Conference on Transport Research (pp. 11-15).

[24] Oliveira, M., Vovsha, P., Wolf, J., Birotker, Y., Givon, D., & Paasche, J. (2011, January). GPS-assisted prompted recall household travel survey to support

development of advanced travel model in Jerusalem, Israel. In 90th Annual Meeting of the Transportation Research Board.

[25] Kohla, B., & Meschik, M. (2013). Comparing trip diaries with GPS tracking: Results of a comprehensive Austrian study. Transport survey methods: best practice for decision making, 305-320.

[26] Rasmussen, T. K., Ingvardson, J. B., Halldórsdóttir, K., & Nielsen, O. A. (2013, August). Using wearable GPS devices in travel surveys: A case study in the Greater Copenhagen Area. In Proceedings from the Annual Transport Conference at Aalborg University) ISSN (pp. 1603-9696).

[27] Horak, R. (2007). Telecommunications and data communications handbook (Vol. 25). New York: Wiley-Interscience.

[28] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779.

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

[30] Song, C., Koren, T., Wang, P., & Barabási, A. L. (2010). Modelling the scaling properties of human mobility. Nature Physics, 6(10), 818.

[31] Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., & Barabási, A. L. (2015). Returners and explorers dichotomy in human mobility. Nature Communications, 6, 8166.

[32] Çolak, S., Lima, A., & González, M. C. (2016). Understanding congested travel in urban areas. Nature Communications, 7, 10793.

[33] Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. Science, 328(5981), 1029-1031.

[34] Frias-Martinez, V., Virseda, J., Rubio, A., & Frias-Martinez, E. (2010, December). Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development (p. 11). ACM.

[35] Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011, July). Prediction of socioeconomic levels using cell phone records. In International Conference on User Modeling, Adaptation, and Personalization (pp. 377-388). Springer, Berlin, Heidelberg.

[36] Zhao, Y. (2000). Mobile phone location determination and its impact on intelligent transportation systems. IEEE Transactions on intelligent transportation systems, 1(1), 55-64.

[37] Zhou, C., Ludford, P., Frankowski, D., & Terveen, L. (2005, April). An experiment in discovering personally meaningful places from location data. In CHI'05 Extended Abstracts on Human Factors in Computing Systems (pp. 2029-2032). ACM.

[38] Byon, Y. J., Abdulhai, B., & Shalaby, A. S. (2007). Impact of sampling rate of GPS-enabled cell phones on mode detection and GIS map matching performance (No. 07-1795).

[39] Amin, S., Andrews, S., Apte, S., Arnold, J., Ban, J., Benko, M., ... & Dodson, T. (2008). Mobile century using GPS mobile phones as traffic sensors: A field experiment. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.152.8548&rep=rep1&type=pdf.

[40] Wiehe, S. E., Carroll, A. E., Liu, G. C., Haberkorn, K. L., Hoch, S. C., Wilson, J. S., & Fortenberry, J. (2008). Using GPS-enabled cell phones to track the travel patterns of adolescents. International journal of health geographics, 7(1), 22.

[41] Work, D. B., Tossavainen, O. P., Jacobson, Q., & Bayen, A. M. (2009, June). Lagrangian sensing: traffic estimation with mobile devices. In American Control Conference, 2009. ACC'09. (pp. 1536-1543). IEEE.

[42] Guido, G., Vitale, A., Astarita, V., Saccomanno, F., Giofré, V. P., & Gallelli, V. (2012). Estimation of safety performance measures from smartphone sensors. Procedia-Social and Behavioral Sciences, 54, 1095-1103.

[43] 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. Transportation Research Record: Journal of the Transportation Research Board, (2354), 59-67.

[44] Leber, J. (2013). How wireless carriers are monetizing your movements. MIT Technology Rev. Retrieved from https://www.technologyreview.com/s/513016/how-wireless-carriers-are-monetizing-your-movements.

[45] BITRE. (2014). New traffic data sources – An overview. Retrieved from https://bitre.gov.au/events/2014/files/NewDataSources-BackgroundPaper-April%202014.pdf.

[46] Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. GeoJournal, 72(3-4), 137-148.

[47] De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., & Yu, C. (2010, June). Automatic construction of travel itineraries using social breadcrumbs. In Proceedings of the 21st ACM conference on Hypertext and hypermedia (pp. 35-44). ACM.

[48] Sui, D.Z. and Goodchild, M.F. 2001. Are GIS becoming new media?. International Journal of Geographical Information Science, 15(5): 387–390.

[49] Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. International Journal of Geographical Information Science, 25(11), 1737-1748.

[50] Naaman, M. (2011). Geographic information from georeferenced social media data. SIGSPATIAL Special, 3(2), 54-61.

[51] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

[52] Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., & Xie, X. (2015, February). You are where you go: Inferring demographic attributes from location check-ins. In Proceedings of the eighth ACM international conference on web search and data mining (pp. 295-304). ACM.

[53] Riederer, C. J., Zimmeck, S., Phanord, C., Chaintreau, A., & Bellovin, S. M. (2015, November). I don't have a photograph, but you can have my footprints.: Revealing the Demographics of Location Data. In Proceedings of the 2015 ACM on Conference on Online Social Networks (pp. 185-195). ACM.

[54] Bao, J., Lian, D., Zhang, F., & Yuan, N. J. (2016). Geo-social media data analytic for user modeling and location-based services. SIGSPATIAL Special, 7(3), 11-18.

[55] Liu, Q., Wu, S., Wang, L., & Tan, T. (2016, February). Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In AAAI (pp. 194-200).

[56] Riederer, C., Kim, Y., Chaintreau, A., Korula, N., & Lattanzi, S. (2016, April). Linking users across domains with location data: Theory and validation. In Proceedings of the 25th International Conference on World Wide Web (pp. 707-719). International World Wide Web Conferences Steering Committee.

[57] Gu, Y., Yao, Y., Liu, W., & Song, J. (2016, August). We know where you are: Home location identification in location-based social networks. In Computer Communication and Networks (ICCCN), 2016 25th International Conference on (pp. 1-9). IEEE.

[58] Liao, Y., Lam, W., Jameel, S., Schockaert, S., & Xie, X. (2016, September). Who Wants to Join Me?: Companion Recommendation in Location Based Social Networks. In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (pp. 271-280). ACM.

[59] Troped, P. J., Oliveira, M. S., Matthews, C. E., Cromley, E. K., Melly, S. J., & Craig, B. A. (2008). Prediction of activity mode with global positioning system and accelerometer data. Medicine and science in sports and exercise, 40(5), 972-978.

[60] Gilbert, E., & Karahalios, K. (2009, April). Predicting tie strength with social media. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 211-220). ACM.

[61] Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011, July). Prediction of socioeconomic levels using cell phone records. In International

Conference on User Modeling, Adaptation, and Personalization (pp. 377-388). Springer, Berlin, Heidelberg.

[62] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3, 1376.

[63] Zhang, S., Yin, D., Zhang, Y., & Zhou, W. (2015). Computing on base station behavior using Erlang measurement and call detail record. IEEE transactions on emerging topics in computing, 3(3), 444-453.

[64] Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from GPS data: a literature review on the existing methodologies. Procedia-Social and Behavioral Sciences, 138, 557-565.

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

[66] Tsui, S., & Shalaby, A. (2006). Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. Transportation Research Record: Journal of the Transportation Research Board, (1972), 38-45.

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

[68] Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L., & Perez, R. (2010). Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. IET Intelligent Transport Systems, 4(1), 37-49.

[69] Zhang, L., Dalyot, S., Eggert, D., & Sester, M. (2011). Multi-stage approach to travel-mode segmentation and classification of GPS traces. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: [Geospatial Data Infrastructure: From Data Acquisition And Updating To Smarter Services] 38-4 (2011), Nr. W25, 38(W25), 87-93.

[70] Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems, 36(2), 131-139.

[71] Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P. (2014). Supporting large-scale travel surveys with smartphones–A practical approach. Transportation Research Part C: Emerging Technologies, 43, 212-221.

[72] Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., & Axhausen, K. (2004). Eighty weeks of global positioning system traces: approaches to enriching trip information. Transportation Research Record: Journal of the Transportation Research Board, (1870), 46-54.

[73] Stopher, P., Clifford, E., Zhang, J., & FitzGerald, C. (2008). Deducing mode and purpose from GPS data. Institute of Transport and Logistics Studies, 1-13.

[74] Elango, V., & Guensler, R. (2010, May). An automated activity identification method for passively collected GPS data. In 3rd Conference on Innovations in Travel Modeling. Phoenix, AZ.

[75] Huang, L., Li, Q., & Yue, Y. (2010, November). Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks (pp. 27-30). ACM.

[76] Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Applications, 40(8), 3299-3311.

[77] Shen, L., & Stopher, P. R. (2013). A process for trip purpose imputation from Global Positioning System data. Transportation Research Part C: Emerging Technologies, 36, 261-267.

[78] Kim, Y., Pereira, F. C., Zhao, F., Ghorpade, A., Zegras, P. C., & Ben-Akiva, M. (2014, August). Activity recognition for a smartphone based travel survey based on cross-user history data. In Pattern Recognition (ICPR), 2014 22nd International Conference on (pp. 432-437). IEEE.s

[79] Oliveira, M., Vovsha, P., Wolf, J., & Mitchell, M. (2014). Evaluation of two methods for identifying trip purpose in GPS-based household travel surveys. Transportation Research Record: Journal of the Transportation Research Board, (2405), 33-41.

[80] Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., & Das, K. (2017). Real-time trip purpose prediction using online location-based search and discovery services. Transportation Research Part C: Emerging Technologies, 77, 96-112.

[81] Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. (2014). Trip purpose identification from GPS tracks. Transportation Research Record: Journal of the Transportation Research Board, (2405), 16-23.

[82] Lu, Y., & Zhang, L. (2015). Imputing trip purposes for long-distance travel. Transportation, 42(4), 581-595.

[83] Xiao, G., Juan, Z., & Zhang, C. (2016). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transportation Research Part C: Emerging Technologies, 71, 447-463.

[84] Gong, L., Kanamori, R., & Yamamoto, T. (2017). Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. Travel Behaviour and Society.

[85] Lu, X., & Pas, E. I. (1999). Socio-demographics, activity participation and travel behavior. Transportation Research Part A: policy and practice, 33(1), 1-18.

[86] Altshuler, Y., Aharony, N., Fire, M., Elovici, Y., & Pentland, A. (2012, September). Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. In Privacy, Security, Risk and

Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom) (pp. 969-974). IEEE.

[87] Auld, J., Mohammadian, A., Simas Oliveira, M., Wolf, J., & Bachman, W. (2015). Demographic characterization of anonymous trace travel data. Transportation Research Record: Journal of the Transportation Research Board, (2526), 19-28.

[88] Roy, A., & Pebesma, E. (2017, April). A Machine Learning Approach to Demographic Prediction using Geohashes. In Proceedings of the 2nd International Workshop on Social Sensing (pp. 15-20). ACM.

[89] Feng, T., & Timmermans, H. J. (2016). Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. Transportation Planning and Technology, 39(2), 180-194.

[90] Pourret, O., Naïm, P., & Marcot, B. (Eds.). (2008). Bayesian networks: a practical guide to applications (Vol. 73). John Wiley & Sons.

[91] Wen, C. H., & Koppelman, F. S. (2001). The generalized nested logit model. Transportation Research Part B: Methodological, 35(7), 627-641.

[92] Boyd, J. H., & Mellman, R. E. (1980). The effect of fuel economy standards on the US automotive market: an hedonic demand analysis. Transportation Research Part A: General, 14(5-6), 367-378.

[93] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

[94] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[95] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[96] Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

[97] Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), 651-674.

[98] 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(1), 25.

[99] Krause, C. (2015). Short Term Travel Behavior Prediction Through GPS and Land Use Data (Doctoral dissertation). Retrieved from: https://drum.lib.umd.edu/handle/1903/17333.

[100] Tang, L., Pan, Y., & Zhang, L. (2018). Trip Purpose Imputation Based on Long Term GPS Data (No. 18-05010). Retrieved from:

http://amonline.trb.org/2017trb-1.3983622/t005-1.4000488/241-1.4001221/18-05010-1.3991828/18-05010-1.4001228.

[101] Ramsey, K., & Thomas, J. (2012). EPA's Smart location database: A National dataset for characterizing location sustainability and urban form. Retrieved from: https://metrocouncil.onlinegroups.net/groups/research/files/f/26322-2012-02-28T213936Z/SLD_v02_report.pdf.

[102] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[103] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[104] 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(1), 25.

[105] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. BMC bioinformatics, 9(1), 307.

[106] Janitza, S., Strobl, C., & Boulesteix, A. L. (2013). An AUC-based permutation variable importance measure for random forests. BMC bioinformatics, 14(1), 119.