Automating and Optimizing a Transportation Mode Classification Model for use on Smartphone Data

by

Akram Omar Nour

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2015

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Portions of this thesis have been previously published in a conference preceding, submitted for journal publication, or presented at conference. The following provides a listing of the sections of the thesis that have been previously published/submitted for publication in whole or in part and the citation for the publication.

Submitted for Journal Publications

- Content of Chapter 6

  Nour, A., B. Hellinga, and J. Casello, "Transportation Mode Classification based on Smartphone data: Enhancing Accuracy Using Spatial Statistics and GIS", (Submitted in March 2015), Journal of Transport Geography. Under Revision.

Peer Reviewed Conference Proceedings and Presentations

- Content of Chapter 4 (Section 4.2) and Chapter 5

  Nour, A., J Casello, and B. Hellinga, "Developing and Optimizing a Transportation Mode Inference Model Utilizing Data from GPS Embedded Smartphones", 2015, 94th Annual Meeting of the Transportation Research Board, Washington, D.C., January 11-15.

Abstract

As transportation engineering and planning evolve from "data poor" to "data rich" practices, methods to automate the collection and translation of data to information become increasingly important. Advances in wireless communications and technologies provide the opportunity to collect detailed data on travel trajectories using smartphones equipped with GPS and accelerometers. These types of smartphones are ubiquitous and, as such, present an opportunity to conveniently collect spatial and temporal data at regular time intervals. This can be useful to utilize as a method to document trip attributes of interest, namely origin, destination, departure time, route choice, trip purpose, and mode choice. Though some of these attributes can be relatively easily extracted from the smartphone data, inferring transportation mode(s) used by the trip maker remains a challenging problem.

This research presents a data-driven classification model to infer the transportation mode(s) used by trip makers on the basis of data collected with GPS equipped smart phones. Rather than making *a priori* assumptions, we instead employ an optimization method to objectively produce the following classifier components and methods: a ranked feature vector based on the power of differentiation between different modes; the classification technique between the range of candidate classifiers; the number of ranked attributes to include in the feature vector; data formatting; and optimal model parameters. The model is trained and tested using labelled trip data. The calibrated model is evaluated by testing its ability to classify travel mode correctly for GPS data at a different level of disaggregation than the one used in the model training step. The model provides an accuracy of approximately 86% at the disaggregated level (*e.g.,* Walk, Bike, Transit, and Private Automobile) and approximately 94% at aggregated level (*e.g.,* Non-Motorized and Motorized.)

The results obtained from the optimized model are supplemented with a GIS based model to improve the identification of transit trips. The method employed integrates GIS data such as the locations of transit stops and signalized intersections with observed travel patterns from the GPS embedded smartphone data. The combination of these two data sources generates new classification features that, when applied to the collected data, demonstrate that this technique vastly improves the accuracy of the classification model for identifying transit mode usage.

Acknowledgment

It is a great pleasure to acknowledge the wonderful people in my life whose scientific and moral support made this work possible. First, I would like to express my deepest love and gratitude for my two loving parents, Mr. Omar Noor and Mrs. Sabah Baqasi, who taught me nothing is impossible with a positive attitude, perseverance, and faith. Their support, advice, and guidance have been the key factors in getting me this far.

I am deeply indebted to my advisors, Prof. Bruce Hellinga and Prof. Jeff Casello, for their invariable support, guidance, and encouragement throughout my program and this research. Their comprehensive knowledge, dedication, and personal integrity helped me set a higher standard for myself not only on an academic and professorial level, but also on a personal level. This thesis could not have been completed without their extremely insightful observations, knowledge, sincerity and critical attention that motivated me to do my best.

I am sincerely grateful to all of my advisory committee members: Prof. Darren Scott from McMaster University, Prof. Jean Andrey, Prof. Carl Hass, and Prof. Liping Fu for their time, valuable comments, and critical remarks towards the improvement of this research.

I would also like to extend my sincere appreciation to the faculty and staff of the Department of Civil and Environmental Engineering, especially Prof. Frank Saccomanno, Prof. Liping Fu, Prof. Carl Haas, and Prof. Susan Tighe for their advice and support during my study in UW. They provided me with an incredibly enlightening and humbling experience.

I have also been privileged to have gained a very talented group of friends and colleagues during my study in the University of Waterloo. I would like to especially express my utmost thanks to my best friends Soroush Salek and Reza Noroozi. I wish to also thank Dr. Babak Mehran, Amir Zarinbal, Mazen Melebary, Roshanak Taghipour, Ehsan Bagheri, Ibrahim Almohanna, Mohab Elhakim, Amin Hamdi, Usama Shahda, Amir Ghods, Kevin Yeung, Ian Dunlop, and all of my colleagues in UW-TSRG for their kind support and great friendship.

My sincerest thanks and gratitude goes to Dr. Nabeel Koshak for his endless invaluable support and guidance; words cannot express how thankful and grateful I am.

Finally, the greatest appreciation goes to my wife Reham and my sons Omar, Ahmad, and Yousif. Without their patience and sacrifice, I would not have been able to pursue my aspirations and complete this work; thank you for being by my side.

Dedication

This dissertation is dedicated to my amazing family.

My parents:

Omar and Sabah

My brothers:

Ahmad and Anas

My sisters:

Samah and Samar

The light of my life my lovely wife:

Reham

My three angels:

Omar, Ahmad, and Yousif

I love you all!

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Transportation systems are essential components of human societies. They facilitate the mobility and accessibility of students to schools, workers to employment, clients to businesses, etc. The continued growth in population and economy has placed growing pressure on transportation systems with a significant increase in travel demand. This leads to additional vehicles, trips, and traffic congestion on road networks. Traffic congestion is one of the major transportation challenges that negatively impacts society through increasing delays, fuel consumption, and air pollution. In Canada, the delay, wasted fuel, and extra Green-House-Gas (GHG) emissions caused by congestion on average cost the country over $5 billion, $0.32 billion, and $0.07 billion per year, respectively (in 2000 dollar value) (iTrans, 2006).

Transportation agencies recognize that it is not feasible to build their way out of congestion by constructing new roads or expanding current networks due to limited public funding (Yack, 2010). Furthermore, there is increasing concern about the sustainability of an auto-centric transportation system due to: rising costs of providing and maintaining road infrastructure; increasing air pollution and associated adverse health impacts; growing levels of child and adult obesity; and improving and understanding of the societal costs of automobile collisions. Therefore, there is increased interest to increase the efficiency of the existing transportation infrastructure and to change travel behavior by enhancing the attractiveness of higher efficiency and lower impact transportation modes such as public transit, cycling, and walking.

1.1 Travel Behavior

The performance of the transportation network is the result of an interaction between the supply of capacity and travel demand. Supply is a function of the transportation infrastructure (*e.g.*, roads, public transport routes, etc.). Demand is the aggregation of travel decisions made by individuals. For an individual making a trip, the sequence of travel decisions such as destinations, modes, and paths is called travel behavior. The result of travel behavior is often a series of connected segments called a “trip”. An illustration of segments comprising a hypothetical trip is shown in (Figure 1-1).

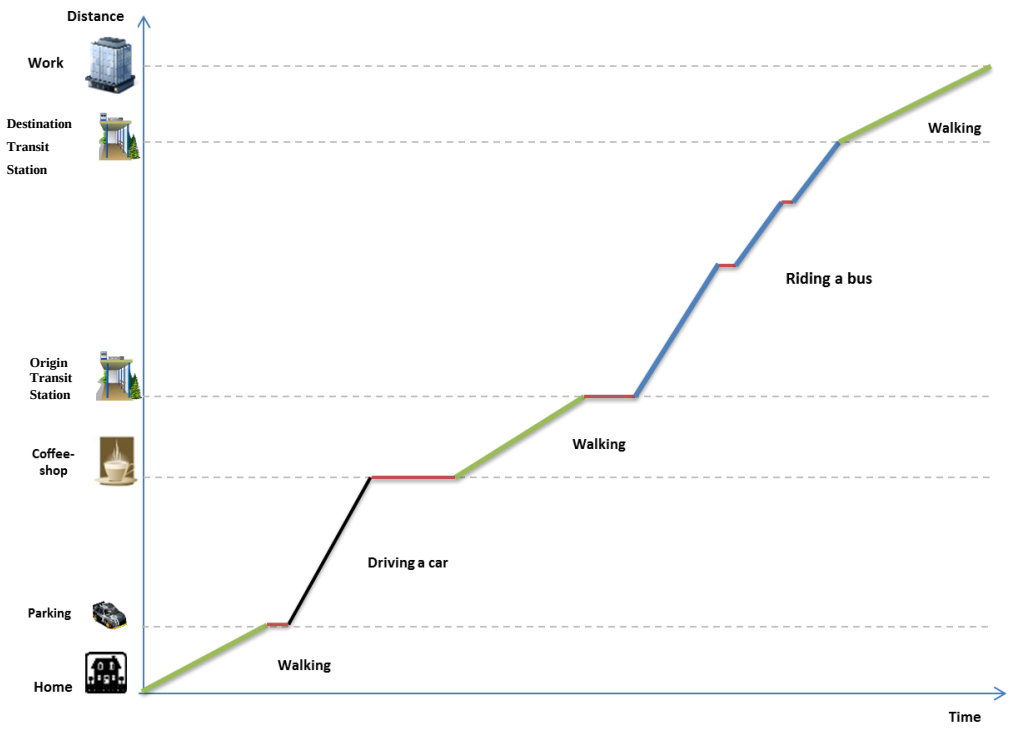


Figure 1-1 A space - time diagram of disaggregated travel behavior data

Generally, the decisions for these choices are made on the basis of a complex behavioral process which is influenced by different factors such as a trip maker's needs, socioeconomic standing, and demographic characteristics. Transportation planners and engineers try to model this decision process in order to develop tools that can be used to evaluate different transportation infrastructure investments or policies.

Widely used travel forecasting models are based on the framework of the multinomial logit (MNL) model shown in Equation 1-1 developed by McFadden (McFadden *et al.*, 1977) and random utility theory (RUT). Utility theory postulates that a decision maker chooses a transportation mode, departure time, route, etc. from a choice set such that the alternative chosen maximizes/minimizes their perceived utility/generalized cost. The utility function of each alternative used within conventional travel forecasting models for a given trip is impacted by several factors including trip maker socioeconomics (income, cars ownership, age, etc.) and alternatives' attributes (in-vehicle time, access time, out of pocket cost, etc.). Mathematically, the logit model is written as

$$P_{im} = \frac{e^{U_m}}{\sum_{m=1}^{M} e^{U_m}} \tag{1-1}$$

where

P_{im} Probability of trip maker *i* choosing alternative *m*

U_m The utility function associated with choosing alternative *m*

M Set of transportation mode alternatives available to the trip maker

Generally, the utility function has a linear form combining all variables considered as mode attributes and individual characteristics (de Dios Ortúzar & Willumsen, 2001) (see Equation 1-2). Variables included in the utility function have different units. Therefore in order to estimate the relative importance of each variable included in the utility function to others, all the variables are converted to a common unit. This conversion process results in a linear function of the summation of all converted variables into a so-called generalized cost (GC) function (Kumar *et al.*, 2004). This generalized cost function replaces the utility function in Equation 1-2.

$$U_m = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_b X_b + \alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_2 \ldots\ldots + \alpha_a Y\_a \quad (1\text{-}2)$$

where

U_m The utility function associated with choosing alternative transportation mode *m*

X_1, X_2, X_b Alternative attributes

Y_1, Y_2, Y_a Individual's characteristics

$\beta_1, \beta_2, \beta_b$ Parameters for alternative's attributes

α_1, α_2, a_a Parameters for individual's characteristics

As a consequence, it can be stated that the likelihood of a trip maker choosing a particular mode (*e.g.*, transit) is based on the generalized cost (GC) of using that mode relative to the generalized cost of all other available modes. Typically, GC functions include in-vehicle time, out of vehicle time, out of pocket cost, transfer penalties, and a mode-specific constant (bias). The mode-specific constant is supposed to represent the factors that are difficult to quantify such as comfort of ride, reliability, etc.

These formulations of mode choice make two important assumptions:

1. Trip makers are able to accurately determine the trip attributes associated with the mode that they have chosen to use; and

2. Trip makers are able to accurately determine the trip attributes associated with modes that they have not used, but which are available to them for their trip.

However, there is substantial evidence in the literature that both of these assumptions are incorrect. Substantial research in transportation planning suggests that, when evaluating transportation alternatives to fulfill their travel needs, many travelers ignore alternative modes due to lack of information about the availability of the service and wrong perceptions about service quality (Brög, 1998; Lucas *et al.*, 2011; Stopher *et al.*, 2009). In order to improve travel forecasting models and the way in which transportation infrastructure investments are evaluated and made, improved traveler behavior data are required.

1.2 Travel behavior data collection

Measuring travel behavior has typically been done through the use of travel diaries in which the survey participant is required to record details of each trip that they make in a period of time such as day, week, etc. Survey participants are asked to log spatial and temporal trip attributes and other characteristics such as trip origin and destination, trip purpose, departure and arrival time, trip duration, transportation mode, etc. These diaries, which can be paper-based or web-based, may be supplemented with other survey data collected via face-to-face interviews, mail-back surveys, web-based surveys, or telephone surveys.

There are several challenges associated with collecting detailed travel behavior data.

1. The distribution, collection, data logging, and analysis of trip diary surveys are resource intensive.
2. It has been observed that the data collected through conventional travel survey techniques suffer from poor accuracy (Ettema *et al.*, 1996; Stopher & Collins, 2005.) Participants tend to under-report short activity stops as well as trips that do not end or start at home. In addition, auto users tend to underestimate travel time and public transportation users tend to overestimate travel time (Ettema *et al.*, 1996; Stopher & Collins, 2005).
3. Survey participation rates substantially decrease as the effort required in completing the survey increases (Doherty *et al.*, 2001).

As a result of these issues, there has been interest in using technologies that would automate the process of collecting and analyzing travel behavior data. Global Positioning System (GPS) have been proven to be a potential technology to replace conventional data collection techniques (Casas & Arce, 1999; Wolf *et al.*, 2001).

Global Positioning System is a satellite-based positioning system maintained by the United States government. The system consists of 24 satellites that orbit the earth in 12 hours and communicate with GPS ground-based receivers. These satellites send radio signals to earth that carry the satellites' information. When a GPS ground-based unit (receiver) receives these signals from at least three satellites, the position of the GPS unit can be determined by processing the satellites' locations and distance to each satellite estimated from communication time.

GPS devices, receivers, are enabled to record position data (latitude and longitude) in addition to speed, heading, and elevation periodically based on the device's settings. The data can be recorded as often as every second when the GPS antenna is receiving clear signals from a sufficient number of satellites. The accuracy of GPS location measurements typically range between 3 and 50 meters.

GPS data are subject to a number of sources of error that affect the quality of the obtained data as will be discussed in the next chapter; the most significant are:

1. Loss of data when the GPS signals are obstructed (*e.g.*, when the receiver is in a tunnel or building, etc.); and
2. Signal multipathing which occurs when there is not a direct line of sight to the GPS satellite and signals reflect off of nearby buildings. This typically occurs when the receiver is located near tall buildings, such as is often the case in downtown urban areas.

Despite these accuracy problems, in most situation GPS provides significantly more accurate spatiotemporal data for travel behavior compared to paper and phone interview surveys (Casas and Arce, 1999). Recently, researchers have carried out travel behavior studies in which participants are provided a GPS data logger or an in-vehicle GPS unit which records their spatiotemporal data at pre-set time intervals.

With the advances in wireless communications and technologies during the last five years, most smartphones are equipped with GPS and are available at affordable prices to average consumers due to the mass production of this technology (Barbeau *et al.*, 2009). This has created an opportunity for researchers to use GPS embedded smartphones to collect travel behavior data. GPS-embedded smartphones are able to determine users' location with higher accuracy compared to regular GPS loggers. The accuracy of the provided data has proven to be in the range of 3-30 meters in position requests and within 0.2 meters/sec in speed (Witte & Wilson, 2004). GPS-embedded smartphones use an Assisted GPS (A-GPS) system that improves the start-up performance or what is called time-to-first-fix[1] (TTFF) of GPS. In addition to radio signals from satellites, A-GPS utilizes cellular network resources in poor signal conditions to increase the accuracy of acquired spatiotemporal data.

However, the challenge with this approach is that the collected GPS data do not directly provide travel behavior; they provide only positions as a function of time. Models are required to transform these data into travel behavior data including trips, trip origin and destination, intermediate and main activity stops, transportation modes, trip purpose, etc. Figure 1-2 shows the data transformation to produce valuable travel information from raw GPS data. The focus of this thesis - transportation mode inference - is shown in bold.

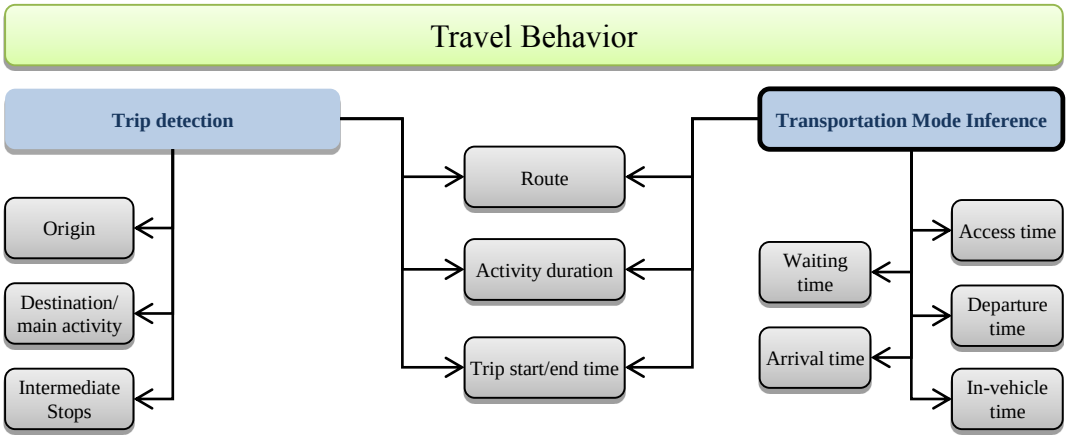


Figure 1-2 Travel behavior information to infer from raw smartphone data

[1] TTFF is the time required for a GPS receiver to acquire satellite signals and provide position information for the unit.

1.3 Research Goals and Objectives

The goal of this research is to develop an automated method to infer a traveler's transportation mode from data collected via GPS equipped smartphones. The first step in the process is to develop a method to transmit the data gathered on smartphones to a server for processing. This work has been completed by many other researchers, including Taghipour (2012). The next step in the work is to apply methods to first eliminate / correct any GPS-related errors – erroneous x, y or z values – as a result of lost communications. When calibrating and validating mode inference models, it is necessary to have labelled data (i.e. in addition to the GPS data, we must also know what mode was actually used). It is typical for the labelling to be done by the trip maker while they are making their trip (using the smartphone) or after the trip has been completed (using a web interface). Regardless of the method used, these labels typically contain errors. For example, some users identified a switch from auto to pedestrian modes several seconds (or in some cases minutes) prior to actually making that transition. As such, a number of points in the data set are labeled as walking with speeds that can only be achieved by motorized travel. To correct these errors, a systematic method was developed to identify the actual mode transfer point – defined to coincide with at least one point with speeds that were nearly zero.

With valid trip data, an optimization approach is applied to determine the types and combination of mathematical techniques that produce the best results in terms of correctly identifying transportation modes. Finally, shortcomings in the developed model's performance are addressed with the integration of additional spatial statistics and GIS data. The outcome of the work is a robust self-optimizing model that can convert raw GPS data, passively collected through smartphones, to a set of trips with highly accurate transportation mode labels. The overall research framework is illustrated in Figure 1-3.

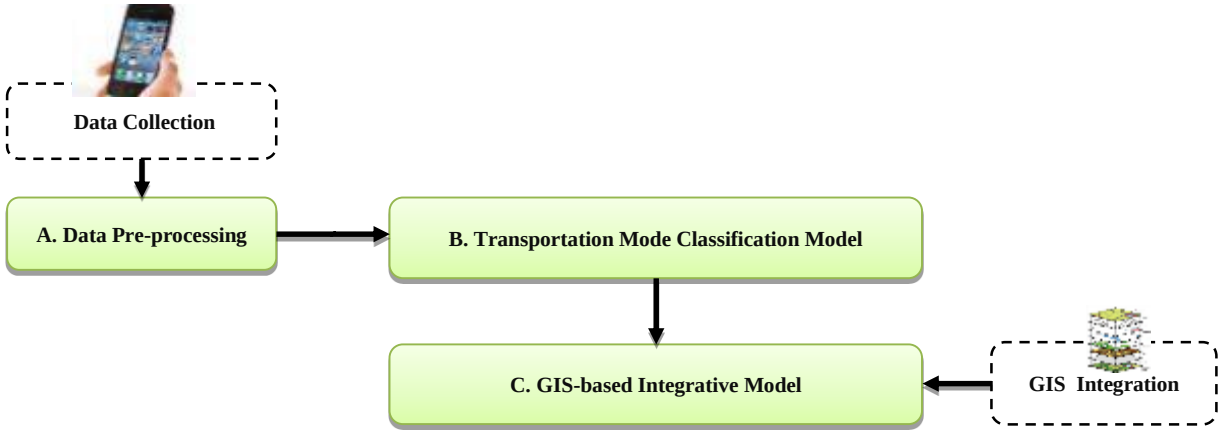


Figure 1-3 Conceptual research framework

On the basis of these objectives, the following contributions are made:

A-1. Develop a custom software tool, including a Graphical User Interface (GUI), which can be used to view and manipulate the spatiotemporal attributes of trips including space-time diagrams, speed-time diagrams, and labelled versus model predictions of mode. This tool enables the analysis to:

a. understand specific type of behavior associated with each of the different transportation modes;

b. comprehend the behavior of the developed model and algorithms;

c. identify, investigate and tackle any weaknesses associated with the developed model; and

d. identify and justify the chosen criteria in the developed algorithms.

A-2. Develop an algorithm to automatically filter the feasible trips based on the data collected via smart phones.

A-3. Develop an automated method for identifying errors in labelled data and in particular errors in the identifying of Mode Transfer Points (*MTP*), the points where users change from one to another transportation mode.

A-4. Develop an algorithm to automatically correct suspicious *MTPs* by shifting the mode transfer label forward or backward along the time axis. Having correctly labelled *MTPs* is essential for accurately evaluating the performance of the mode inference model.

B. Develop and evaluate an optimized classification model by which transportation mode can be identified automatically based on the available data collected via smart phones. It must be noted, that these methods must be configured so that they do not require the trip maker to record or identify trip attributes.

C. Develop and evaluate a GIS-based method to enhance the performance of the transportation mode classifier, particularly for identifying transit modes.

1.4 Dissertation Outline

The reminder of this dissertation is organized as follows:

Chapter 2 provides a review of the relevant literature, including (1) a brief overview of existing techniques in collecting travel behavior via GPS; (2) an overview of existing algorithms that analyze GPS data to identify travel behavior; and (3) existing trip planner applications. Chapter 3 reviews the data collection instrument, web application, and the instruction given to the participants. Chapter 4 discusses the errors associated with the collected GPS data and highlight the importance of exploring and understanding the nature of the data to the overall goal of developing a highly accurate classification model. The general statistics about the data are presented later in this chapter.

Chapter 5 addresses the challenge of identifying the transportation modes used for trips (or portions of trips) solely on the basis of the GPS data. In this chapter, an objective optimization method is introduced based on multiple regression and factorial analysis to investigate an exhaustive combination of factors to calibrate the classification model. In addition, a method in which the features are ranked and selected based on the differentiation power between different classes is introduced. At the end of the chapter, the optimized model is evaluated and the results are presented. Chapter 6 introduces and demonstrates a GIS based method to improve upon the model's performance for transit trips. The evaluation results of the developed GIS model are shown also at the end of the chapter.

Chapter 7 summarizes the research conclusions and contributions of this research and provides the recommendations for further studies.

A more detailed framework of the research to be conducted is shown in Figure 1-4.

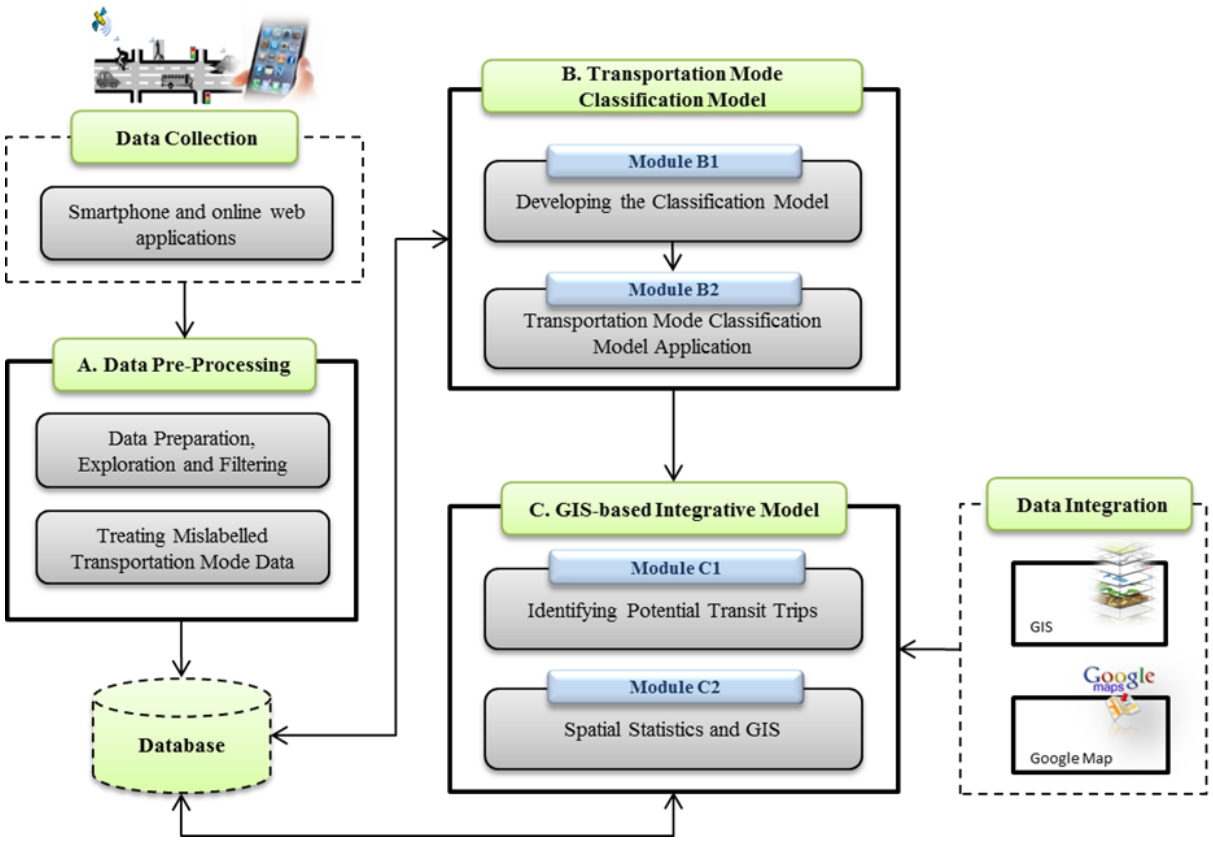


Figure 1-4 Proposed Framework for Transportation Mode Classification Model

Chapter 2

Literature Review

As discussed in the introduction, there are significant opportunities to enhance the way in which traveler behavior data can be gathered and interpreted with an ultimate goal of enhancing transportation planning and engineering. The use of GPS-embedded smartphones has the potential to generate significant traveler behavior data that can inform infrastructure investment and operational decision-making. These passive, automated methods significantly reduce the need for user input and improve accuracy compared to previous methods.

This chapter provides a review of the most relevant literature on inferring travel behaviour using GPS data. The review is divided into two main sections:

1. Use of GPS data for collecting travel behaviour data.
2. Techniques for inferring travel behaviour from GPS data.

2.1 Travel behaviour data collection using GPS

GPS has been utilized in different studies as the next-generation travel survey technique that involves less respondent burden and exhibits excellent results (Murakami & Wagner, 1999). Lee-Gosselin (2002) categorized the implementation of GPS travel survey into three general categories as follows:

1- Imitating conventional approach;

2- Passive approach;

3- Hybrid approach.

The first approach is basically to imitate traditional travel surveys. Participants are provided with either handheld GPS loggers (*e.g.*, Draijer *et al.*, 2000) or in-vehicle GPS units along with a handheld computer *e.g.*, personal digital assistant. From the GPS data, a trip's spatiotemporal details such as departure time, origin, travel time, destination, etc. are extracted with high accuracy. On the other hand, trip characteristics *e.g.*, trip purpose and mode choice can be collected through handheld computers. The two sources of information in this approach are

valuable; however, it is insufficient for long term data collection effort due to burden on participants in acquiring entries for daily trip characteristics.

The second approach is to passively collect data though GPS units. The central idea of this approach is to collect travel data without interacting with the study's participants. This approach is very suitable from the participants' perspective in long term data collection; however, inferring travel behaviour and trip characteristics from raw GPS data are challenging.

The third approach is considered a combination between the first and second approaches, a hybrid. GPS data are collected through either handheld or in-vehicle GPS units; then, participants are asked to review their GPS traces and annotate their travel components with types of activity stops, trip purpose, modes chosen, etc.

2.1.1 GPS based data collection study cases

Individual travel behaviour survey has gained increasing attention in transportation research since the first GPS studies were conducted in the mid-1990s (Casas & Arce, 1999; Murakami & Wagner, 1997; Wolf *et al.*, 2001; Yalamanchili *et al.*, 1999) to investigate the use of automated travel diaries with GPS. In 1996, the U.S. Federal Highway Administrative (FHWA) sponsored a study in Lexington area to test the feasibility of GPS for personal travel survey for 100 participants (Wagner, 1997). Casas and Arce (1999) conducted a study to compare trip reporting in household travel diaries to GPS collected data for 150 participants. Doherty *et al.* (2001) integrated GPS travel diary data and interactive computer based travel behaviour surveys. The GPS travel behaviour data collection studies fall into two general categories: in-vehicle based or person based.

2.1.1.1 In-vehicle GPS

In in-vehicle based studies, participants' vehicles were equipped with GPS devices to collect data when a traveler is performing a trip (*e.g.*, Du & Aultman-Hall, 2007; Ogle *et al.*, 2002; Schönfelder *et al.*, 2006). However, in-vehicle based GPS data has some shortcomings. First, a multi-modal trip will not be captured. Second, origin and destination cannot be accurately specified unless they are presented by the first and last location of the vehicles. In addition, the traveler who performed the trip cannot be characterized.

2.1.1.2 Handheld GPS

In handheld GPS based studies, participants are asked either to carry GPS loggers during their trip or, in the case of having a GPS embedded smart phone, installing an application in their smart phone to collect travel diaries (*e.g.*, Asakura & Iryo, 2007; Barbeau *et al.*, 2009; Chung & Shalaby, 2005; De Jong & Mensonides, 2003; Ohmori *et al.*, 2005; Tsui & Shalaby, 2006). Using a device associated with the individual trip maker (*e.g.*, dedicated GPS logger or smart phone) avoids many of the shortcomings associated with using vehicle-based GPS loggers.

2.1.2 GPS data cleaning and filtering

The location accuracy of GPS receivers is typically in the range of 5-10 meters from the actual position (Wolf, 2004). GPS errors can be categorized into two groups: systematic and random errors.

There are two main sources for systematic errors:

1- Number of satellites in view: some GPS receivers can provide the number of satellites in view for each record which gives an indication of how accurate the point is. If the number of satellites is less than four, the point most likely is classified as inaccurate. Therefore, the higher the number of satellites in view, the higher the position accuracy is likely to be;

2- Horizontal dilution of precision (HDOP): When the available satellites are widely spaced across the visible sky, then the estimated position is more accurate. However, when the available satellites are closely aligned, then the estimated position is less accurate (Stopher *et al.*, 2005). HDOP is a measure of this effect; a low value of HDOP indicates that the satellites are widely dispersed and the associated recorded location is expected to be more accurate than when HDOP is large.

There are a number of sources of random errors including: clock or receiver problems, atmospheric and ionospheric effects, urban canyoning errors or signal blocking (Jun *et al.*, 2006). The urban canyoning errors, also called multi-path signal blocking, is the one of the most common sources of random errors in transportation data collection, especially in an urban area. The recorded points are scattered around the actual position due to GPS signals reflecting off of buildings, walls, and surfaces. Signal blocking occurs when the GPS signal cannot be

received and therefore no location data can be obtained. The likelihood of signal blocking is a function of location of the GPS receiver (*e.g.*, placed on a vehicle's dash-board versus placed in a knapsack) and the environment (*e.g.*, presence of tall buildings, indoors vs outdoors, etc.) (Draijer *et al.*, 2000).

Filtering and smoothing techniques have been widely used to address errors in GPS position measurements (*e.g.*, Chung & Shalaby, 2005; Jun *et al.*, 2006; Ogle *et al.*, 2002; Schüssler *et al.*, 2008). Filtering techniques deal with the errors caused by systematic factors and smoothing techniques remove the noise caused by random errors (Schüssler *et al.*, 2008).

Wolf *et al.* (2001) and Ogle *et al.* (2002) indicated that the number of satellites in view along with the HDOP value provide an indication of systematic errors. They filtered and removed the data point if the number of satellites is less than 4 satellites and HDOP value is greater than or equal to 4. Stopher *et al.* (2005) suggested that data should be filtered to include only GPS records for which HDOP is less than 3 and the satellites in view number is greater than 4. Chung and Shalaby (2005) eliminated all the GPS records with less than three satellites in view and HDOP greater than 5. Schüssler *et al.* (2008) did not have access to HDOP or the number of satellites in view and therefore used other criteria applied to altitude, speed and acceleration that they claimed reliably detected systematic position errors. In their study, which was conducted in Switzerland, they omitted all the points with an altitude value less than 200 and more than 4200 meters above sea level. They also filtered the data and removed all the points with speeds that exceed $50\ m/sec$ or acceleration more than $10 m/sec^2$.

The problem of missing data points due to blocked signals is commonly solved by interpolating positions for the last recorded point and the first consecutive recorded points after the signal loss gap (Ogle *et al.*, 2002; Stopher *et al.*, 2005). This method might be reliable for data missing during a short period of time.

Regarding random errors, different smoothing techniques have been used to pre-process GPS data to decrease the impact of this type of error; however, each technique provides different results. Therefore, the choice of smoothing technique is based on the identified problem that needs to be solved and the type of information provided by GPS receivers. A moving average smoothing is commonly used with time series data that creates a series of averages of subsets

data of the full data. This approach was applied by Ogle *et al.* (2002) and Chung *et al.* (2005) on the speed data using Equation (2-1).

$$v_i^s = Average\ (v_{i-1}, v_i, v_{i+1}) \quad (2\text{-}1)$$

where v_i^s = smoothed speed at time *i*;

v_i = raw speed data at time *i*

Jun *et al.* (2006) applied and evaluated three smoothing techniques by applying them to a time series of speed, acceleration, and cumulative distance observations. The first technique is the least-square spline approximation which minimizes the residual sum-square errors (RSS). The second technique is the Kernel-based smoothing method which adjusts the probability of occurrence in the data; this approach had been applied also by Schüssler *et al.* (2008). The third technique is the Kalman filter smoothing method where the data are smoothed by modifying the error iteratively. In their study, the authors modified the conventional Kalman filter to provide better smoothed GPS data. The results show that the modified Kalman filter provided the smallest difference between GPS data and data obtained from the vehicle speed sensor.

2.2 Travel behaviour inference using GPS data

This section presents the relevant techniques reported in the literature for inferring travel behaviour including activity detection, trip purpose, mode inference, etc. A basic assumption is that when individuals are engaged in an activity, transfer from one transportation mode to another, stop at signalized intersections, or are stuck in congestion they are (primarily) stationary – with GPS observations producing speeds within some threshold of 0. Therefore, the data collected from GPS can be generally categorized, based on the recorded speed, into two main categories:

1. Stationary: speed is zero – or close to zero;
2. Non-stationary (moving): speed is more than zero.

Figure 2-1 illustrates a speed-time diagram for a trajectory that includes two hypothetical trips showing the stationary and non-stationary observations. The challenge of this task is

differentiating between the stops where a traveler has participated in an activity from stops associated with a trip such as a delay due to traffic jam, waiting for a bus, etc.

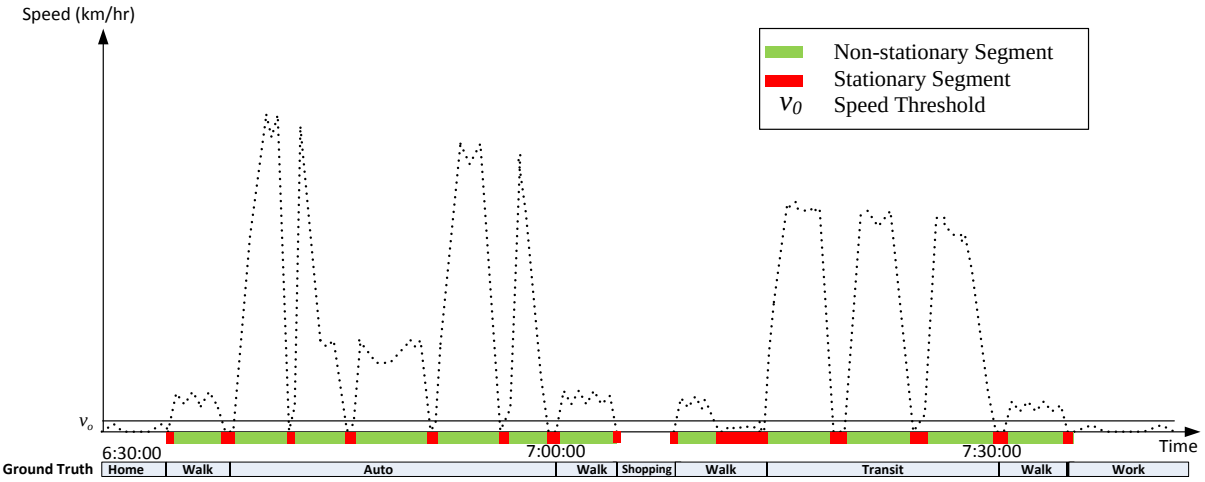


Figure 2-1 Speed-Time diagram for hypothetical trajectory data

2.2.1 Activity stop detection

Trip origin, trip destination, and intermediate activity stops can be detected by identifying activities (stops). The most common approach that has been used to detect an activity is the *Dwell-Time Threshold* approach. The central premise of this approach is that the trip maker is stationary when engaged in the activity. Thus, the technique consists of identifying the times when the GPS data indicate the trip maker is stopped, and when the stopped time exceeds a predefined threshold value the trip maker is assumed to be engaged in an activity.

Wolf *et al.* (2000) carried out a study using in-vehicle GPS units. They tested different dwell-time thresholds (i.e. 60s, 90s, and 120s) and found that if the threshold is set too high, then short duration stops such as pick up or drop off stops will not be identified. On the other hand, if the threshold is too low, non-activity stops (*e.g.*, stops caused by congestion or at signalized intersections) will be misclassified as an activity stop. They found that for their data set optimal performance was found for a threshold equal to 120 seconds.

Doherty *et al.* (2001) identified some cases that the time alone cannot be used to capture activity stops (*e.g.*, a short duration time stop or when the GPS unit was not turned off at the end of the trip). Therefore, they set a location threshold in addition to dwell-time threshold to detect activity stops. In their revised model, if the GPS reported locations remain within the

location threshold of 20 meters for a time that exceeds a threshold of 3 minutes, then the associated sequence of GPS points will be identified as an activity stop. Stopher *et al.* (2004) followed the same methodology but changed the location threshold and dwell-time threshold to 7.4 meters and 2 minutes, respectively.

Axhausen *et al.* (2004) report on the Trip Identification and Analysis system (TIAS) which uses three thresholds to identify activity stops. If the dwell time is greater than 300 seconds (5 minutes), TIAS classifies the occurrence of an activity stop as "confident"; if dwell time is greater than 120 seconds and less than or equal 300 seconds, the occurrence of an activity stop is classified as "probable"; if dwell time is greater than 20 seconds and less than or equal to 120 seconds, activity stop occurrence is classified as "suspicious". The second and third classified categories (i.e. "Probable" and "Suspicious) are overlaid on the GIS road network layer for further classification analysis.

Having the activity stops detected, the next step trip characterization can be extracted. Activity locations can be inferred from the first point's spatial attributes (latitude and longitude) identified in a sequence of non-movement points representing a stop. The trip's origin location can also be inferred from the first point recorded by a GPS logger in a trip. However, the first point recorded by GPS may not be the first point of a trip due to GPS delay of acquiring a signal. Stopher (2004) found that the delay of acquiring a signal varies between 15 seconds to 5 minutes depending on different factors such as movement speed, cold starts, etc. Schonfelder *et al.* (2002) addressed this problem by assuming that the destination location of the previous trip is the origin location of the current trip. The time spent in performing an activity can also be estimated as the difference in time between the first and last points of identified stop (non-movement) sequence of points which presents the arrival and departure time, respectively.

Although all developed methodologies were reported by their authors to provide reasonable performance, these methods were developed primarily for GPS data collected from vehicles and therefore are not able to correctly capture some activity stop situations when the person is using a non-auto mode. Furthermore, these models don't distinguish between a stop associated with a transportation mode choice (*e.g.*, waiting at a bus stop) and activity-stops.

Once the start and end times of each activity stop are identified, the location of each activity stop can be identified. Several techniques have been proposed in the literature, namely:

1. The location (latitude and longitude) of the first GPS point identified in the sequence of activity stop points is designated as the stop location;
2. The first point recorded by a GPS logger in a trip is identified as the origin and the last location recorded for a trip is identified as the destination. This approach assumes the trip maker turns on the GPS logger at the start of each trip (and waits until a position lock is achieved before beginning to travel) and turns off the GPS logger at the end of each trip.

2.2.2 Trip purpose

The second trip characteristic to infer from the data is activity stop purpose (also called trip purpose). This is considered one of the most challenging processes in GPS data analysis. Existing methods have primarily focused on combining the GPS data with GIS (land use and point of interest (POI)) data and the individual trip maker's attribute (*e.g.*, home and work locations, employment status, etc.) in addition to general travel characteristics for the region (*e.g.*, time of day travel patterns for home-to-work trips).

For most of these approaches reported in the literature, one or more manual steps are required. For example, Wolf (2000) proposed fusing the in-vehicle based GPS data with land use information to identify activity/trip purpose. However, a separate database must be constructed to associate each land use type with a primary, secondary, and tertiary (when possible) trip purpose.

Schonfelder and Samage (2003) developed a probabilistic method to infer trip purpose that considered the participants' socio-economic demographic information along with the existing land use data in trip purpose identification using Swedish survey data. They classified the trip purpose as "Home" if the identified trip-end/activity point was within 200 meters of a participant's house location; however, they admitted the possibility of misclassifying activity stops within this radius. Second, trips were assigned as "Work" if the participants are full-time workers and (1) the destination is the second most frequented of all, (2) the temporal trip attributes match the national travel surveys for work purpose, and (3) the trips were performed

on weekdays. For other trip purposes, they identified Points of Interest (POI) within a 300 meter diameter centered at the vehicle destination (Figure 2-2).

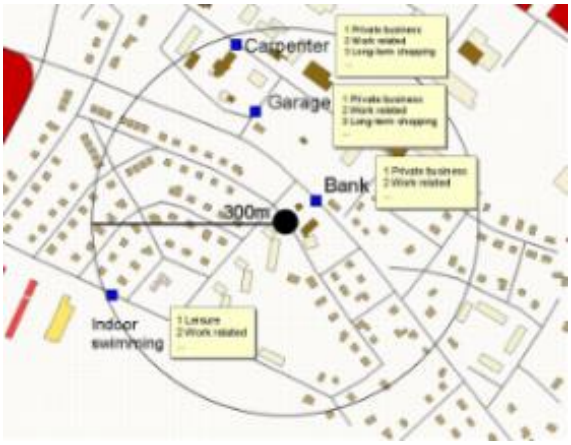


Figure 2-2 Potential trip purpose inference by land use (Source: Schonfelder and Samaga, 2003)

Stopher *et al.* (2008) used hand-held GPS to collect travel data and determined trip purpose on the basis of GIS information, trip characteristics, and information about the locations of the trip maker's most frequent destinations (including work, home, school, etc.). These destinations were obtained directly from the trip maker via a separate survey.

2.2.3 Mode inference

Transportation mode is the third key component of travel survey to infer from the data. This research aims to distinguish between different transportation modes and detect the mode(s) chosen in performing a trip. This can be classified as a pattern recognition problem that can be solved using different machine learning classification algorithms. Therefore, the main goal of this dissertation builds upon machine learning algorithms and techniques.

2.2.3.1 Machine learning

Machine learning is a discipline in which the development of classification algorithms is based on learning from input data (i.e. labelled data) and then these classification algorithms are applied to predict label of previously unseen observations. The transportation mode inference problem can be approached using machine learning classifiers such as Decision Tree (DT), Bayesian Network (BN), Discriminate Analysis (DA), k-Nearest Neighbors (k-NN), etc. Recently, Bayesian network and k-NN have received significant attention to solve

classification problems due to their flexibility and ability to graphically represent the problem, as well as relatively easy interpretation of the results. Both models are briefly described in the following subsections.

Bayesian Network

Bayesian network, $B =< G, X, P >$, is a probabilistic graphical model consisting of directed acyclic graph (DAG), $G =< X, P >$, representing a set of nodes corresponding to random variables, $x_i \in X$, (*e.g.*, attribute, feature, event, etc.) and links corresponding to informational or a causal dependencies among variables (Pearl, 1988). Figure 2-3 demonstrates a BN example where $x_1 : x_4$ are random variables; the node for variable x_1 is a parent for the two child nodes of variables x_2 and x_3; nodes x_2 and x_3 are parents of the node x_4. The dependency for each node, random variable, is typically specified by a conditional probability distribution (CPD) given its parents, $p(x_i|x_{pa(i)}) \in P$, as they are demonstrated in Figure 2-3. For discrete random variables, the conditional probability is represented by a table which lists the probability of each feasible value in a child node given each combination of values of its parents. A BN is a graphical representation that encodes a joint probability distribution ($P(X)$) factored over the set of random variables (X) given the parent nodes of each variable ($x_{pa(i)}$) as follows:

$$P(X) = \prod_{i=1}^{n} P(x_i \,|x_{pa(i)}) \tag{2-2}$$

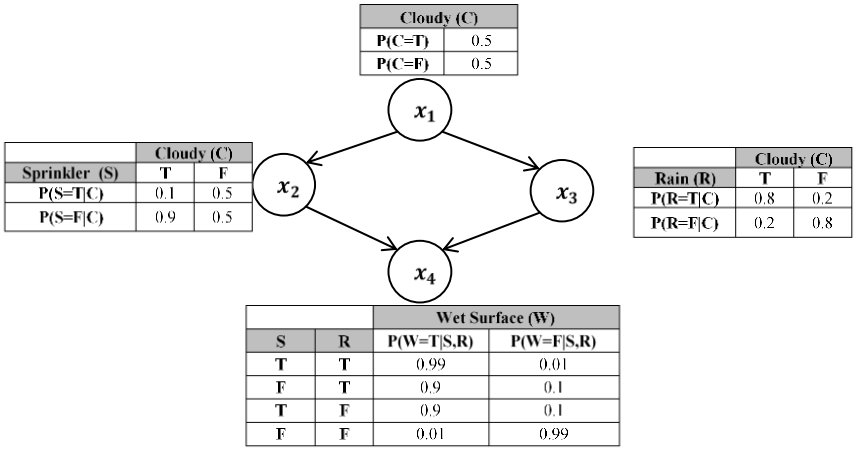


Figure 2-3 B N representing CPT for five random variables (Source: Murphy, 1998)

BN can be utilized as a classifier due to its ability to estimate the conditional probability for a node given the values assigned to the parent/child nodes. The main goal of a BN classifier is to correctly predict the class label for an instance described by a vector of features. Two main processes are involved in applying BN as a classification model: learning and inference. In the learning process, the BN model attempts to learn from a given set of labeled data represented by a set of attribute variables in order to determine the structure of the network and the CPD parameters. The BN inference process classifies instances based on the trained model by applying Bayes rule to compute the posterior probability of class $c \in X_C$ given specific features $x_i \in X_i$, as follows:

$$P(c \mid x_i) = \frac{p(x_i, c)}{p(x_i)} = \frac{p(x_i \mid c)\, p(c)}{\sum_c p(x_i \mid c)\, p(c)} \tag{2-3}$$

A simple form of BN is a Naïve-Bayes (NB) classification model whereas the class node (X_C) is a parent node of all other nodes (X_i) (Duda and Hart, 1973), as shown in Figure 2-4. NB classifier assumes that all feature variables are independent from each other. Although this assumption appears to be problematic, it has been shown that the NB model outperforms more sophisticated classifiers (Langley *et al.*, 1992). In addition, the NB classifier is flexible enough to be extended in order to overcome this limitation. Then, the Bayes rule will be applied to compute the posterior probability of class $c \in X_C$ given the feature vector for a particular instance of $X_1, X2, \ldots, X_i$ as follows:

$$P(Xc \mid X_i) = \prod_{i=1}^{n} P(X_i \mid X_C) \tag{2-4}$$

The classification result would be the predicted class with the highest posterior probability for each mode segment obtained from the previous algorithm.

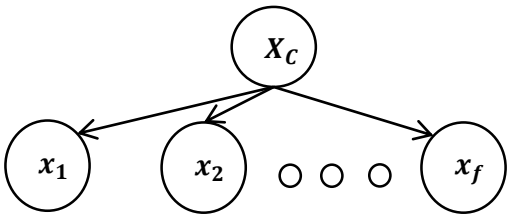


Figure 2-4 Naïve Bayes network

k-Nearest Neighbors

k-Nearest Neighbors (k-NN) is a non-parametric method that consistently achieves high performance among the various methods of supervised statistical pattern recognition (Dasarathly, 1990 and Fukunga, 1973). First referenced by Fix and Hodges (1951), k-NN has been widely used since the early 1970s, most likely because it has a solid theoretical foundation, can be easily interpreted, and performs well. The simple principle behind k-NN is to calculate the distance between the *k* nearest number of training cases (nearest neighbors) to the new sample in the feature space. Then, the new sample –observation- will be classified based on an appropriate estimator to the most common class among the *k* nearest neighbours. Figure 2-5 illustrates a simple example of k-NN to classify one observation, x_1, where k=5 (nearest neighbors).

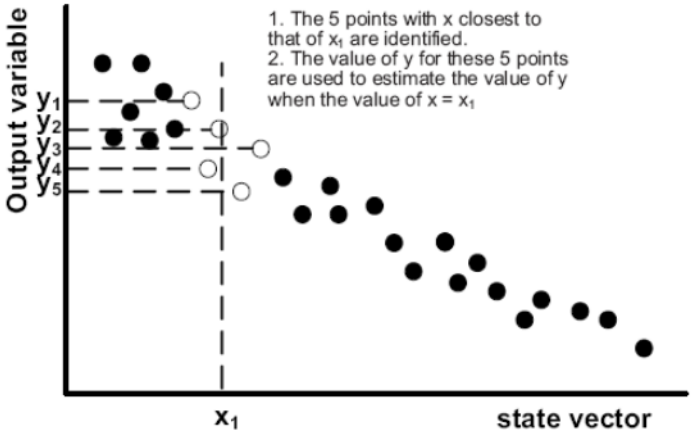


Figure 2-5 Example of k-NN classification mode with k=5 (cited from Steve, 2005)

The performance of k-NN relies heavily on a set of parameters *i.e.* number of nearest neighbors (*k*), distance metric, and feature vector. Fukunaga and Hostetler 1973 have identified the optimum *k* as a function of sample size, number of feature space, and the underlying distribution of dependent variables against the independent variables.

Regarding distance metric, there are various metrics to determine the distance between the sample point (observation) and the set of training data. The standard Euclidean distance is regularly used in order to identify the nearest neighbors of an observation. However, Fukunaga and Hostetler (1973) suggest the Mahalanobis distance as the optimal metric; this distance metric considers both variance and covariance of each variable with other variables in order to

measure the correlation between the feature vector of the current data and the training data (Mahalanobis, 1936).

The selection of the set of variables (attributes) in the feature vector is also an important aspect in the context of k-NN. For example, choosing a large set of attributes might result in irrelevant attributes dominating relevant ones, which negatively impacts the performance of the model. Therefore, in the case of highly dimensional data (*e.g.*, number of attributes is more than 10), dimension reduction and feature extraction are usually performed. Feature extraction process aims to transform the high-dimensional data (original input data) into a fewer dimensional space (reduced set of features) that represents the relevant information from the original data. Both processes should be applied to the raw data prior to calibrating or applying the k-NN model.

Principle Component Analysis (PCA) is a technique to achieve dimensionality reduction and feature extraction in one step. PCA can be viewed as a transformation of the original data from a coordinate axes to a new set of axes that are aligned with the variability in the data (Abdi *et al.*, 2010 and Alter *et al.*, 2000). The transformed data preserves the variability. Other researchers have used Principal Components Analysis (PCA) to reduce the dimensionality of the data used in classification; for instance, the study conducted by Reddy *et al*. (2010). However, it should be noted that the use of PCA does not necessarily improve the performance of the classification model. In some cases PCA improves the classification accuracy; in other cases, the PCA produces poorer results.

2.2.3.2 Case studies

Chung and Shalaby (2005) identified the points where travelers transfer from one mode to another as a mode transfer point (*MTP*). They assumed walk mode can be used as an intermediate mode in any mode transfer. Therefore, they developed an algorithm to classify MTPs in one of three categories: end-of-walk (EOW), start-of-walk (SOW), and end-of-gap (EOG), where the gap is the period when a loss of GPS signals occurred. The feature vectors used to categorize SOW and EOW points are speed ($> 10\ km/hr$) and acceleration ($< 0.1\ m/s^2$). In addition, they applied several rule-based algorithms to assure the quality of the previous classification model. Then, they used a fuzzy logic-based model to classify mode segments using the following fuzzy variables from GPS records: average speed, 95^{th} percentile

maximum speed, positive median acceleration, and number of satellites in view as a measure of data quality.

Tsui and Shalaby (2006) extended the work by Chung and Shalaby (2005) by using GIS map and transit route service information. The fundamental approach is to match a travel segment (from the GPS) with the presence of a transit route (from GIS). Quantitatively, they developed a route searching algorithm that is only activated when the resulting membership of cycling and bus from the fuzzy logic classifier exceeds a threshold. When the route searching algorithm matches at least one transit route, the segment is labeled as transit. The addition of this route searching algorithm to their original classification model improved the accuracy of the classifier from 76% to 80%.

Gong *et al.* (2011) and Schüssler (2010) also followed Chung and Shalaby in identifying transportation mode segments. To further distinguish transit trips, these authors classified travel segments as transit trips when a segment's start and end points (origins and destinations) are sufficiently close to transit stations - the term *transit station* indicates any location at which a transit vehicle is scheduled to stop to board and discharge passengers.

Other researchers attempted to improve their classification models' performance by acquiring additional information. For example, Stenneth *et al.* (2011) utilized not only GIS data on transit route and station locations, but also temporal information – a comparison of transit schedules with the recorded times from the GPS. The authors report very high identification accuracy. Other researchers have used household transportation attributes, such as automobile and bike ownership to enhance classification accuracy.

Moiseeva *et al.* (2010) developed a system called "TraceAnnotator" that uses the Bayesian belief network (BBN) to automatically detect the transportation mode from GPS trace data and GIS database. In addition, the system identifies activity type where GPS data are fused with (personalized) land use data. The BBN represents the dependencies between different spatial and temporal variables with other factors to estimate the conditional probability that a particular mode has been used. The variables considered in the BBN are as follows: distance to railway track, average and maximum acceleration, average and maximum speed, speed deviation, distance accumulated, car and bike ownership, etc. There are seven different modes into which the system has to classify the data into: walk, run, bike, motorcycle, car, bus, train.

An evaluation was conducted using approximately 1500 trips compiled from 8 different individuals. The authors report an overall accuracy level of approximately 92%. However, most trips were made by non-motorized modes and the misclassification rate (precision) for some modes was relatively high.

Gonzalez *et al.* (2008) developed a smartphone application called TRAC-IT which collected GPS data and supplemental information from the user (*e.g.*, number of occupants in the vehicle, transportation mode, trip purpose etc.). The data are sent to a server to be analyzed through a Neural Network (NN) classifier to detect the transportation mode associated with the trip.

The method was tested on a very small data set (114 trips). The reported accuracy of the system was quite high (91%); however only 3 modes were considered, and only for a limited number of travel conditions and trip types.

Reddy *et al.* (2008 and 2010) developed a transportation classification system that runs on a mobile phone equipped with GPS and 3-axis accelerometer. The classifier relies only on speed from the GPS and accelerometer data. The classifier distinguishes between five different transportation modes (Still, Walk, Running, Bike, and Motorized). The classifier is composed of a Decision Tree (DT) followed by a discrete Hidden Markov Model (DHMM) that analyses the speed from GPS data along with variance and frequency components of the accelerometer. DT is applied to differentiate between the characteristics' boundaries of transportation modes. Then, DHMM eliminates the noise based on the knowledge of the previous transportation mode and the likelihood of transition into the next mode.

This method was evaluated with a very small data set; data were collected from 16 individuals (8 males and 8 females) who were asked to gather 15 minutes of data for each of the five modes in a given consecutive sequence. The results presented by the authors indicate an overall accuracy of 93.6%; however, the accuracy varies depending on the mode being identified. In particular, the model has difficulties differentiating between walking and biking and between different motorized modes (*e.g.*, auto versus bus).

The most significant constraint with this model is the use of a mode transition matrix in the HMM. This matrix must be calibrated from labeled data. If certain mode transitions are not captured in the labeled data used for calibration, then these mode transitions cannot be predicted by the model. Furthermore, this transition matrix is assumed to be constant over

time which implies that transportation network changes or policy initiatives have no impact on these probabilities. This implication is almost certainly false.

Zheng *et al.* (2008) used the same technique as proposed by Reddy *et al.* but used different variables in the feature vector and computed these features on the basis of the GPS data after the data had been divided into a sequence of fixed interval time steps. They carried out an evaluation study in which they tested different discretization interval durations (i.e. for segmenting the GPS data) and found an overall accuracy of 75%. However, this approach uses a state transition matrix and therefore suffers from the same limitations identified previously.

Rasmussen *et al.* (2013) described a method to identify transit segments using the percentage of stops occurring at transit stations. They establish appropriate thresholds based on the characteristics of individual transit routes; as the level of service on the route increases (implying less time delayed at signalized intersections or in congestion), the percentage of stops should also increase. Other researchers have used household transportation attributes, such as automobile and bike ownership to enhance classification accuracy. Stopher *et al.* (2008) developed an algorithm that only assigns car or bike as a mode for a trip if the household indicates ownership of a car or bike.

These methods have demonstrated improvements in identifying transit segments. However, opportunities exist to advance these approaches. For example, a more robust model will be able to correctly classify walking or cycling segments that occur immediately adjacent to a transit line.

2.3 Summary

This chapter has focused on the work that has been directed to infer travel behaviour from GPS data, mainly handheld GPS devices.

From the literature review the following observations can be made:

1. Most of the trip-end/activity stop detection models using GPS data have been developed based on in-vehicle GPS units. For these models, the mode is known and locations are constrained to the road network.

2. Several proposed models assume that trip makers turned on the GPS logger at the beginning of the trip and turned off the GPS at the end of the trip and therefore the trip origin and trip destination are known (or at least more easily identified).
3. Identifying short duration activity stops and distinguishing these stops from non-activity stops cannot be done with adequate accuracy on the basis of basic features, such as speed, time, etc. Instead, additional spatiotemporal features must be incorporated.
4. The differentiation between activity stops and activities associated with using specific modes *e.g.*, waiting time in a bus station for a bus has not been addressed.
5. In mode detection, much of the previous work depends upon assumptions about trip constructs – *e.g.*, all trips must begin and end with a walking segment that lasts longer than one minute. In addition, walking mode has been considered as the transition mode between any two modes having a minimum duration of 60s. Consider the example where someone is dropped off at a transit stop and boards an arriving vehicle shortly thereafter. In this case, a very short (in time and distance) walking trip will be observed.
6. Most of the previous researchers have employed GPS data gathered at very short intervals (i.e. 1 second), primarily using data loggers.
7. Previous work has relied on what can be called a "static" feature set – a pre-determined set of travel attributes – to identify the transportation mode.
8. Travel profile of different travel modes is affected by traffic conditions which has a direct influence on the descriptive features of each travel mode, especially the motorized ones. Differentiating between motorized and non-motorized modes is much more challenging when the motorized mode is operating in highly congested traffic.
9. Machine learning classification models have been applied to the problem of transportation mode detection and classification and have shown promising results.
10. The literature reviewed has shown that various researchers have employed different classification techniques, parameters values, and data formats and have reported a wide range of accuracies. There does not appear to be a clear indication of which

classification technique, which set of parameter values, or which data format, should be used.

11. The integration of detailed GIS databases (*e.g.*, traffic network, transit infrastructure, etc.) appears to be necessary to automatically infer transportation mode with sufficient accuracy to be of practical value.

12. A comprehensive framework to collect, analyze, and infer traveler transportation mode from a fully passive GPS-embedded smartphone, to the best of my knowledge, has not been developed.

The research proposed herein is aimed to infer travel behaviour information from GPS embedded smartphone (i.e., GPS and accelerometer) by developing an automated framework to optimize the choice of the classification model and the associated parameters.

In our work, we propose a method that classifies transportation modes using fewer and less restrictive assumptions on trip structure. We also develop the model based on data gathered at longer intervals – five seconds. The developed model is automated such that multiple combinations of classification techniques, parameter values, and data formats are concurrently employed. From the literature, the conditions under which PCA improve the model's function are not explicitly known. As such, in our work, we train our model using original data and only principal components produced from the PCA. This approach allows us to determine *a posteriori* whether to include PCA. This gives our modeling method more flexibility and increases our model's performance with a wide range of data inputs.

The performance of the classification models have been improved by integrating spatial statistics and GIS for the transportation infrastructure in identifying a sequence of points as a transit segment. Finally, the performances of the classification models were evaluated on larger data sets than those used in some of these previous publications such as Chung and Shalaby (2005).

In the next chapter, the travel behavior data collection process and system are introduced as the starting step in the proposed framework (Figure 1-4).

Chapter 3

Travel Behaviour Data Acquisition

As discussed in previous chapters, the conventional techniques of travel behaviour data collection such as paper-based and face-to-face survey suffer from the following weaknesses 1) the low number of survey participants due to the burdens on users; 2) lack of precision associated with the recorded information due to the delay between the time of conducting the survey and the time of the trip. The previous chapters have also explained the benefits of using GPS-embedded smartphones to gather these data passively and with enhanced accuracy.

In this research, accurate spatiotemporal data for individuals' travel have been gathered using smartphones. The full framework to be explained in this thesis is shown in Figure 3-1 highlighting the focus of this chapter which describes in more detail the travel behaviour data acquisition. The chapter is divided into three main sections. In the first section, we briefly describe the system used to collect the data and its components *i.e.* the user and server sides. The second section describes the data collection procedures. In the third section, general statistics representing the collected data are presented.

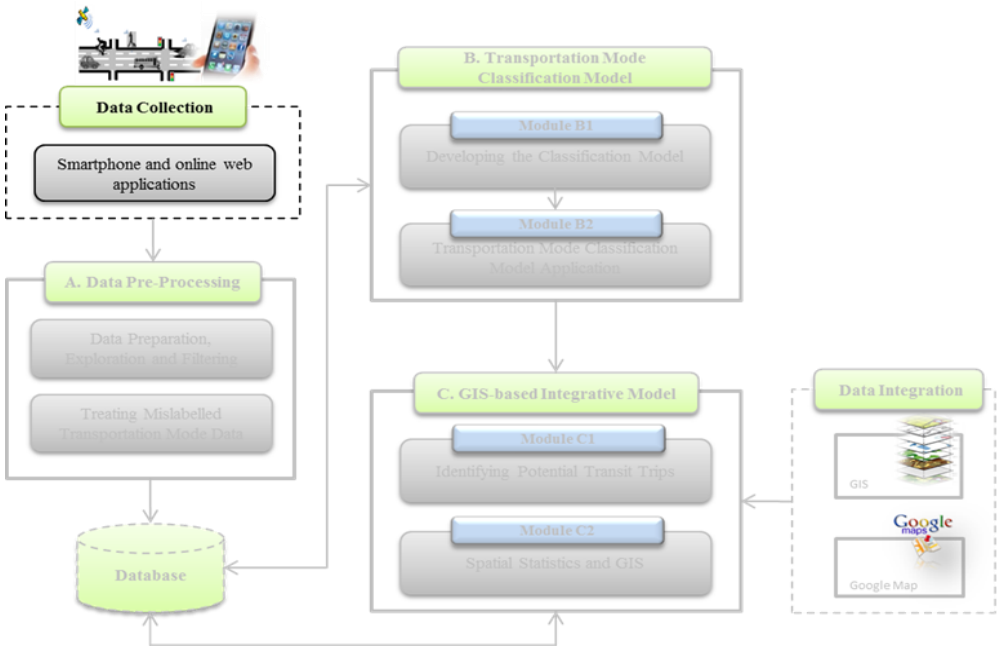


Figure 3-1 Proposed Transportation Mode Classification Framework

(Data Collection)

3.1 Travel Behaviour Data Acquisition System

The data acquisition system used in this study is called TrackBerry - a custom travel survey application developed by the **I**ntelligent **T**ransportation **S**ystems **R**esearch **G**roup (ITSRG) at the University of Waterloo, Canada (Taghipour, 2012). The system consists of two main components as shown in Figure 3-2:

1- User side: a smartphone application to collect and label the data in real time, and a web-based application to verify data points and label after completion of a trip; and

2- Server side: a database to store the smartphone data.

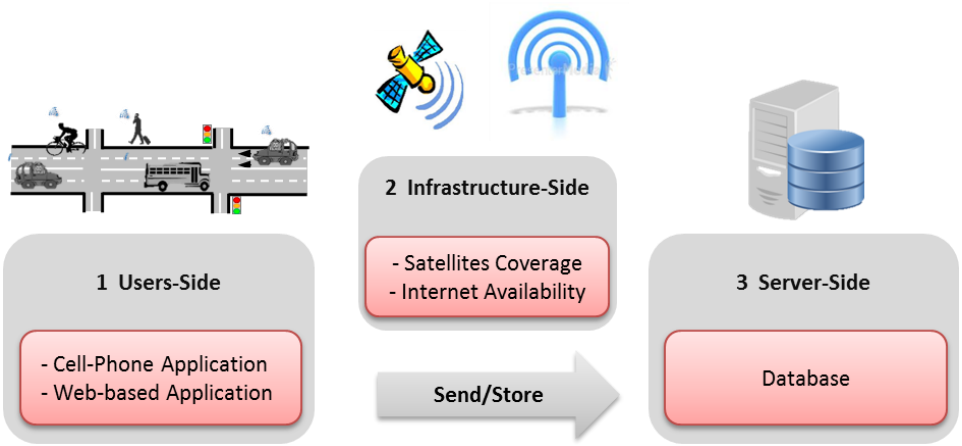


Figure 3-2 TrackBerry travel survey system component

On the user side, the smartphone application can be downloaded to Blackberry smartphones equipped with a GPS receiver. Once installed, the application is designed to automatically record real time spatial-temporal data from GPS at pre-set frequency. In addition, the application allows survey participants to label their current transportation mode from a pre-set list of labels *i.e.* Walk, Bike, Transit, Auto, Activity, or Waiting (Figure 3-3). This feature is accessible by users at any point of time during the trip to allow them to change the label whenever they switch from one mode to another transportation mode. Once the data are collected, they are transmitted wirelessly to a server to be stored in a database for further analysis.

Figure 3-3 Transportation mode alternatives list in TrackBerry application

The second part of the user side is the web application. Survey participants have online access that allows reviewing, modifying and verifying the recorded trip information in order to ensure the labels are correct (Figure 3-4).

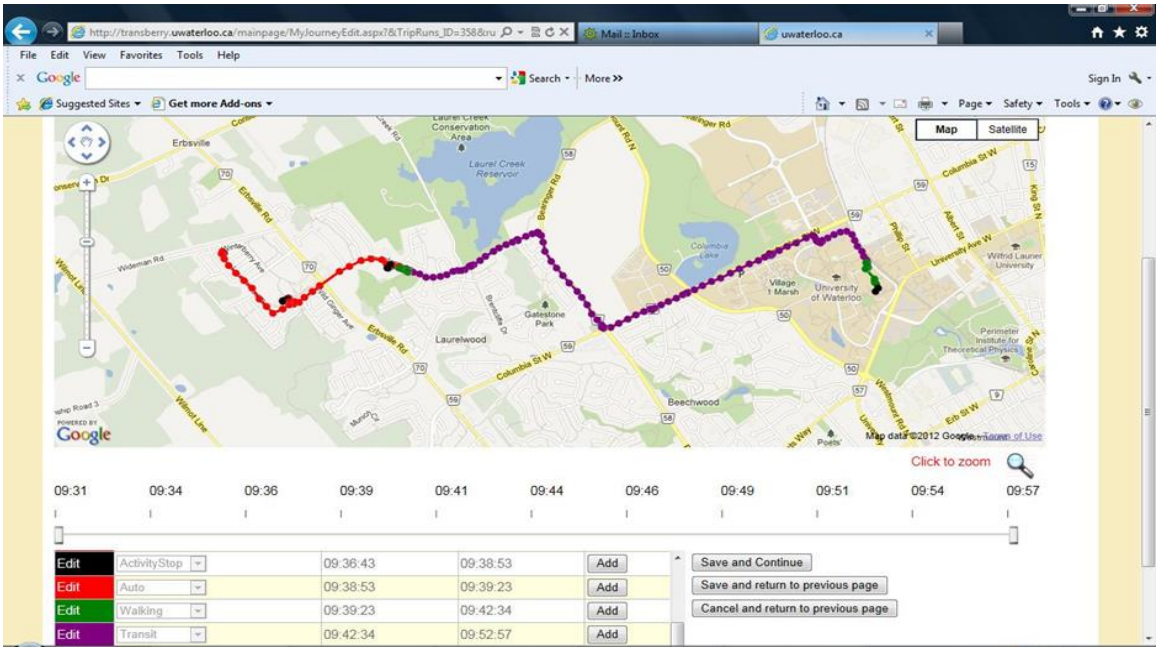


Figure 3-4 Online web-interface to verify components of users' trips

On the server side, a database is created to store the sent data from all participants. These data are a potential source for identifying essential trip elements and extract travel behaviour information as will be discussed in the following chapters.

3.2 Data collection procedures and termenology

Survey participants began by creating an account in the system and each was assigned a unique traveler ID (*r*). Then, all participants were given simple instructions for installing and operating the mobile application in their smartphones. They were asked to turn on the application before they leave their homes in the morning and turn it off at least 15 minutes after they arrived back at end of the day. In addition, participants were asked to carry their smartphones all the time during their travel.

The smartphone application was programmed to record real time spatiotemporal data every 5 seconds. Simultaneously, participants were asked to label the time and location when they switch to a new transportation mode, for example switching from walking to driving a car. At the end of each day, participants were advised to review the recorded transportation mode labels of their performed trips through the web-based application and apply any necessary correction to their labels.

The data obtained from a smartphone are stored in the database according to the unique traveller ID (*r*). Each record contains a single observation which consists of time stamp; position (*x*,*y*,*z*); acceleration; and transportation mode label. A time series of consecutive observations for a traveller *r* is referred to as a trajectory[2]. Normally, a trajectory is expected to contain data for an entire 24 hour period (or the length of time that the smart phone is on). Therefore, the trajectory consists of data reflecting all trips and activities performed by traveler *r* during the day *d*.

A trip is identified when a traveler moves from one location (origin) to another location (destination) to perform an activity. Hence, a trip is defined as the movement between two activity locations. The first data point of a trip is defined as start-of-trip (*SOT*) and this point is

[2] A new trajectory ID is generated each time the software application is started on the smartphone and at the beginning of each day. Consequently, if a traveller is collecting data and the collection period spans midnight, these data will be divided into two trajectories. Also, if the traveller closes and then restarts the application, the data collected after the application is restarted will be associated with a new trajectory ID.

also an end-of-activity (*EOA*) point. A trip can involve one or more transportation modes to travel from an origin to a destination.

The portions of a trip that contain sequential data points labeled as having been completed by the same mode is defined as Transportation Mode Segments (*TMS*). When a trip is comprised of a single mode, the entire trip will consist of one *TMS*. Alternatively, when a trip involves multiple transportation modes (and therefore multiple *TMS*), a data point *i* can be identified as the point when a traveler transfers from one mode to another. This point is called a mode-transfer-point (*MTP*). Figure 3-5 illustrates a portion of a hypothetical trajectory (*e.g.*, only the data from 6:30 AM to 7:30 AM) on which the trip components are labeled.

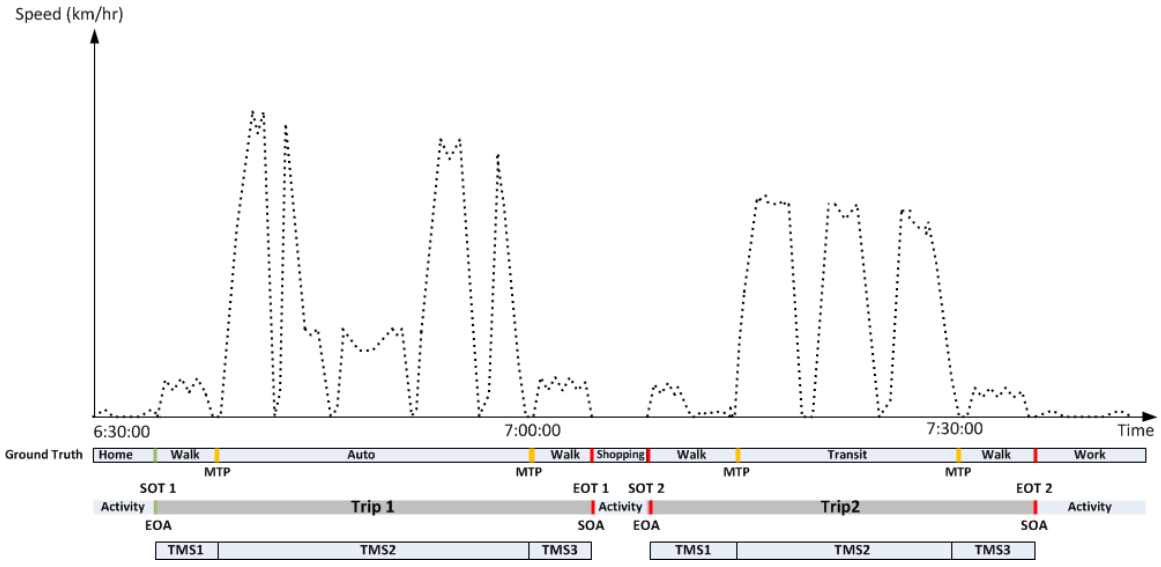


Figure 3-5 Illustration of terminology for a portion of a hypothetical trajectory

3.3 Data Collection Results

To gather the data necessary for the research, 20 people volunteered, 14 males and 6 females, to collect GPS data from smartphones between October 2011 and May 2013. A significant proportion of the data was gathered between October 2011 and May of 2012. After 2012, recognizing a need for more data, I and several colleagues conducted a second data collection exercise. In total, 326 trajectories were verified and stored which include different transportation modes and activities. The ultimate goal of the research is to identify transportation modes based solely on automatically collected data from smart phones. The

model should be capable of correctly identifying travel modes from data generated by a large number of travelers, often with unique trip-making characteristics: origin destination pairs, walking speeds, cycling speeds, driver behavior, etc.

The data set used in this model is limited slightly by both the spatial limits of the data – most of the data were gathered in the Regional Municipality of Waterloo – and by the small and relatively homogeneous population from which the travel data are generated – the majority of data producing volunteers were between 25 and 35 year old students from the University of Waterloo. As such, a potential limitation of the model is that these input data do not reflect a full diversity of traveler behaviors. However, the trip patterns observed in the data set do reflect significant diversity in many categories: mode, time of day, duration of trips, and congestion levels. Perhaps most importantly, the stopping patterns – i.e. the occurrence of stationary segments that define transportation mode segments – demonstrate considerable variability and randomness, such that the analysis is started with confidence that the observations are sufficiently uncorrelated to demonstrate the strength of the mode classification model. Additional information about the variability in the data is contained in Chapter 4 (Figures 4.12 through 4.14).

Table 3-1 provides statistical information about the trajectories. In total, 642 trips, 791 transportation mode segments, and 128,820 data points were collected. Figure 3-6 illustrates a sample trip trajectory that includes four activities and three different transportation modes (Walk, Auto, and Bus).

The ultimate goal of the research is to identify transportation modes based solely on automatically collected data from smart phones. The model should be capable of correctly identifying travel modes from data generated by a large number of travelers, often with unique trip-making characteristics: origin destination pairs, walking speeds, cycling speeds, driver behavior, etc.

The data set used in this model is limited slightly by both the spatial limits of the data – most of the data were gathered in the Regional Municipality of Waterloo – and by the small and relatively homogeneous population from which the travel data are generated – the majority of data producing volunteers were between 25 and 35 year old students from the University of Waterloo. As such, a potential limitation of the model is that these input data do not reflect a

full diversity of traveler behaviors. However, the trip patterns observed in the data set do reflect significant diversity in many categories: mode, time of day, duration of trips, and congestion levels. Perhaps most importantly, the stopping patterns – i.e. the occurrence of stationary segments that define transportation mode segments – demonstrate considerable variability and randomness, such that the analysis is started with confidence that the observations are sufficiently uncorrelated to demonstrate the strength of the mode classification model. Additional information about the variability in the data is contained in Chapter 4 (Figures 4.12 through 4.14).

Table 3-1 Statistics and information of investigated trajectories'

Transportation Mode	**Duration (min)**	**# points**	***#TMS***
Walk	2986	21754	224
Bike	1979	18947	105
Transit	1111	10463	65
Auto	8807	77656	397

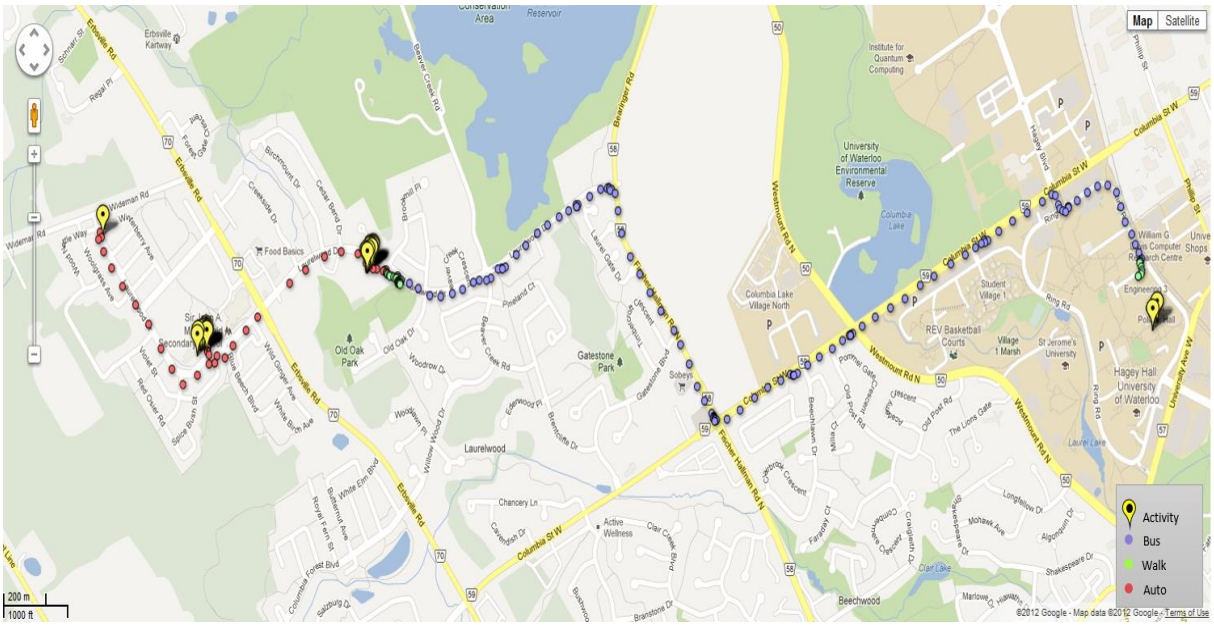


Figure 3-6 GPS trace for a verified sample trip

One of the main contribution of this research is to improve upon existing methods to identify automatically modes of travel from GPS data. Speed is one of the key attributes for

distinguishing between different transportation modes (*e.g.*, an observation with a speed of 110 km/h is not a walk mode). As noted in the literature review, there are generally two approaches to quantifying the properties of movement. The first is to generate a series of attributes based on individual data points. In these instance-based models the speed attribute in the feature vector consists of the instantaneous speed for the individual GPS point. However, all modes, including auto and transit, experience low speeds at least some of the time. Consequently, observing that the speed for a particular observation is low does not provide enough information to make a reliable estimate of the mode.

The second method, applied here, is to investigate the transportation mode based on the attributes of a series of observations – using mode segments rather than individual observations. Using the average speed from each transportation mode segment (*TMS*) is expected to demonstrate greater (and therefore more distinguishable) differences between motorized and non-motorized travel. Further, the ranges of observed values of speed are typically much smaller. A necessary step to completing this approach is identifying activity stops and mode transfer points, hence the emphasis in the literature review of previous methods of stop or activity identification.

To demonstrate the difference between instance based and segment based approaches, consider the data set introduced in the previous table in this section.

Table 3-2 provides a summary of the conventional and proposed (i.e. segment level) attributes for each category of mode. The mean instantaneous speed and the mean average segment speed for each mode are very similar. However, the standard deviation computed on the basis of the instantaneous speeds is much larger than the standard deviation computed from the average segment speeds. Though this is not unexpected, it demonstrates that the distributions of speeds between the conventional approach and the proposed approach are quite different. Minimizing the variance of data points has significant benefits for mode inference.

Table 3-2 Speed characteristics of different transportation modes (at individual observation and segment levels)

	Conventional (Instances)			Proposed (Segment)		
Mode	**# of Obs.**	**Mean (m/s)**	**Standard Deviation (m/s)**	**# of Obs.**	**Mean (m/s)**	**Standard Deviation (m/s)**
Walk	21,754	1.10	0.79	224	1.05	0.86
Bike	18,947	2.87	2.47	105	3.36	0.97
Bus	10,463	5.76	5.35	65	6.42	2.71
Auto	77,656	13.29	12.20	397	11.07	6.41
Total	128820			791		

Figure 3-7 illustrates the boxplot of speed over each mode based on the instantaneous speed. The boxes are bound by the 25^{th} and 75^{th} percentile observations for each mode; the horizontal line in the center of the box represents the median value. The dashed extensions above and below the box represent the 2.7σ and -2.7σ, respectively. The data points outside of these ranges are considered outliers.

As expected, all modes have data points that are common with very low speed, which makes differentiation quite challenging in this range. The 25^{th} percentile of the recorded speed from both Auto and Bus data are very similar to the Walk and Bike speed distributions. In this case, if instantaneous speed is used to classify mode, the model will classify these points incorrectly. Further, when comparing motorized modes, the distribution for Transit is completely contained within the range for Auto. This overlap demonstrates the challenges for mode inference even within a subset of modes.

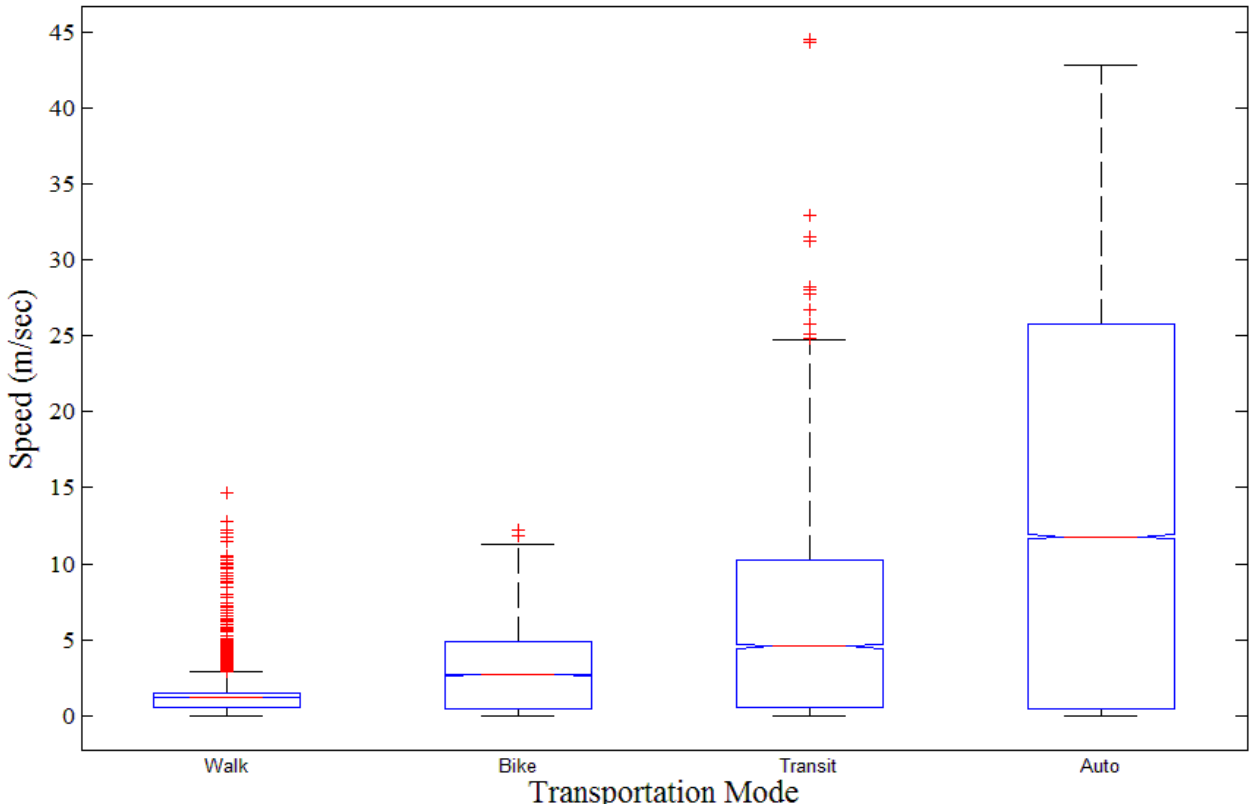


Figure 3-7 Boxplot of instantaneous speed by transportation mode

Figure 3-8 presents a similar boxplot, but using the mean speed data from the transportation mode segments rather than the instantaneous speeds. The most important observation is the significant reduction in overlap between the 25th and 75th percentile boundaries for all modes. In fact, no overlap exists between walking, cycling and transit; and though the overlap between transit and auto remains, it is also much less than at the instantaneous speed level.

Given these observations, two conclusions can be drawn about mode inference at the segment level:

1. Using speed (or other attributes discussed in subsequent chapters) as a differentiating variable at the segment level reduces the problem of misclassifying some Auto and Transit points as Walk or Bike modes (or vice-versa).

2. Even at the segment level, reliable classification between Transit and Auto may require additional attributes be incorporated within the feature vector.

While the boxes in Figure 3-8 exhibit stronger differentiating attributes, a problem remains with the range of data observed beyond the whiskers which extend to +/-2.7σ and are considered outliers. While some of these data may be valid, others may be the result of incorrect labels provided by the user. The next chapter describes the methods by which these suspicious points are systematically investigated.

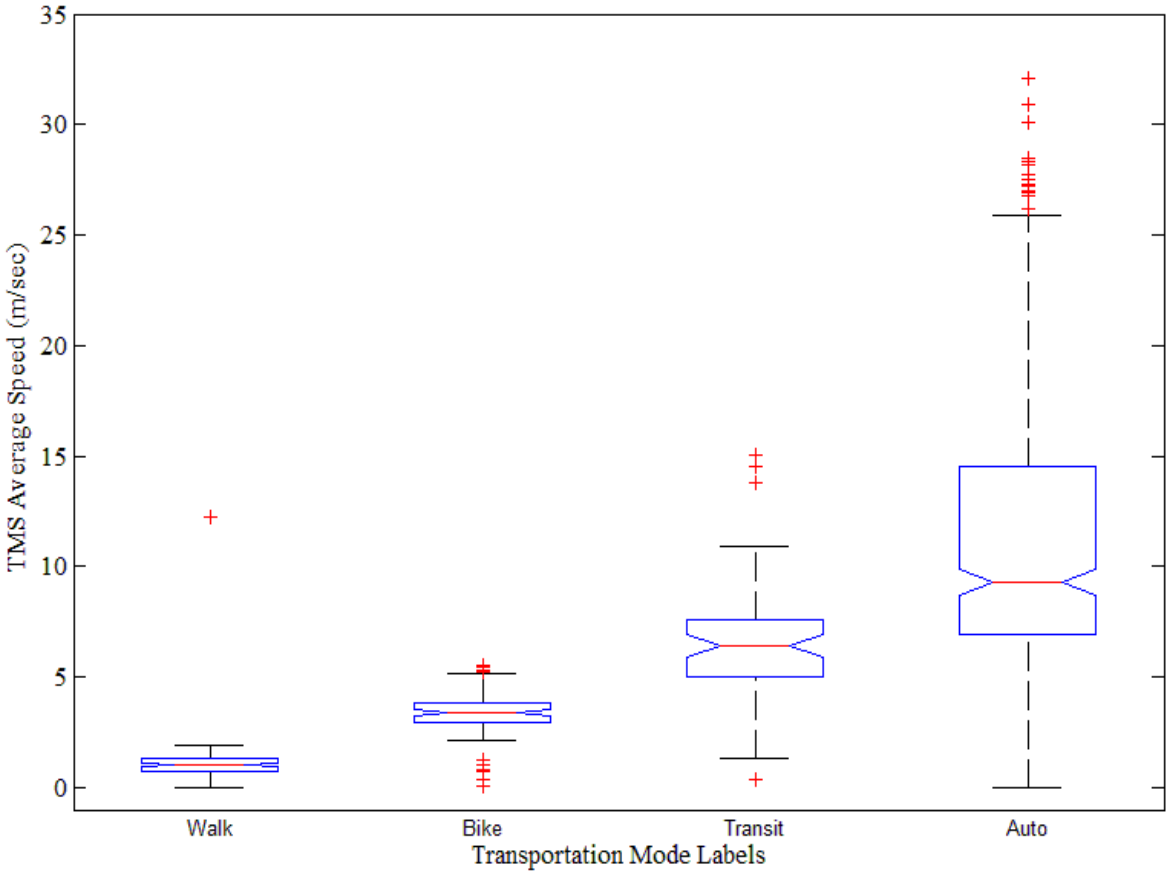


Figure 3-8 Boxplot of mean transportation mode segment speeds for different transportation modes

3.4 Summary

In this chapter, the process by which data are gathered, labeled, and transferred to a server is explained. Moreover, the chapter describes statistically the data that were gathered and that will form the basis of the analysis in subsequent chapters. Next, the necessary terminology was introduced including key concepts of trajectory, trip, activities, transportation mode segments and mode transfer points. Finally, the chapter demonstrates the importance of completing the mode inference at the segment level rather than using an instance (or point-based) approach.

Once the GPS smartphone data are collected and successfully transferred to the server, the database is ready for processing. The database contains observations of trips collected at 5 seconds frequency. Each observation includes trajectory ID, trip ID, time stamp, position (longitude and latitude), speed, and acceleration. The collected and stored data are subject to different sources of error that might have a significant impact on the modeling process. Therefore, the next chapter will explain the custom software tool and Graphical User Interface that was developed and applied to investigate the suspicious data first introduced in Figure 3-6. These data are systematically pre-processed to remove, or at least reduce, the impact of errors (from various sources) on the overall modeling process.

Chapter 4

Data pre-processing (Module B)

The primary objective of this research is to develop a robust transportation mode classification model. However, as discussed in earlier chapters and illustrated in Figure 4-1, there is a need to pre-process the data acquired from the GPS equipped smart phones before using these data to calibrate or apply a mode classification model. This chapter describes the data pre-processing methods as well as the custom software tool that was developed to facilitate the data pre-processing. The data pre-processing, shown in the highlighted portion of Figure 4-1, consists of two steps aimed at filtering, cleaning, and treating the data by developing:

1- a software tool to investigate and review the trips in the dataset to ensure the quality of saved trips and remove unreliable trips (trips that include extreme outliers or trips with long period(s) of missing data); and

2- an algorithm to identify incorrectly labelled transportation mode transfer points and to automatically correct these errors (either in terms of time and/or location) in these labels.

The following sections describe these two steps in more detail.

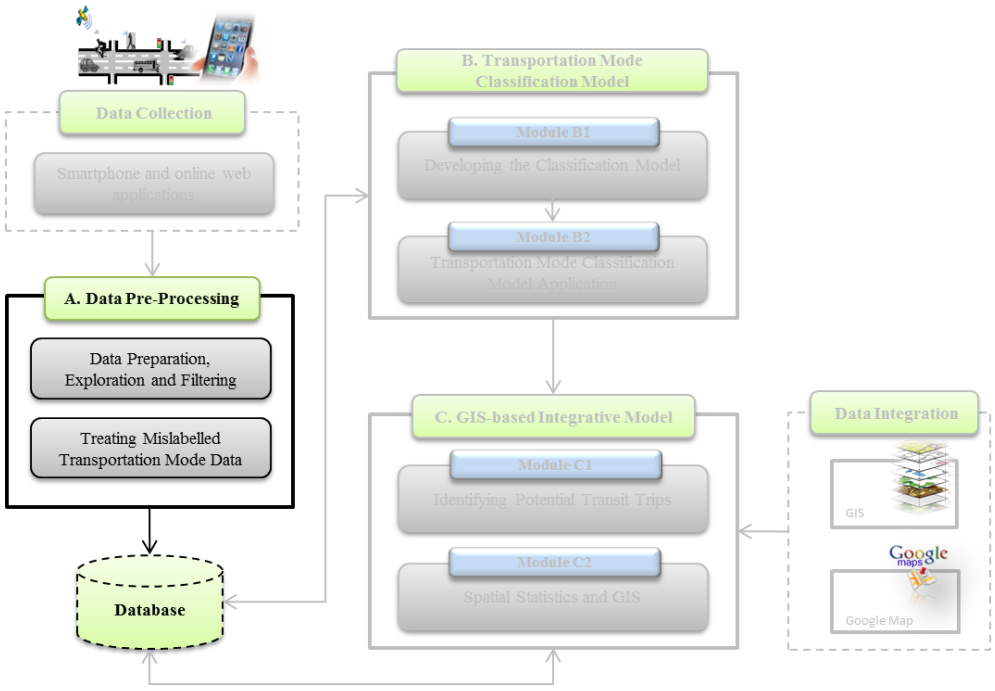


Figure 4-1 Proposed Transportation Mode Classification Framework

(A. Data Pre-processing)

4.1 Data Preparation, Exploration and Filtering

The database generated by the Trackberry application and used in this study contained 451,307 location records stored as (latitude, longitude, elevation, time) for each traveler, *r*, on a given day, *d*. To prepare the data for the transportation mode classification, the location data are transformed from World Geodetic System (WGS84) - latitude and longitude - into the Cartesian (x,y) coordinate system by applying the following formulas (Defense Mapping Agency, 1987):

1. Convert Geodetic coordinates from decimal degrees to radians

$$lon = (Longitude) \times \frac{\pi}{180} \quad (4\text{-}1)$$

$$lat = (Latitude) \times \frac{\pi}{180} \quad (4\text{-}2)$$

2. Convert Geodetic coordinates in radians into Cartesian coordinates in meters

$$x = (N + alt) \times cos(lat) \times cos(lon) \quad (4\text{-}3)$$

$$y = (N + alt) \times cos(lat) \times sin(lon) \quad (4\text{-}4)$$

where

$$N = \frac{R}{\sqrt{(1 - e^2) \times sin(lat)^2}} \quad (4\text{-}5)$$

R: the earth radius in meter (6378137 m)
e: WGS 84 ellipsoid constant ($81.81919e^{-3}$)
lat: geodetic latitude (radians)
lon: longitude (radians)
alt: height above WGS84 ellipsoid (m)

Next, an initial filtering is applied to identify and remove duplicate records. In the database, due to either errors in GPS or in the communication protocol, some records are duplicated with the same location and time recorded. In these cases, one of the duplicate observations was deleted. Approximately 1500 observations (0.2% of the dataset) were identified as duplicate records and were deleted.

Given 714,615 initial data points, and a desire to ensure data quality in subsequent analysis, the first step in data exploration was to use naive testing to identify missing data points or suspicious points. The data were evaluated at both the trip and the segment-level (*TMS*) assuming user labels were correct. The evaluation of trips resulted in the deletion of all data associated with a trip when at least one of the following two conditions was met:

1. The trip is less than five minutes in length (47 out of 642 have been identified); or
2. The trip has no transportation mode labels (6 out 595 trips have been identified).

A total of 53 trips were eliminated from the 642 in the original dataset (8%) resulting in 589 trips for further analysis.

The remaining trips were investigated at the *TMS* level. The approach taken was to calculate the average speed for every *TMS*. Those that were outside of the 25^{th} and 75^{th} percentile were labeled as suspicious and requiring further investigation.

4.1.1 Custom Data Visualization Software Tool

A significant challenge when dealing with a large dataset is being able to simultaneously perform data visualization and quantitative analysis. Data visualization helps to explain most of the ambiguous behaviour of the data by visually investigating the interrelationship between different attributes. It is a powerful technique to convey useful information in a quick, efficient, and effective manner. There are many existing software tools that can do either the quantitative analysis or the visualization, but few that can do them simultaneously. Therefore, a custom software tool was developed to help explore the collected trips and investigate suspicious points, segments or trips.

The developed tool produces:

1- Speed-time diagrams: the interface generates plots of speed – both actual instantaneous and average moving speed (a smoothing over seven points) as a function of time. The slope represents measures of acceleration.

2- Space-time diagrams: the interface generates a plot of distance traveled as a function of time. The slope represents velocity.

3- Observations of actual labels over time: the interface generates a plot of user-labels as a function of time.

4- A plot of a trip map: the user can choose to plot the trips traces (x,y) on a google map to investigate trip attributes relative to the spatial attributes. For example, for autos, segments with zero speeds can be correlated with the location of intersections; similarly long periods of very low speeds occurring on freeways can be identified more easily as auto trips in congestion rather than walking trips. In addition, the GUI enables users to plot a buffer around the traces and specify the transportation system infrastructure within the buffer i.e. traffic signal, transit bus stations, etc.

An additional feature of the diagrams is that they are interactive. If a user plots a trip segment containing many points, the user is able to graphically select a subset of the points. This subset is automatically highlighted on all diagrams containing the appropriate data. The subset of data is also highlighted in the source table.

Figure 4-2 illustrates the developed GUI. First, the user starts with choosing the trajectory ID and trip ID - in case the trajectory includes more than one trip - from the drop-down lists. Then, the user has the capability to investigate the observations associated with either all transportation modes or a specific mode. Next, the type of diagrams to be generated is specified such as space-time, speed-time, map, etc.

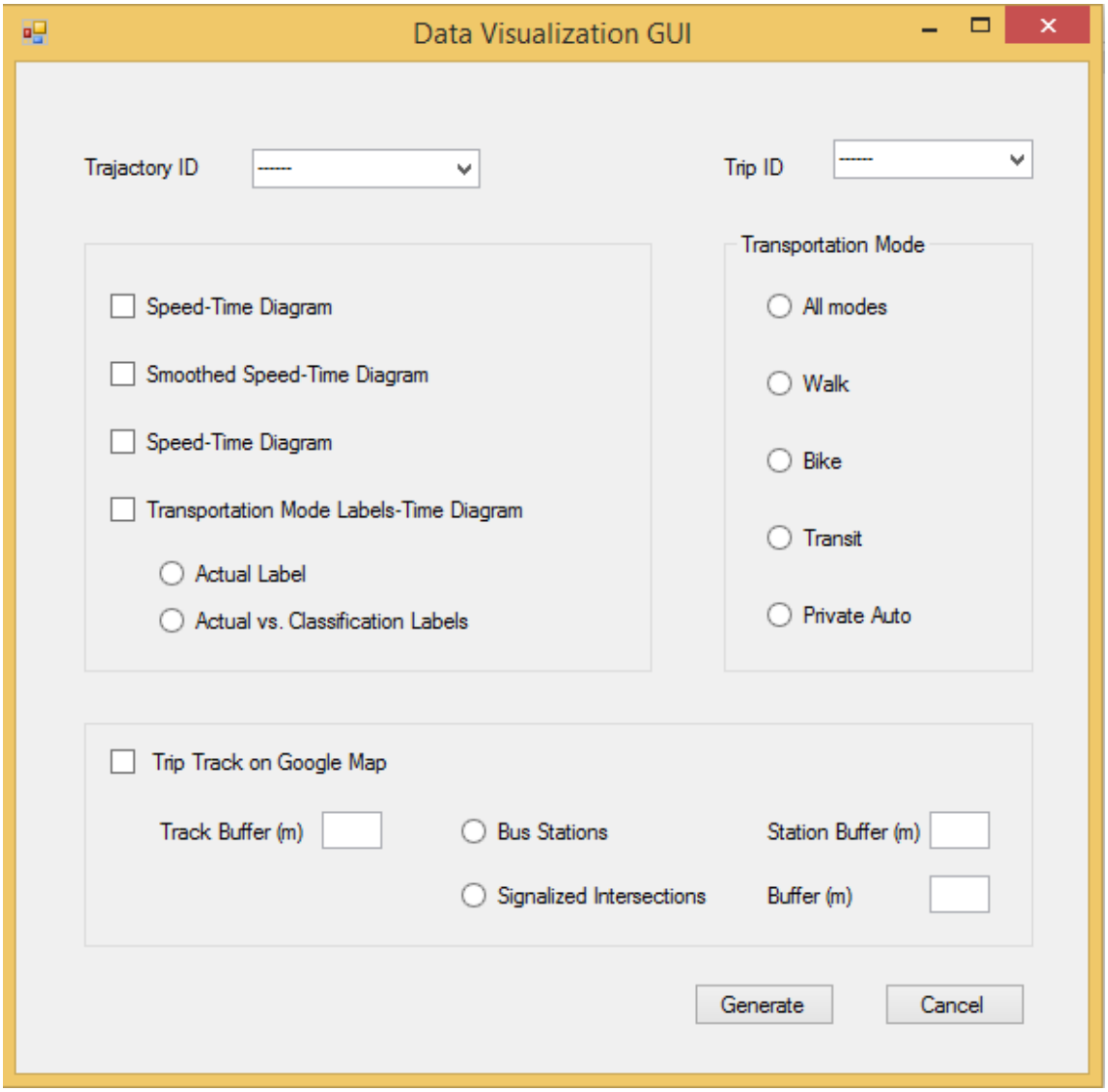


Figure 4-2 Data visualization Graphical User Interface (GUI)

Figure 4-3 illustrates an example of generating a space-time diagram, a speed-time diagram, a smoothed speed-time diagram, and transportation mode labels for trajectory 1193346 and trip ID 1. All transportation modes used in this trip will be investigated at this time.

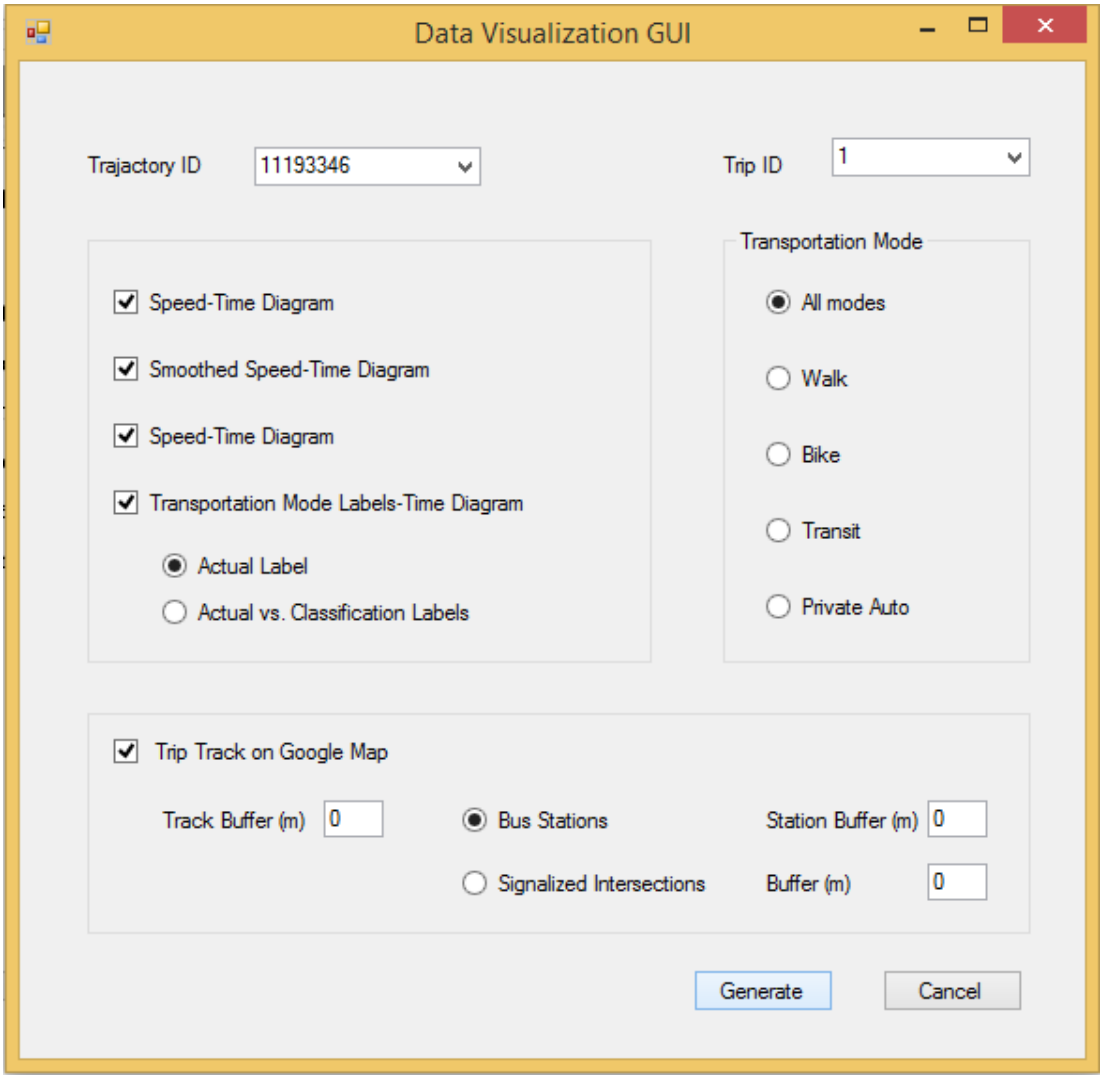


Figure 4-3 Data visualization GUI example for a trip (1) in a trajectory (11193346)

Figure 4-4 shows the diagrams generated from the previous step. The title of the figure refers to the trajectory ID. As noted, all diagrams are linked and interactive. The dashed red box in Figure 4-4 demonstrates how a user can select a subset of points – in this case a series of very low speed points – to investigate their properties. In this case, the user highlights the points on the speed time diagram and the same points are highlighted on smoothed speed, the distance and the mode label diagrams. Also, the points are highlighted on the map and the source table shown in Figure 4-5.

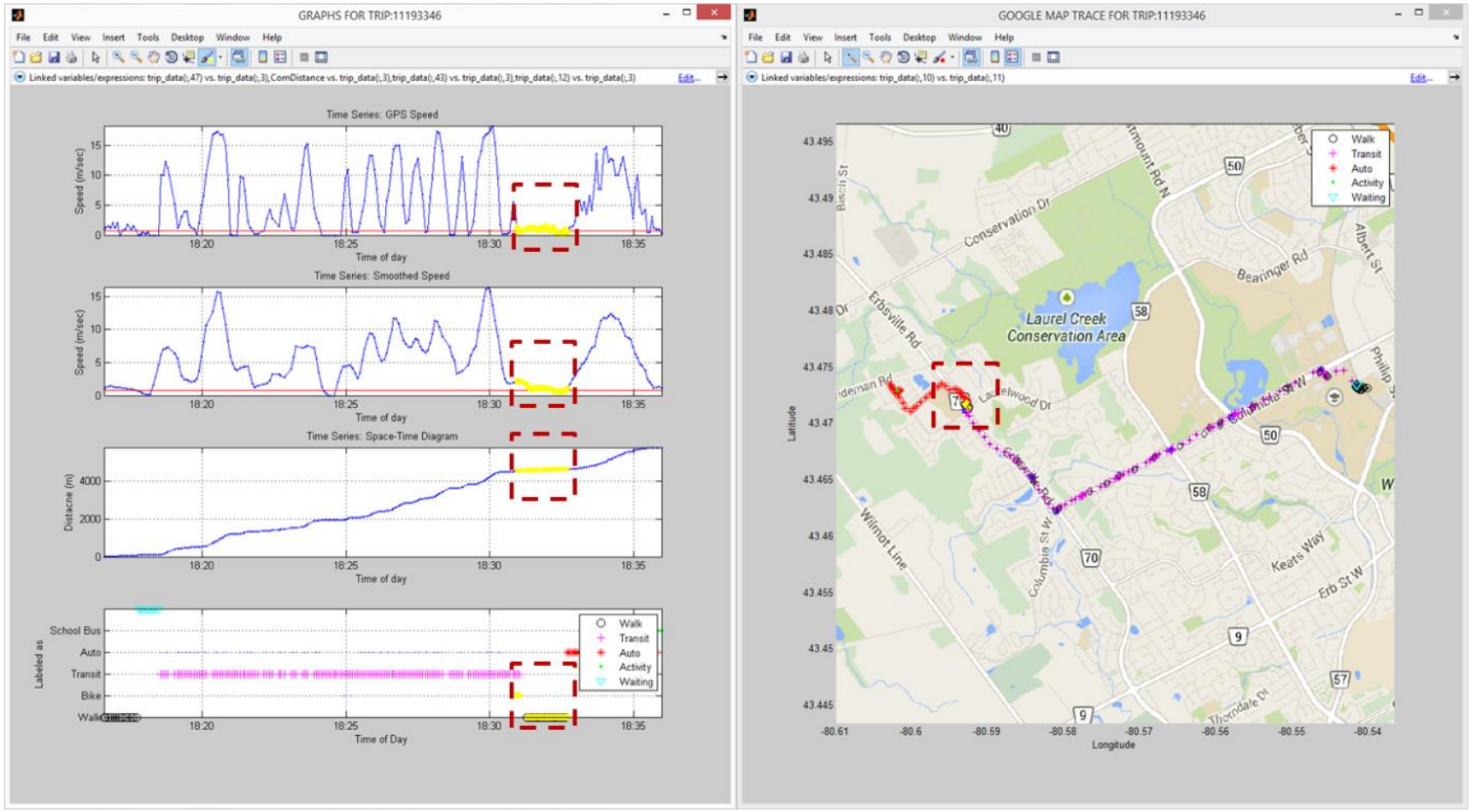
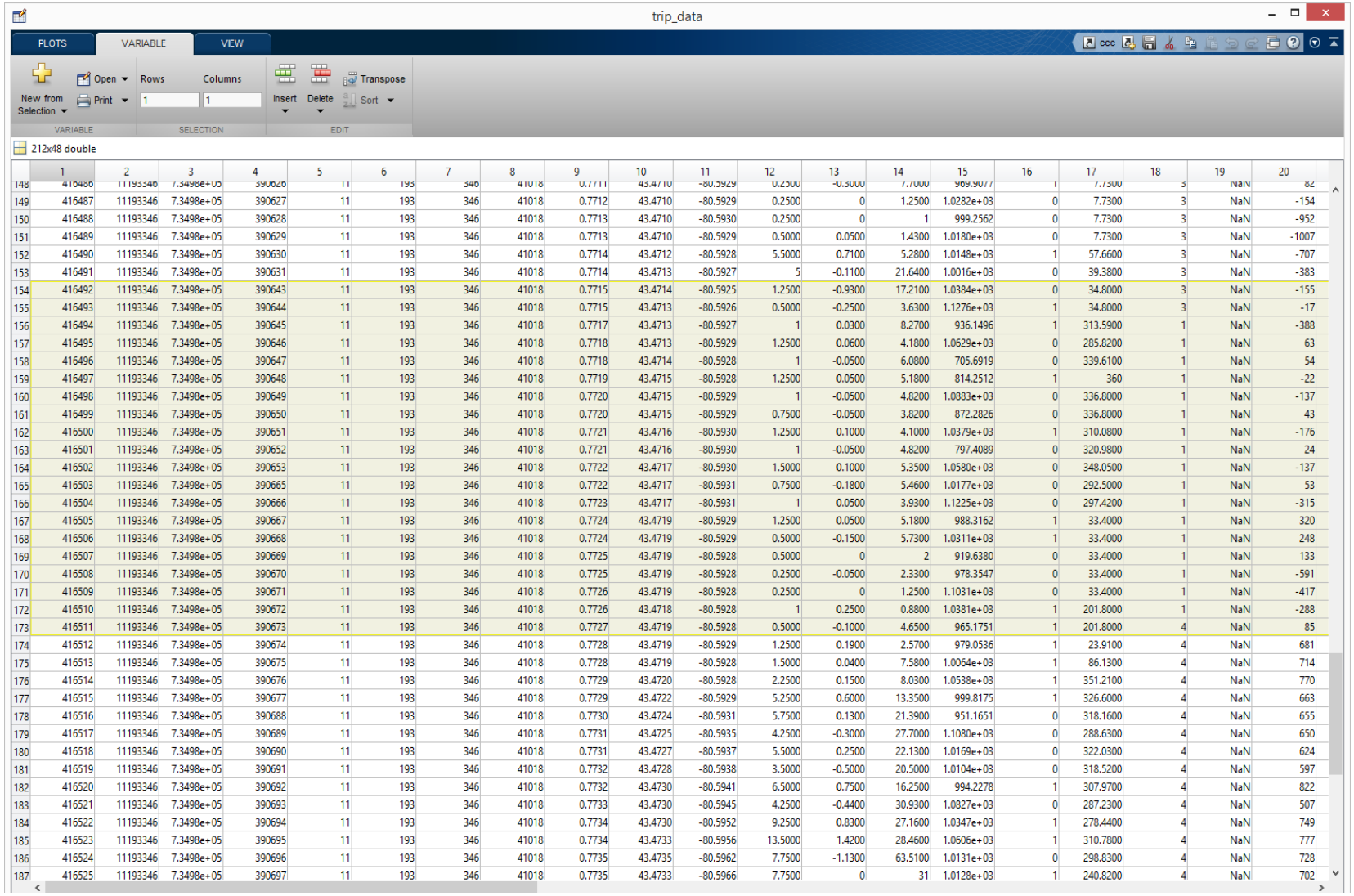


Figure 4-4 The generated interactive plots from GUI



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
148	416486	11193346	7.3498e+05	390626	11	193	346	41018	0.7711	43.4710	-80.5929	0.2500	-0.3000	7.7000	969.9077	1	7.7300	3	NaN	82
149	416487	11193346	7.3498e+05	390627	11	193	346	41018	0.7712	43.4710	-80.5929	0.2500	0	1.2500	1.0282e+03	0	7.7300	3	NaN	-154
150	416488	11193346	7.3498e+05	390628	11	193	346	41018	0.7713	43.4710	-80.5930	0.2500	0	1	999.2562	0	7.7300	3	NaN	-952
151	416489	11193346	7.3498e+05	390629	11	193	346	41018	0.7713	43.4710	-80.5929	0.5000	0.0500	1.4300	1.0180e+03	0	7.7300	3	NaN	-1007
152	416490	11193346	7.3498e+05	390630	11	193	346	41018	0.7714	43.4712	-80.5928	5.5000	0.7100	5.2800	1.0148e+03	1	57.6600	3	NaN	-707
153	416491	11193346	7.3498e+05	390631	11	193	346	41018	0.7714	43.4713	-80.5927	5	-0.1100	21.6400	1.0016e+03	0	39.3800	3	NaN	-383
154	416492	11193346	7.3498e+05	390643	11	193	346	41018	0.7715	43.4714	-80.5925	1.2500	-0.9300	17.2100	1.0384e+03	0	34.8000	3	NaN	-155
155	416493	11193346	7.3498e+05	390644	11	193	346	41018	0.7715	43.4713	-80.5926	0.5000	-0.2500	3.6300	1.1276e+03	1	34.8000	3	NaN	-17
156	416494	11193346	7.3498e+05	390645	11	193	346	41018	0.7717	43.4713	-80.5927	1	0.0300	8.2700	936.1496	1	313.5900	1	NaN	-388
157	416495	11193346	7.3498e+05	390646	11	193	346	41018	0.7718	43.4713	-80.5929	1.2500	0.0600	4.1800	1.0629e+03	0	285.8200	1	NaN	63
158	416496	11193346	7.3498e+05	390647	11	193	346	41018	0.7718	43.4714	-80.5928	1	-0.0500	6.0800	705.6919	0	339.6100	1	NaN	54
159	416497	11193346	7.3498e+05	390648	11	193	346	41018	0.7719	43.4715	-80.5928	1.2500	0.0500	5.1800	814.2512	1	360	1	NaN	-22
160	416498	11193346	7.3498e+05	390649	11	193	346	41018	0.7720	43.4715	-80.5929	1	-0.0500	4.8200	1.0883e+03	0	336.8000	1	NaN	-137
161	416499	11193346	7.3498e+05	390650	11	193	346	41018	0.7720	43.4715	-80.5929	0.7500	-0.0500	3.8200	872.2826	0	336.8000	1	NaN	43
162	416500	11193346	7.3498e+05	390651	11	193	346	41018	0.7721	43.4716	-80.5930	1.2500	0.1000	4.1000	1.0379e+03	1	310.0800	1	NaN	-176
163	416501	11193346	7.3498e+05	390652	11	193	346	41018	0.7721	43.4716	-80.5930	1	-0.0500	4.8200	797.4089	0	320.9800	1	NaN	24
164	416502	11193346	7.3498e+05	390653	11	193	346	41018	0.7722	43.4717	-80.5930	1.5000	0.1000	5.3500	1.0580e+03	0	348.0500	1	NaN	-137
165	416503	11193346	7.3498e+05	390665	11	193	346	41018	0.7722	43.4717	-80.5931	0.7500	-0.1800	5.4600	1.0177e+03	0	292.5000	1	NaN	53
166	416504	11193346	7.3498e+05	390666	11	193	346	41018	0.7723	43.4717	-80.5931	1	0.0500	3.9300	1.1225e+03	0	297.4200	1	NaN	-315
167	416505	11193346	7.3498e+05	390667	11	193	346	41018	0.7724	43.4719	-80.5929	1.2500	0.0500	5.1800	988.3162	1	33.4000	1	NaN	320
168	416506	11193346	7.3498e+05	390668	11	193	346	41018	0.7724	43.4719	-80.5929	0.5000	-0.1500	5.7300	1.0311e+03	1	33.4000	1	NaN	248
169	416507	11193346	7.3498e+05	390669	11	193	346	41018	0.7725	43.4719	-80.5928	0.5000	0	2	919.6380	0	33.4000	1	NaN	133
170	416508	11193346	7.3498e+05	390670	11	193	346	41018	0.7725	43.4719	-80.5928	0.2500	-0.0500	2.3300	978.3547	0	33.4000	1	NaN	-591
171	416509	11193346	7.3498e+05	390671	11	193	346	41018	0.7726	43.4719	-80.5928	0.2500	0	1.2500	1.1031e+03	0	33.4000	1	NaN	-417
172	416510	11193346	7.3498e+05	390672	11	193	346	41018	0.7726	43.4718	-80.5928	1	0.2500	0.8800	1.0381e+03	1	201.8000	1	NaN	-288
173	416511	11193346	7.3498e+05	390673	11	193	346	41018	0.7727	43.4719	-80.5928	0.5000	-0.1000	4.6500	965.1751	1	201.8000	4	NaN	85
174	416512	11193346	7.3498e+05	390674	11	193	346	41018	0.7728	43.4719	-80.5929	1.2500	0.1900	2.5700	979.0536	1	23.9100	4	NaN	681
175	416513	11193346	7.3498e+05	390675	11	193	346	41018	0.7728	43.4719	-80.5928	1.5000	0.0400	7.5800	1.0064e+03	1	86.1300	4	NaN	714
176	416514	11193346	7.3498e+05	390676	11	193	346	41018	0.7729	43.4720	-80.5928	2.2500	0.1500	8.0300	1.0538e+03	1	351.2100	4	NaN	770
177	416515	11193346	7.3498e+05	390677	11	193	346	41018	0.7729	43.4722	-80.5929	5.2500	0.6000	13.3500	999.8175	1	326.6000	4	NaN	663
178	416516	11193346	7.3498e+05	390688	11	193	346	41018	0.7730	43.4724	-80.5931	5.7500	0.1300	21.3900	951.1651	0	318.1600	4	NaN	655
179	416517	11193346	7.3498e+05	390689	11	193	346	41018	0.7731	43.4725	-80.5935	4.2500	-0.3000	27.7000	1.1080e+03	0	288.6300	4	NaN	650
180	416518	11193346	7.3498e+05	390690	11	193	346	41018	0.7731	43.4727	-80.5937	5.5000	0.2500	22.1300	1.0169e+03	0	322.0300	4	NaN	624
181	416519	11193346	7.3498e+05	390691	11	193	346	41018	0.7732	43.4728	-80.5938	3.5000	-0.5000	20.5000	1.0104e+03	0	318.5200	4	NaN	597
182	416520	11193346	7.3498e+05	390692	11	193	346	41018	0.7732	43.4730	-80.5941	6.5000	0.7500	16.2500	994.2278	1	307.9700	4	NaN	822
183	416521	11193346	7.3498e+05	390693	11	193	346	41018	0.7733	43.4730	-80.5945	4.2500	-0.4400	30.9300	1.0827e+03	0	287.2300	4	NaN	507
184	416522	11193346	7.3498e+05	390694	11	193	346	41018	0.7734	43.4730	-80.5952	9.2500	0.8300	27.1600	1.0347e+03	1	278.4400	4	NaN	749
185	416523	11193346	7.3498e+05	390695	11	193	346	41018	0.7734	43.4733	-80.5956	13.5000	1.4200	28.4600	1.0606e+03	1	310.7800	4	NaN	777
186	416524	11193346	7.3498e+05	390696	11	193	346	41018	0.7735	43.4735	-80.5962	7.7500	-1.1300	63.5100	1.0131e+03	0	298.8300	4	NaN	728
187	416525	11193346	7.3498e+05	390697	11	193	346	41018	0.7735	43.4733	-80.5966	7.7500	0	31	1.0128e+03	1	240.8200	4	NaN	702

Figure 4-5 The generated Interactive data source table from the GUI

4.2 Treating Mislabelled Transportation Mode Data

As mentioned earlier, survey participants had online access to a web interface that allowed for reviewing, modifying and verifying the recorded trip information after uploading the data. However, the verified trips may still suffer from errors in the labelling process. The most common source of error arose from participants mislabelling multiple points as "activity" when transportation was occurring and, the opposite, labelling points as transportation during activities. This largely resulted from participants making label changes at times not coinciding with the beginning and end of an activity.

The method by which activity beginning and end points were corrected is as follows. When an activity is preceded or followed by a motorized mode, it is relatively straightforward to identify the boundaries of the activities using travel speed values. When an activity occurs between walking or cycling trips, the speed differential itself fails to provide sufficient differentiation to identify the limits of the activity. To resolve this challenge, the actual trip path in the x,y plane was examined. The limits of the activity were identified as the points at which travel became much more circuitous and less direct.

A second source of error occurred when the traveller changed from one transportation mode to another mode, but did not change the transportation mode labels at the correct time (i.e. they changes the label earlier or later than when the change in transportation mode actually occurred).

This error in labelling is illustrated in Figure 4-6, which shows velocity as a function of time for a hypothetical trajectory. The traveller makes two trips. In each trip, the traveller walks from the trip origin to access a motorized transportation mode (in trip 1 this is auto and in trip 2 this transit), and then walks from the point of departure from the motorized mode to the destination. Thus there are a total of eight mode transfer points (*MTPs*). The figure illustrates the ground truth as well as the labelled data. The labelled *MTPs* are identified with an orange numbered arrow. It is evident from a comparison of the labelled MTPs and the ground truth that there is a discrepancy in the times at which the mode transfers occurred. For some of the *MTPs*, the labelled time of transfer is too early and for others, it is too late.

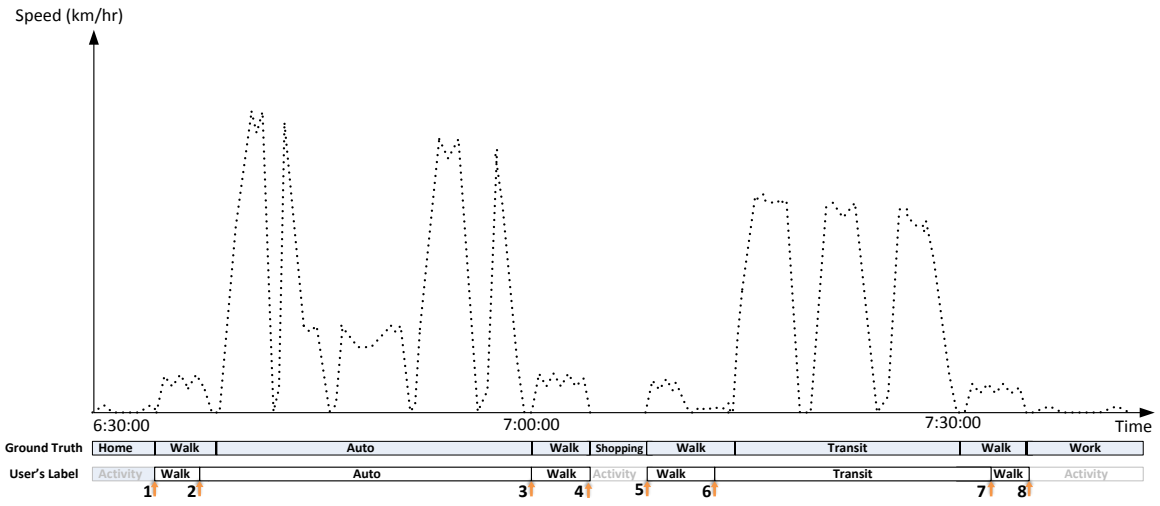


Figure 4-6 Errors in user labelling of *MTP*s for a hypothetical trip

It was found that these errors in *MTP* labelling existed even after the travellers made manual corrections to their data using the web interface. Therefore, an automatic algorithm was developed to screen the data to identify and correct erroneous labels. A basic assumption is that individuals transferring from one transportation mode to another do so while stationary (as indicated by speeds close to or equal to zero). A series of (at least two) consecutive points with speeds below a certain threshold (v_{th}) is defined as a stationary sub-segment (*SS*). If the term Non-Stationary Sub-Segment (*NSS*) is defined to mean a portion of a transportation mode segments in which the trip maker is moving using any one of the available transportation modes, then it can be stated that *MTPs* should only occur during a stationary sub-segment when that sub-segment is immediately followed by a Non-stationary sub-segment labeled with a transportation mode which is different from the preceding *TMS* (these conditions are considered to be the *MTP validity criteria*). For convention, *MTPs* are designated to be defined as the last point in a stationary sub-segment.

Figure 4-7 illustrates a conceptual example for a trip with two transportation mode segments (*TMS*) where the first *TMS* is subdivided into two sub-segments i.e. non-stationary (moving) and stationary, respectively. The figure shows also the correct location and time for the *MTP* where the transfer to the new mode occurred.

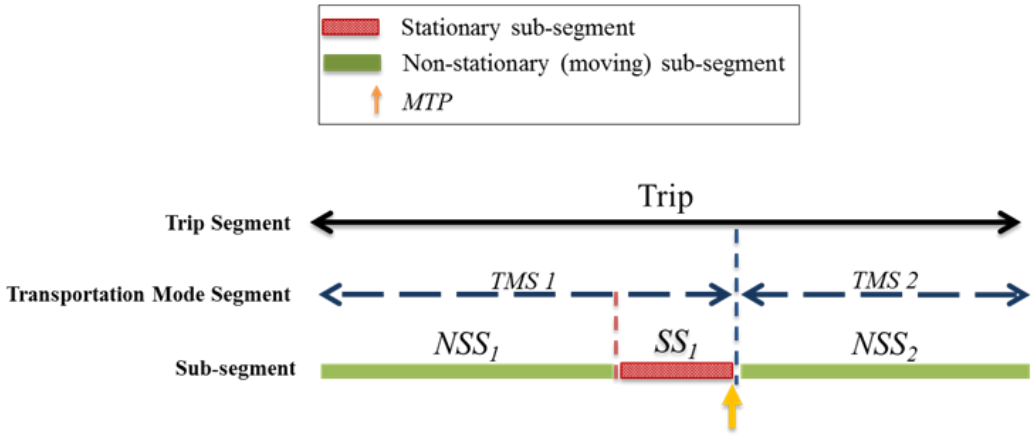


Figure 4-7 Conceptual example for a correct *MTP* location at end of stationary segment

Therefore, the expected possible cases where *MTPs* were misallocated by respondents can be summarized in the following points and also shown in Figure 4-8:

1. Users change the label earlier than when they should during the moving segment.
2. Users change the label earlier than when they should either at the end of the moving segment or beginning of the stationary segment.
3. Users change the label earlier than they should during the stationary segment.
4. Users change the label later than they should during the moving segment.

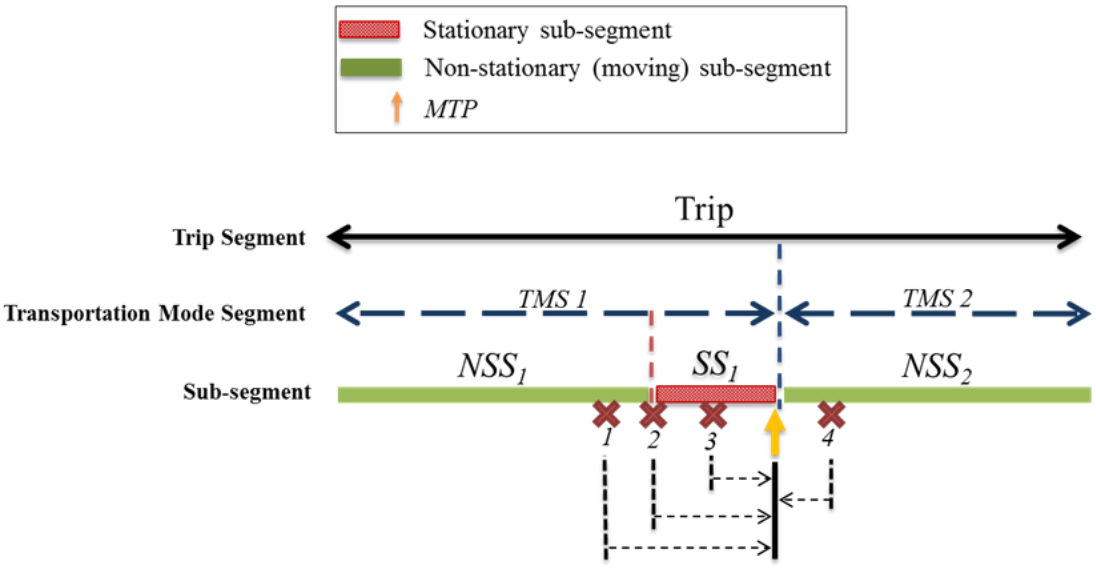


Figure 4-8 Conceptual example for the cases of mislabelling *MTP*

A closer look at the *MTPs* in Figure 4-6 suggests that some are appropriately labeled. *MTPs* numbered 1, 3, 5 and 8 occur when the traveller is at (near) zero speed and at the time that coincides with the end of a stationary sub-segment. However, *MTPs* numbered 2, 6, and 7 exhibit characteristics that are contrary to expectation. *MTPs* numbered 2 and 7 occur with speeds not equal to (near) zero and *MTP* number 6 does not occur at a time coinciding with the end of a stationary sub-segment. The algorithm aims to correct suspicious transportation mode transfer points by shifting the mode transfer label forward or backward along the time axis to the beginning or end of the adjacent stationary sub-segment. The proposed model starts with identifying the stationary and non-stationary sub-segments (*SS*) within each trip in the database, as shown with the hypothetical example in Figure 4-9.

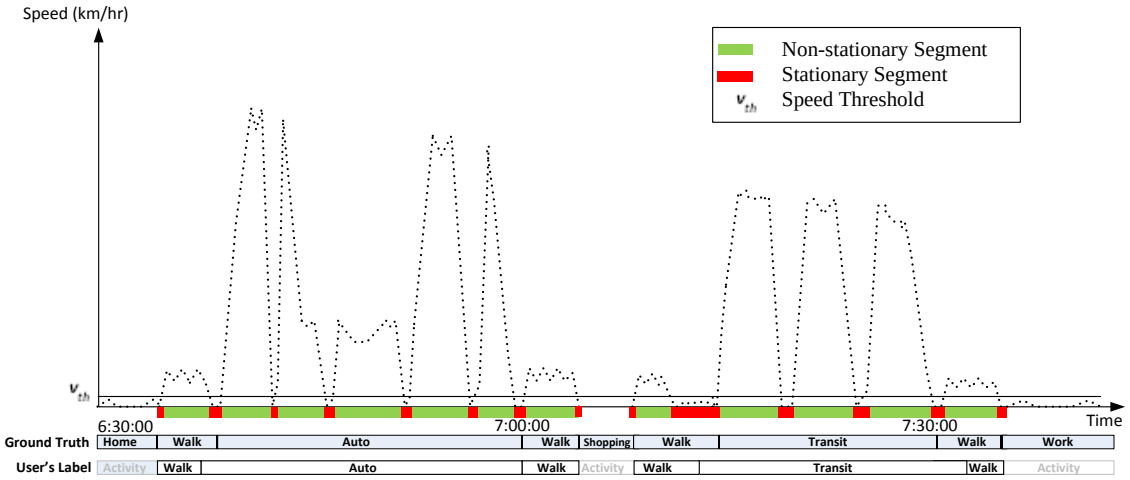


Figure 4-9 Segmentation of the hypothetical trips into stationary and non-stationary sub-segments

Then, all *MTPs* are identified and flagged if they are not correctly labeled (i.e. if they do not satisfy the validity criteria defined earlier). The algorithm corrects the flagged *MTP*s – labeled as points 2, 6, and 7 in the diagram – by shifting the mode transfer label forward or backward along the time axis to the end of the adjacent stationary segment (Figure 4-10).

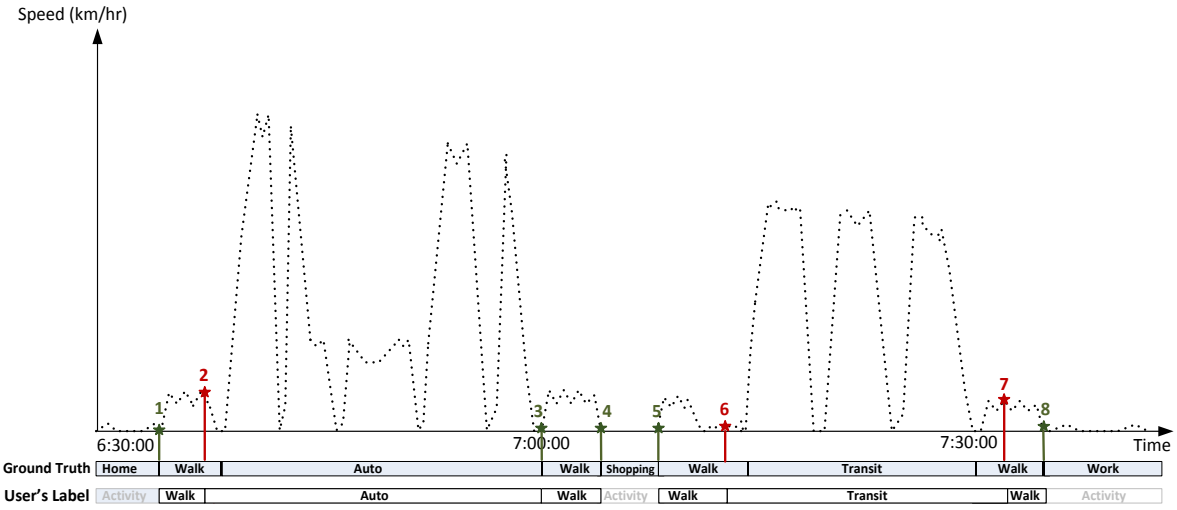


Figure 4-10 *MTPs* are identified and flagged if they are not correctly labeled

Point 2 is initially labeled as an *MTP* that occurs during a non-stationary sub-segment; this point is corrected by moving the *MTP* label to the last point in the subsequent stationary sub-segment. Point 6 is initially labeled as an *MTP* that occurs in a stationary sub-segment, but does not coincide with the last point in the sub-segment. Like point 2, point 6 is moved later in time to the last point in the stationary sub-segment. Finally, point 7 is corrected by moving the *MTP* earlier in time, to the end of the previous stationary sub-segment (or the beginning of the current non-stationary sub-segment). The final, corrected labels are shown in Figure 4-11. The application of the mode transfer point correction algorithm to the study data resulted in approximately 1,000 data points having updated *MTP* labels.

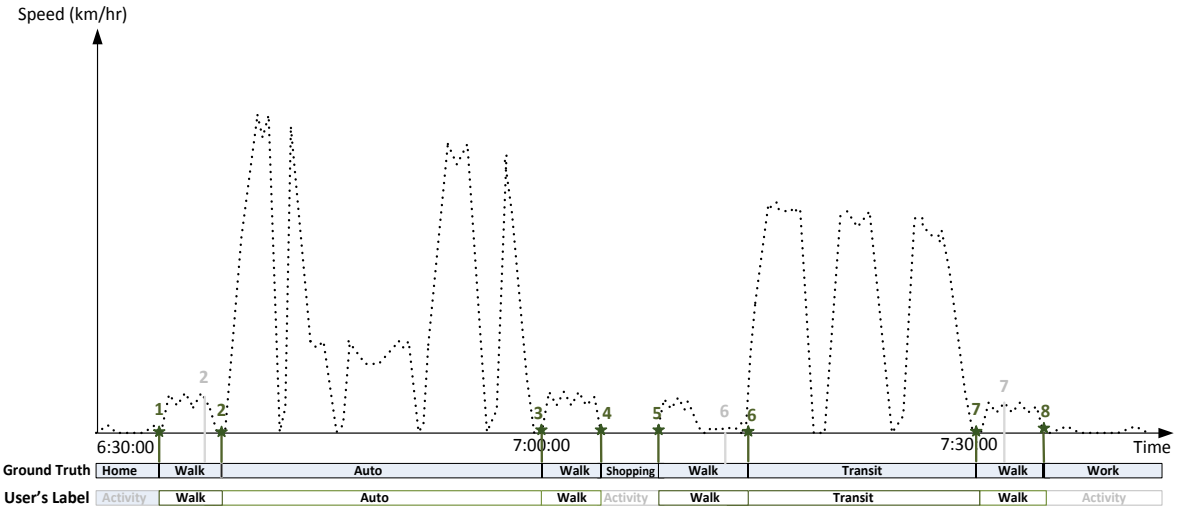


Figure 4-11 Outcomes of *MTP* correction algorithm

4.3 Data Pre-processing Results

The initial data set contained 642 trips. An initial cleaning effort eliminated duplicate points, very short trips and obviously erroneous data reduced the number of trips to 589. Further, detailed assessment of trips using the GUI developed for this research, identified a number of very long trips that were, in fact, multiple trips surrounding an activity. This assessment actually *increased* the number of total trips. The final step in the data pre-processing was to correct mislabelled mode transfer points as described above. The total number of re-labeled observations to new transportation mode is 1042 as shown in Table 4-1. At the end of the pre-processing stage, the data set contains 665 trips, 857 transportation mode segments and 105,638 points.

Table 4-1 Summary of the re-labeled observations to new transportation mode

TRANSPORTATION MODE	NUMBER OF RE-LABELED POINTS
WALK	92
BIKE	3
TRANSIT	79
AUTO	141
ACTIVITY	727

The impacts of the data pre-processing are shown in Figure 4-12. Figure 4-12a illustrates a boxplot showing the statistical properties of the original data, while Figure 4-12b shows the same properties after applying the filtering described in the previous sections. The shapes in the Figures should be interpreted this way:

- The red horizontal line represents the median value;
- The blue polygons are bound by the 75^{th} percentile and 25^{th} percentile on the top and bottom, respectively;
- The extended dashed lines to the whiskers extended to +/-2.7σ of the data;
- The red plus signs reflect individual data points in the extreme tails of the observations.

Two important observations should be made in comparing Figure 4-12a and 4-12b. First, the variability in the data for each transportation mode, reflected by the boundaries of the box plots in 4.12b, is less than in 4.12a. As a result, the overlap amongst ranges between modes is less, particularly in the case of transit and auto. Note that the overlap between the 75^{th} percentile boundary for transit and the 25^{th} percentile for auto is much less in Figure 4.12b than in Figure 4.12a. This reduction in overlap improves the ability to use this attribute to differentiate amongst the modes. The second observation is that the number of extreme values in each mode is substantially less in Figure 4.12b than in 4.12a. This is a result of the correction process where incorrect observations are now labeled correctly.

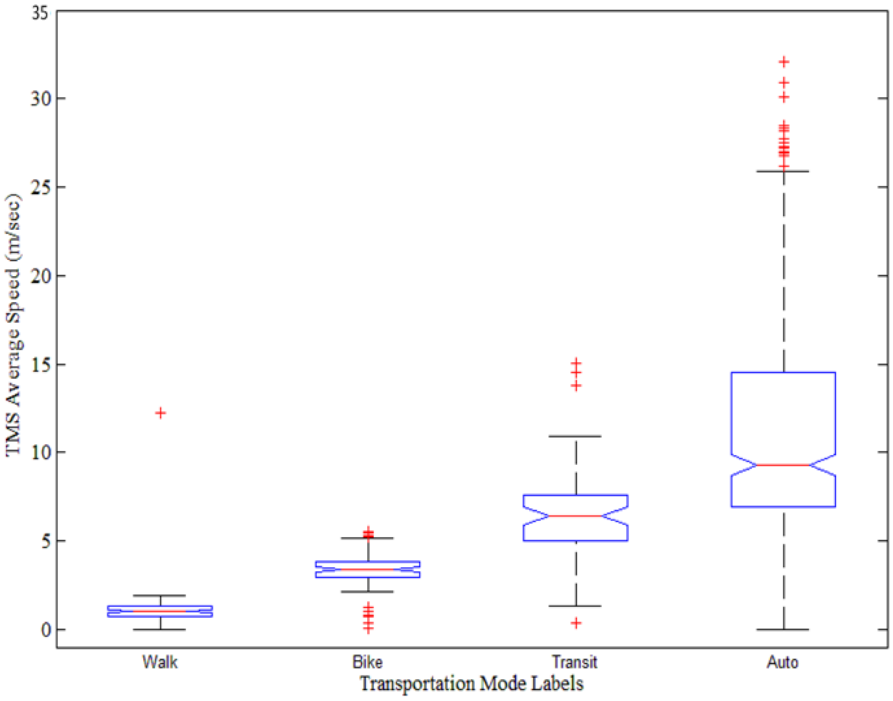


Figure 4-12a Original *TMS* average speed

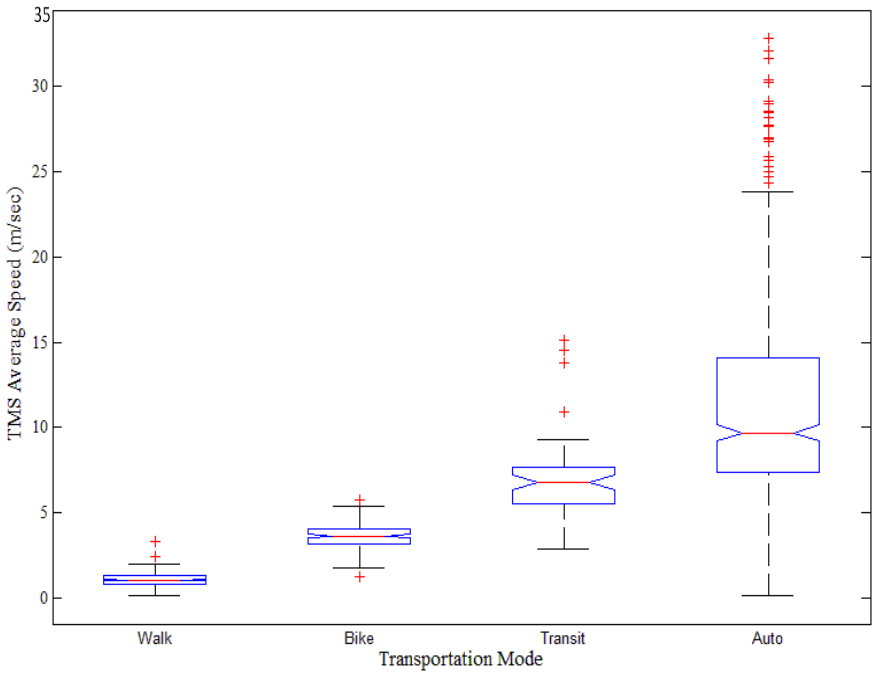


Figure 4-12b Filtered *TMS* average speed

Figure 4-12 The impact of data pre-processing on *TMS* average speed attribute

Table 4-2 summarizes the total number of observations and segments that remain in the data set after pre-processing.

Table 4-2 Statistics and information of pre-processed trajectories

Transportation Mode	Points		Transportation Mode Segments	
	Absolute #	% of total	Absolute #	% of total
Walk	20,028	18.96%	226	26.37%
Bike	14,393	13.62%	109	12.72%
Transit	8,993	8.51%	68	7.93%
Auto	62,224	58.90%	454	52.98%
Total	105,638		857	

Figure 4-13 and Figure 4-14 show the distributions of *TMS* distances and durations for each mode. It is from these data that the mode inference model is built and validated.

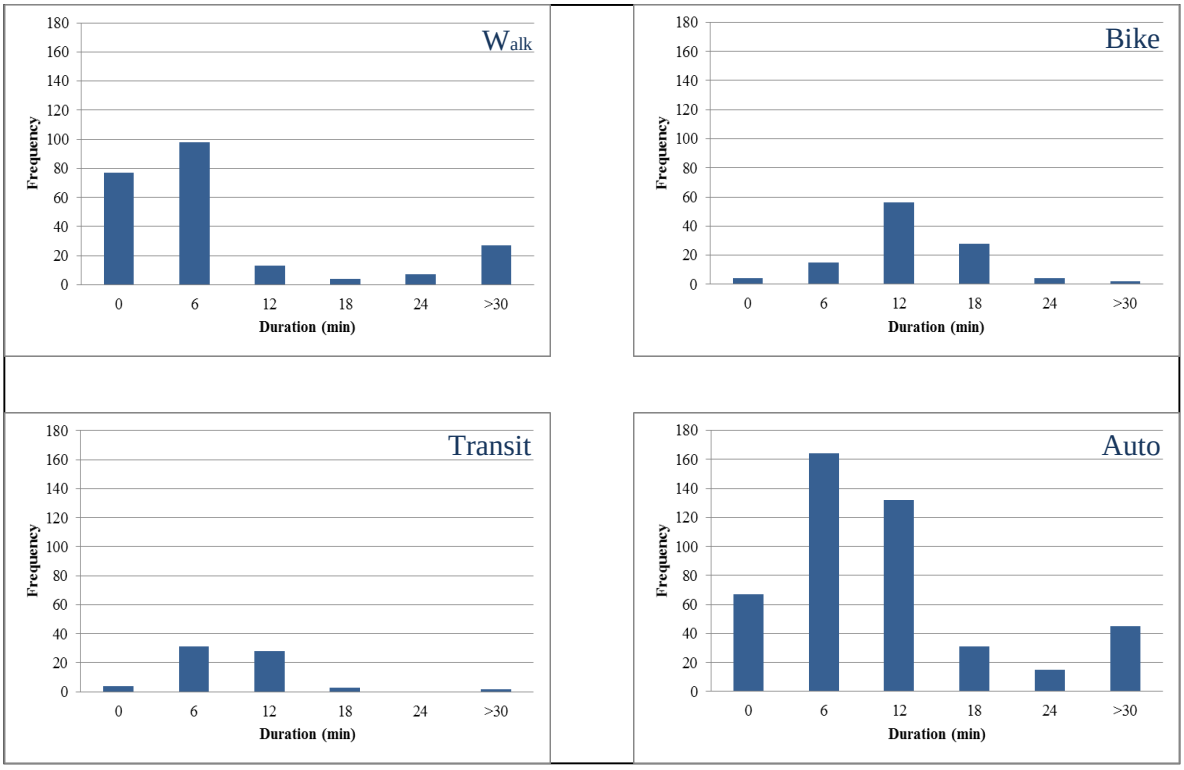


Figure 4-13 The distribution of *TMS* durations

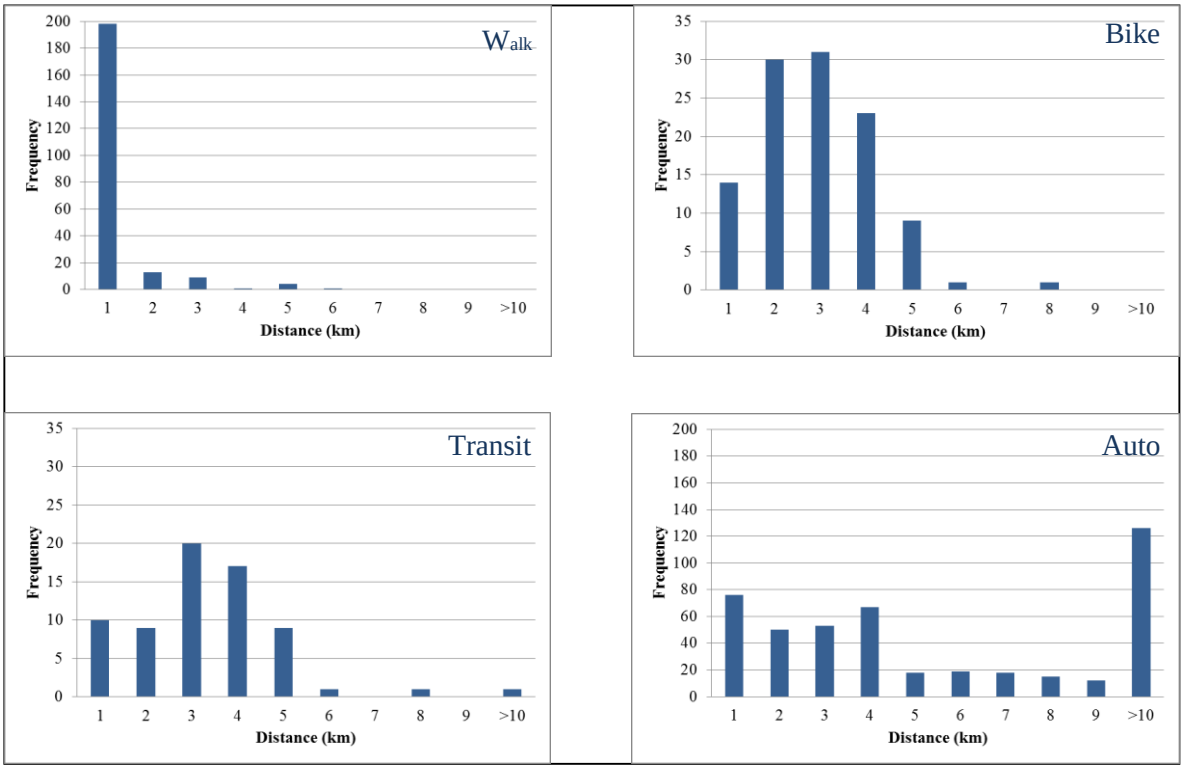


Figure 4-14 The distribution of *TMS* cumulative distance

4.4 Summary

A custom software module was developed to automatically retrieve, process, and re-store sequential GPS data for further analysis. One of the main purposes of this section is to highlight the importance of exploring and understanding the nature of the data to the overall goal of developing a highly accurate classification model. The developed software enables the process of exploring and investigating suspicious and possibly invalid trips. In addition, a naïve criterion based algorithm was developed to eliminate the invalid trips from the database and to correct the mislabelled data which have been missed by survey participants. Moreover, due to participants' subjectivity in specifying the *MTP*, another an algorithm has been developed to systematically correct the location and time of *MTPs*.

The output from the work described in this chapter is a collection of valid trips that are correctly labeled. Beginning with this data set reduces the effect of random error on the overall performance of the proposed mode inference model.

Chapter 5
Transportaion Mode Classification Model

Recall that the overarching goal of this research is to use GPS data to determine the transportation mode of travel. The result of the previous chapter is a data set of cleaned and verified trips. A trip is defined as a time series of GPS points bounded by two activities. Within each trip, the boundaries of travel by a given transportation mode are known from user inputs. Transportation Mode Segments (*TMS*) are defined as portions of a trip that contain sequential data points labeled as having been completed by the same mode. A trip may be comprised of a single mode, in which case the entire trip will consist of one *TMS*. Alternatively, a trip may involve multiple transportation modes (and therefore multiple *TMSs*), in which case one or more Mode Transfer Point *(MTP)* must exist. Therefore, the proposed methodology begins with identifying all *TMS* in the database. The result of this step is a database of segments that are defined by a (GPS) beginning point, end point, and a verified mode label. This database is the input for the work presented in this chapter – the development and implementation of a method to improve upon previous solutions to the mode identification problem.

The proposed approach builds upon previous work but introduces several novel methodological advancements. Generally, the methods are made more robust by limiting the *a priori* assumptions about the data. Instead, a more holistic approach is taken to the analysis that allows the data to determine the appropriate inputs and methods to infer transportation mode.

More specifically, in this work the assumptions around the order of modes used for a given trip and the duration associated with each mode are relaxed. Instead, a technique that identifies sub-segments – both non-stationary (moving) and stationary – is deployed. For the moving sub-segments, the most likely modes for each sub-segment are determined based on segment characteristics that can be computed from the available GPS data: maximum speed, average speed, acceleration rate, and jerk (rate of change of acceleration). Unlike previous researchers, however, travel segments are not classified as belonging to a mode based on a single, weighted combination of travel characteristics. Instead, a classifying algorithm is used that iteratively

introduces attributes from a feature vector such that those characteristics with the strongest differentiating powers between transportation modes are identified and used in the mode classification.

The proposed statistical method is based on a machine learning model. In the literature, researchers have attempted to identify the circumstances under which a given model produces the most accurate mode identification. In this research, multiple regression and factorial analysis have been used to investigate an exhaustive combination of factors to calibrate the classification model in order to optimize the model parameters. This allows the developed framework to produce more robust results for widely varying data sets.

The evaluation method used to assess the performance of the model also differs from previous work. In this research, the model is trained using the user-provided labels that define the beginning point, end point and mode that constitute a transportation mode segment. In the evaluation process, these data are not used (because in practice when attempting to classify unlabelled data these points are not known). Instead, the beginning and end points of a more disaggregate segment definition – non-stationary sub-segment (speed greater than a threshold) and stationary sub-segment (speed less than a threshold) are identified. Then, the mode classification model to infer the transportation mode of non-stationary sub-segments is applied. In essence, this presents the classifier with a "new" data set – the same data but at a more disaggregate level – than what is used to train the model.

The initial sections in this chapter describe the range of options for developing a mode inference model. Given this background, a method is presented by which the optimal combination of these options can be determined automatically, based on the input or training data. Finally, the optimal model for the data that informs this work is presented. Figure 5-1 shows the full research framework introduced earlier, emphasizing the steps included in the transportation mode classification model.

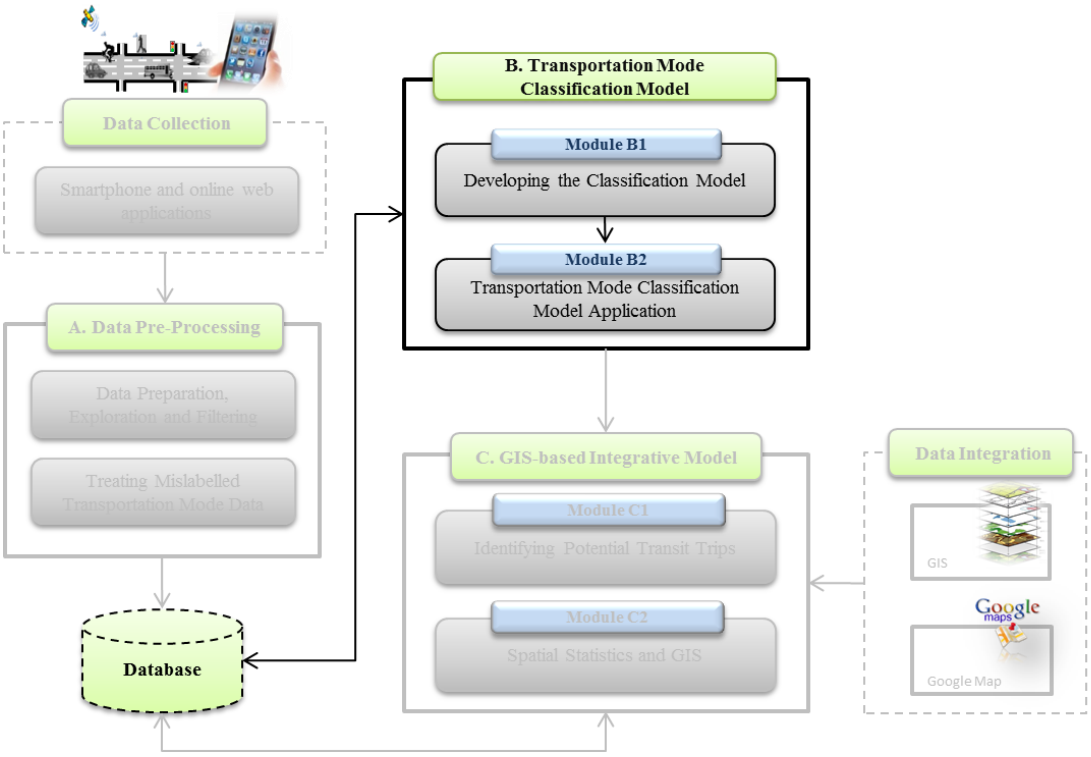


Figure 5-1 Proposed Transportation Mode Classification Framework

(B. Transportation Mode Classification Model)

5.1 Developing the Classification model

As mentioned earlier, transportation mode inference can be solved as a pattern recognition problem using a number of supervised learning models (classifiers). The main goal of any supervised learning model is to build (train) a classifier using a known set of input data (feature vector) and known responses (labels or classes). This "trained classifier" is then capable of generating reasonable predictions for travel mode in response to new data.

To create the model, several initial decisions must be made. These include:

- Determining which attributes of transportation mode segments (*TMS*) are included or excluded from the model;
- Identifying and selecting amongst commonly used machine learning classification techniques;
- Identifying possible data formats for the inputs to the model;

Each of these steps, shown in Figure 5-2, are described in subsequent sections.

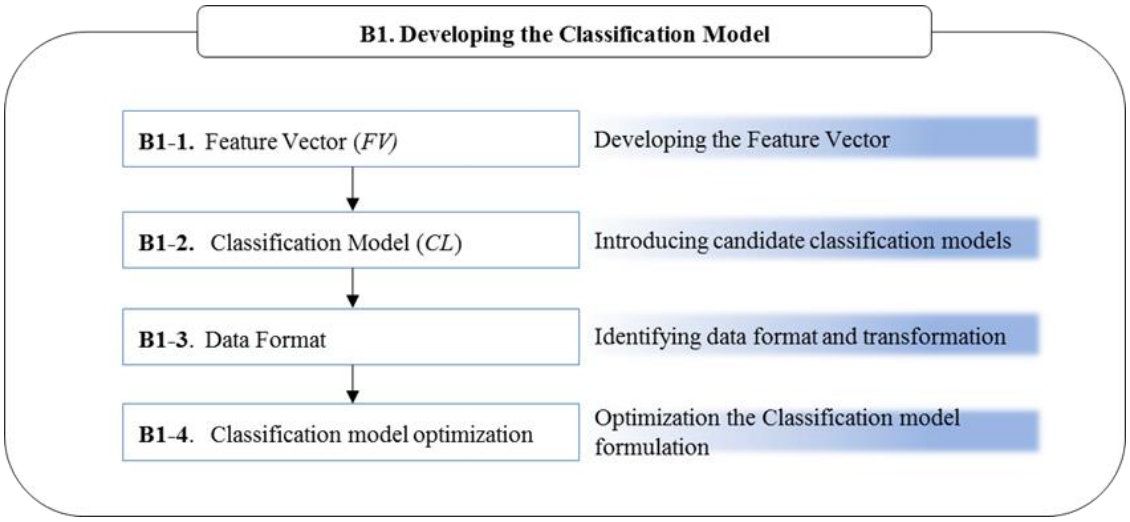


Figure 5-2 The framework of Developing the Classification Model Module

5.1.1 Developing the Feature Vector

For all trips in the data set, the transportation mode segments belonging to mode *m* are extracted amongst a set of candidate modes *M*. Four candidate modes exist: private auto, public transit, cycling and walking. To quantify the characteristics of travel by a given mode, a Feature Vector for mode *m* (FV^m) is developed that contains a set of attributes that are computed from the GPS data. For example, average speed, maximum speed, maximum acceleration, and jerk (the rate of change of acceleration) can be calculated for each mode from the GPS data. Mathematically, a feature vector of dimension *n* for attributes a_i can be defined for a single transportation mode segment *TMS1*:

$$FV^m = \{a^m_{1,TMS1}, a^m_{2,TMS1}, a^m_{3,TMS1} \dots a^m_{n,TMS1}\} \quad \textbf{(5-1)}$$

Alternatively, distributions of values can be developed for each attribute *a* of a given mode *m* when observations are made across all segments, *TMS*. Figure 5-3 shows two hypothetical distributions for attribute *A* developed for two modes, *m=1* and *m=2*. It is also possible to calculate mean values for an attribute belonging to a given mode.

Again, these definitions can be formulated mathematically. Let TMS^m define the number of segments of mode *m*. Then the mean value for an attribute A_1 is given by:

$$\overline{A^m_1} = \frac{1}{TMS^m} \sum_{y=1}^{TMS^m} a^m_{1,TSMSy} \quad (5\text{-}2)$$

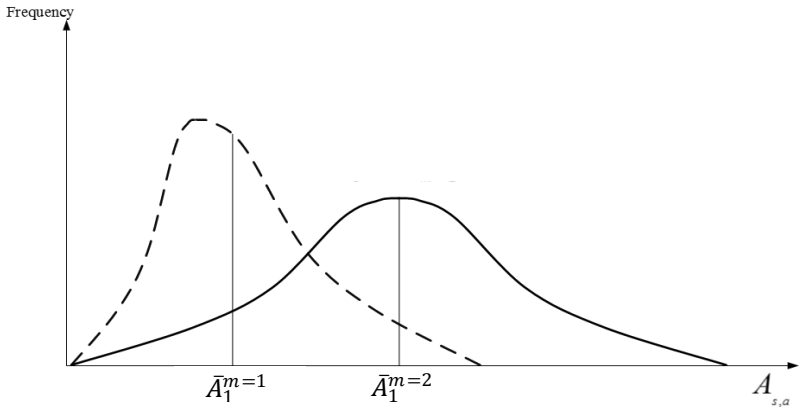


Figure 5-3 Hypothetical distributions for attribute A for modes, m=1 and m=2

Initially, a long list of potential attributes was developed, including multiple metrics of speed, acceleration, and jerk. The full list is shown in Table 5-1. To improve the effectiveness of the mode identification process, it is beneficial to understand which quantitative attributes of transportation mode segments have the greatest differentiating power between modes. For example, it may be reasonable to expect that maximum speed can be the most effective attribute to determine if a trip segment is made by a motorized versus a non-motorized mode.

Table 5-1 The initial attributes considered in the Feature Vector

Speed	**Acceleration**	**Jerk**
Average	Average	Average
Maximum	Maximum	Maximum
Minimum	Minimum	Minimum
Standard Deviation	98th Percentile	98th Percentile
98th Percentile	98th -50th Percentile	98th -50th Percentile
98th -50th Percentile		

Two of the major contributions of this research are to:

1. develop and implement a formal, well-constructed method to determine which attributes to include and exclude from a mode's feature vector; and
2. sort and weigh these attributes based on their differentiating power – i.e. the strength of the attribute in positively identifying belonging to one mode or excluding other, infeasible modes.

To formalize the process of quantifying the differentiating power of an attribute the following approach is taken. For each attribute, a student *t* test (Case 3: two samples with unequal sizes and unknown variances) is calculated to conduct comparisons between the means of distribution to determine if the two sets are significantly different from each other. The resulting p_{value} quantify the probability that the mean value for attribute a for mode 1, $\bar{a}^{m1}$ is equal to the mean value for the same attribute for mode 2, $\bar{a}^{m2}$.

Mathematically, the *t* statistics are calculated for the means of attribute *a* for modes m_1 and m_2: $t_a^{m1,m2}$. The degrees of freedom for the problem, *df*, are then calculated such that the comparison can be made between the calculated *t* statistic and the critical value. In this comparison, if $\left|t_a^{m1,m2}\right| > |t_{crit}|$, the null hypothesis (H0: both samples come from populations with equal means) is rejected at α = 5% confidence level, suggesting the selected attribute is suitable for differentiating between the two transportation modes.

Simultaneously, once a *t* value is estimated, a p_{value} - a scalar value in the range [0, 1] - can be also determined using the Student's t-distribution table. The p_{value} is the probability of observing a test statistic as (or more) extreme than the observed value under the null hypothesis. The statistic (1-p_{value}) is used to quantify the differentiating power (*DP*) of a feature vector attribute. Mathematically:

$$DP_a^{m1,m2} = \left(1 - p_a^{m1,m2}\right) \tag{5-3}$$

Given four transportation modes, $\binom{4}{2}$ or six pairwise comparisons can be calculated for a given attribute as shown in Table 5-2. Ultimately, the goal is to compute a single attribute differentiating power *(ADP)* statistic for each attribute. A simple way to convert the six DP_a values to a single ADP_a value is to take an average. Some pairwise comparisons, however, are more important than others, specifically those attributes that differentiate between motorized and non-motorized modes.

Taking this into account, a new variable, β, is defined. The purpose of this variable is to act as a calibration tool in computing the weighted average of the *ADP* statistic for each attribute *a*. Mathematically, β is included in the *ADP* computation as follows:

$$ADP_a = \frac{\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\left(DP_a^{m_1,m_2}.\beta^{m_1,m_2}\right)}{\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\left(\beta^{m_1,m_2}\right)} \tag{5-4}$$

Where:

β^{m_1,m_2} is the weighting associated with the comparison between mode m_1 and m_2; the default value of β^{m_1,m_2} = 2 for all bolded *DP* as shown in Table 5-2 and $\beta^{m_1,m_2} = 1$ otherwise.

To test the sensitivity of β, *ADP* was computed using values from $\beta=1$ to $\beta=10$ at 0.5 increments. The result of this iterative test was that the value of β failed to significantly influence the ordering of the attributes in the Feature Vector.

Table 5-2 Differentiating Power for all pairwise comparisons

	Walk	**Bike**	**Transit**	**Auto**
Walk		$DP_a^{m1,m2}$	$\boldsymbol{DP_a^{m1,m3}}$	$\boldsymbol{DP_a^{m1,m4}}$
Bike			$\boldsymbol{DP_a^{m2,m3}}$	$\boldsymbol{DP_a^{m2,m4}}$
Transit				$DP_a^{m3,m4}$
Auto				

ADP_a is computed for all attributes. The attributes are then sorted based on their *ADP*; the one with the largest value for *ADP* has the strongest differentiating power, and is ranked as 1st. Amongst *n* attributes, the feature with the lowest *ADP* will have the weakest differentiating power and will be ranked n^{th}.

This initial feature ranking based on *ADP* provides a good, but imperfect ordered list of attributes due to the presence of correlation amongst attributes that can result in skewed differentiating power estimates. Consider the example where maximum speed is a very strong differentiator amongst modes. In this case, it is probable that 98th percentile speed will also have strong differentiating power. But, there is obvious correlation amongst those two attributes. An improved, hierarchical list can be created by accounting for these correlations amongst attributes.

The process of managing correlation begins by classifying attributes into two categories – candidate attributes (initially all those attributes that have been identified) and chosen attributes (initially a null set). The first step is moving the highest ranking attribute from the candidate set into the chosen set. For all remaining attributes, the average correlation between the attribute under consideration in the candidate set and the attribute in the chosen set is

calculated. Using this correlation, the Adjusted Attribute Differentiating Power (*AADP*) Score, is calculated for each attribute in the candidate set as follows:

$$AADP_a = ADP_a(1 - \alpha\rho_a) \quad (5\text{-}5)$$

Where:

α is a user-defined parameter to determine the sensitivity to correlation; and

ρ is the average correlation between each candidate feature and the features in the chosen set.

The attribute in the candidate set with the largest *AADP* score is selected and moved to the chosen set. The process is repeated until all attributes are ranked from highest differentiating power to lowest differentiating power in the chosen set, and the candidate set is empty.

The contribution of the *ADP* and *AADP* is very important. In essence, using this approach eliminates the need to select *a priori* the attributes to be contained in the Feature Vector. Instead, attributes are included in the Feature Vector based on their differentiating power. Each set of data can be analyzed to generate a new, more appropriate Feature Vector customized for that data in the mode inference model; this is a significant improvement over previous approaches. The strength of the *ADP* in producing improved mode inference results is demonstrated quantitatively in section 5.3.

5.1.1.1 Assessing the impacts of *ADP* on Model Performance

As mentioned previously, the *ADP* algorithm was introduced to sort the attributes in the feature vector based on their differentiating power. To demonstrate the importance of the *ADP* technique, the relationship between the classifier accuracy and *ADP* has been investigated; the model was run with an un-informed selection of attributes in the feature vector. In the first case, the model was run 16 times, each time with a single attribute in the feature vector. In the next case, each individual run included 4 attributes; the attributes were selected such that the average *ADP* of these attributes was monotonically increasing.

The model's overall accuracy was calculated for each individual run. Figure 5-4 illustrates the accuracy results for the 16 runs as a function of *ADP*. The relation between classifier accuracy and *ADP* was statistically investigated through a linear regression.

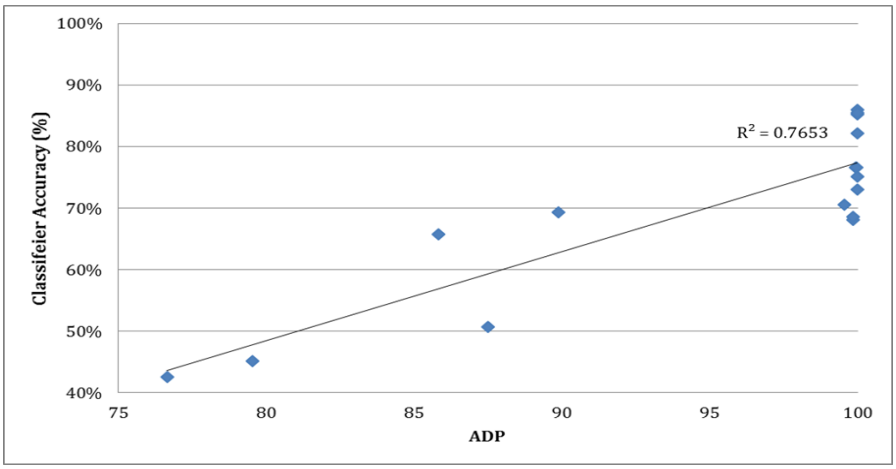


Figure 5-4 Case 1: relationship between classification model accuracy and *ADP*

Table 5-3 summarize the regression analysis results for the model and its components. The model is significant at the 5% significance level with a *p*-value of 9.21E-06. The coefficient of determination (R^2) of 0.7653 means 76% of the variation in the classifier accuracy is explained by the independent variables – *ADP*. Both the intercept and slope coefficients are statistically significant. The slope coefficient of 0.0145 indicates that for every unit increase in *ADP* an increase of 1.45% is expected in the classifier accuracy.

Table 5-3 Regression anlysis results for case 1

Regression Statistics	
Multiple R	0.8748
R Square	0.7653
Adjusted R Square	0.7486
Standard Error	0.0681
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.2116	0.2116	45.6607	9.2104E-06
Residual	14	0.0649	0.0046		
Total	15	0.2765			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.6740	0.2042	-3.3011	0.0053	-1.1119	-0.2361	-1.1119	-0.2361
X Variable 1	0.0145	0.0021	6.7573	0.0000	0.0099	0.0191	0.0099	0.0191

For case 2, the results are plotted in Figure 5-5 as a function of the average *ADP* associated with the chosen set of four attributes in each run. The relation between average *ADP* and

classifier accuracy was statistically investigated through a linear regression providing a coefficient of determination (R^2) of 89.5%.

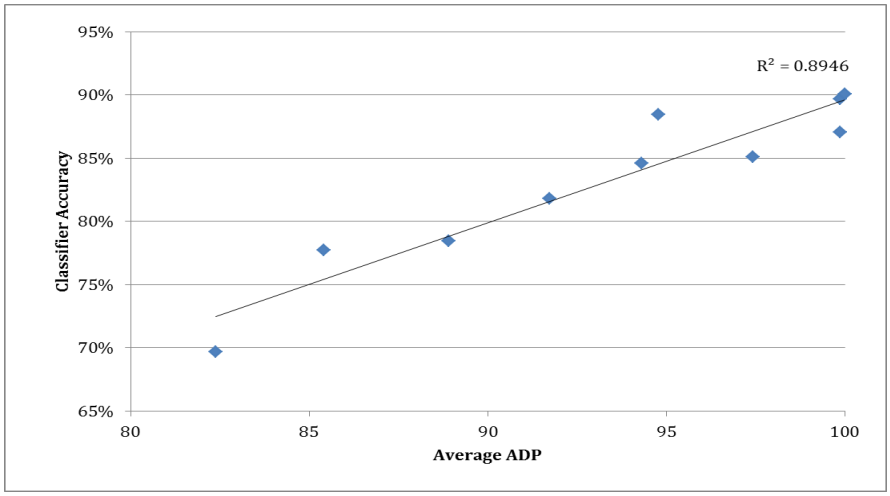


Figure 5-5 Case 2: relationship between classification model accuracy and average *ADP*

Table 5-4 summarize the regression analysis results for the model and its components. It should be noted that the intercept was not statistically significant and therefore the model was re-calibrated forcing the intercept to be equal to zero. The model is significant at the 5% significance level with a *p*-value of 3.53E-05 and the slope is statistically significant. The same conclusion can be made that the classification accuracy linearly increases as average *ADP* increases. Table 5-5 illustrates the combination of attributes in the four attribute runs and their average *ADP*.

Table 5-4 Regression analysis results for case 2

Regression Statistics	
Multiple R	0.9997
R Square	0.9994
Adjusted R Squar	0.8883
Standard Error	0.0216
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.9642	6.9642	14969.8701	2.22593E-14
Residual	9	0.0042	0.0005		
Total	10	6.9684			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
ADP	0.0089	0.0001	122.3514	8.2665E-16	0.00875	0.00908	0.00875	0.00908

Table 5-5 *ADP* effectiveness evaluation in the classification model performance

	Features																
	Avg (V)	StDev (V)	98^{th} Perc. (V)	Max (V)	98^{th}-50^{th} Perc. (A)	98^{th} Perc. (A)	Max (A)	Min (A)	98^{th} Perc. (J)	98^{th}-50^{th} Perc. (J)	Max (J)	98^{th}-50^{th} Perc (V)	Min (V)	Min (J)	Avg (A)	Avg (J)	Average ADP
Run # \ADP	100.00	100.00	100.00	100.00	99.99	99.99	99.97	99.95	99.86	99.85	99.56	89.90	87.50	85.82	79.55	76.65	
1													*	*	*	*	**82.38**
2											*			*	*	*	**85.40**
3										*	*				*	*	**88.90**
4							*		*				*		*		91.72
5							*		*			*	*				94.31
6					*			*		*		*					97.42
7				*			*	*			*						99.87
8				*				*			*				*		94.76
9			*			*			*		*						99.85
10	*	*	*			*											100.00

* Chosen Features

5.1.2 Introducing candidate classification models

Transportation mode inference can be thought of as a pattern recognition problem that can be solved through a number of supervised learning models. Two methods are commonly used to select an appropriate classification technique. The first method involves pre-selecting a classifier based on the characteristics of the data and on the problem definition – see for example Chung and Shalaby (2005). In the second approach, multiple classifiers are trained and tested and the best performing method is selected. Before this stage can be completed, a set of possible classifying models must be identified. Candidate models for this work are introduced here, as shown in Figure 5-6.

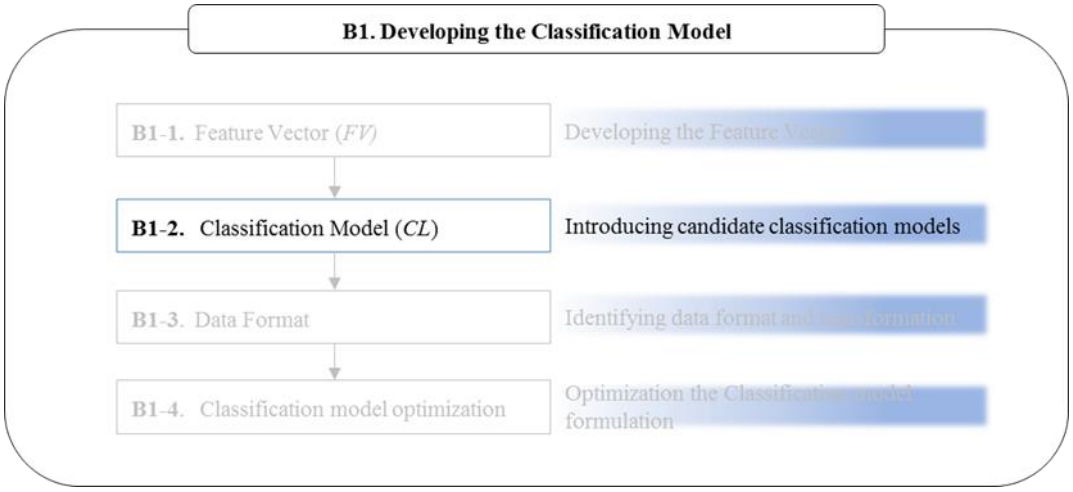


Figure 5-6 Developing the Classification Model (B1-2. Classification Model)

As described in chapter 2, many machine learning techniques exist. From this wide range of models, this research employs three techniques that are suitable to demonstrate the strength of the approach, but also have sufficient flexibility and ease of interpretation of results. The considered models are: Naïve Bayes (NB), Quadratic Discriminant Analysis (QDA), and k-Nearest Neighbor (k-NN). Each of these techniques is briefly described here.

NB classifier is a simple probabilistic method that provides classification results as a probability distribution (degree of certainty) over a set of classes (Russel *et al.*, 2003). The Bayes rule is applied to compute the *posterior* probability of class $c \in X_C$ given the feature vector for a particular instance of $x_1, x_2, \ldots x_N$ as follows:

$$P(X_c|x_i) = \prod_{i=1}^{N} P(x_i|X_c) \quad (5\text{-}6)$$

Conceptually, Naive Bayes generates a higher probability of belonging to a class when: the observed data point has a higher conditional probability of belonging to that class; and when more data are observed from that class. In this case, NB returns higher probability when, for example, the observed speed is consistent with a mode's speed range, and there are significant number of data points belonging to that mode.

The second candidate classifier is Quadratic Discriminate Analysis (QDA), a multivariate statistical technique widely used in pattern recognition and machine learning (Lachenbruch, 1936 and Russel *et al.*, 2003). In QDA, a class discrimination model that separates different classes by a quadratic surface is built. In order to classify a new sample, the trained classifier finds class *c* which maximizes the quadratic discriminate function (Equation 5-7).

$$\hat{c}(x_i) = \text{argmax}_c \, P(x_i \mid c) \tag{5-7}$$

$$P(x_i \mid c) = \frac{1}{(2\pi \mid \Sigma_c \mid)^{1/2}} . \exp\left(-1/2(x_i - \mu_c)^T \Sigma_c^{-1} (x_i - \mu_c)\right) \tag{5-8}$$

where x_i is the feature matrix for the new case;

Σ_c is the covariance matrix for class *c*

Conceptually, the goal of discriminate analysis is to divide an analysis space into two (or more) categories such that the distance from observed data to the dividing (hyper) line (in a linear case) or polynomial (in a quadratic case) is maximized.

The third candidate of the classification models is k-Nearest Neighbors (k-NN). Basically, k-NN calculates the distance between the *k* nearest neighbors of training cases to the new sample in the feature space. While various distance metrics can be used, Mahalanobis distance may be the most appropriate. Mahalanobis distance was chosen as the distance metric as it considers both the variance and covariance of the feature vector variables in order to measure the degree of similarity between the feature vector of the current data and the training data. The consideration of the variance reduces the effect of those variables with high variations while the common effects of correlated variables can be excluded by considering the covariance of the feature vector variables. The Mahalanobis Distance, *MD*, is calculated as:

$$MD_k = \left(\left(x_h^k - x_i^k\right) . \sum_{(X^H)}^{-1} . \left(x_h^k - x_i^k\right)^T \right)^{1/2} \tag{5-9}$$

Where:

MD_k is the Mahalanobis distance between training cases *h* and new instance *i*

x^k is the feature matrix, and

$\Sigma_{(X^H)}$ is the covariance matrix of training feature variables.

After estimating the distance metric, the new data point is classified based on the class that has the majority from *k* neighbors. Therefore in order to avoid ties, *k* is commonly selected as an odd number. In addition, it is logical to discriminate between the *k* nearest neighbors in the prediction in respect to their relevant distances from the new point. Therefore, a weighting function is introduced to increase the influence of the closer neighbors, as defined by Shepard (Shepard, 1968):

$$w_{\bar{i}} = \frac{1}{MD_k^2} \tag{5-10}$$

5.1.3 Identifying Possible Data Formats

In addition to these three statistical techniques, and the parameters that define them, the performance of classifying algorithms is also influenced by the data format. The features in classification learning problem can be either continuous or categorical. Features measured along numerical scales are referred to as "continuous" features. Features expressed with unordered values or defined by qualitative classifications are called "categorical" features. In this research, in order to transform the continuous features to nominal (discrete) features, a supervised discretization following MDL method developed by Fayyad and Irani (1993) is employed that utilizes the class labels to partition the range of the attributes to at least two subranges.

Consider the example shown in Figure 5-7. In the diagram, observations of speed are plotted on the horizontal axis, with corresponding mode labels as illustrated. The speed attribute in Figure 5-7 is partitioned into two discrete categories at a partition boundary – in this case about 8 m/s. Given this transformation, all subsequent conditional probabilities will not be calculated based on a continuous value of speed, but rather based on the belonging (or not) to a speed category. This approach reduces the number of observations with very low conditional probabilities and emphasizes fewer observations with stronger correlations.

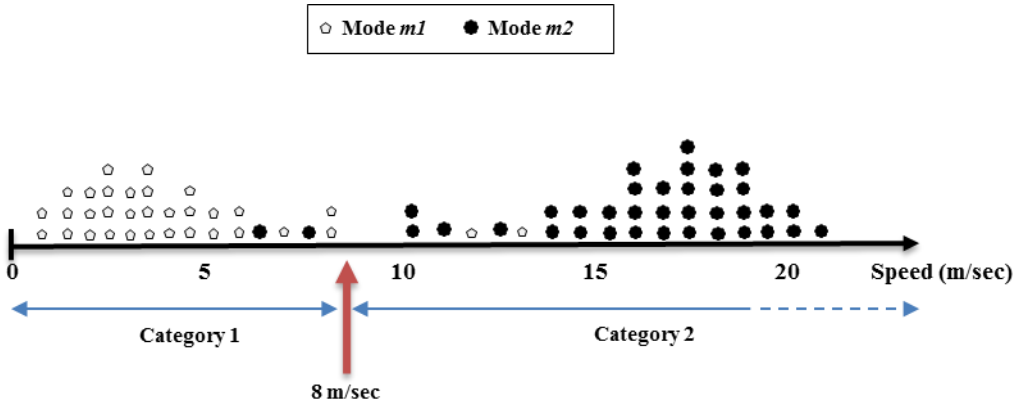


Figure 5-7 Hypothetical example of Speed attribute supervised discretization

Another potential problem with these techniques is in the redundancy that can occur with when a large number of attributes are included in the feature vector. A large set of attributes can sometimes create situations where irrelevant attributes dominate more relevant features, negatively impacting the performance of the model. This condition, referred to as high dimensional data, can be improved by reducing dimensionality. A common technique to achieve this outcome is with Principle Components Analysis (PCA). Conceptually, PCA is a statistical technique that recasts data in new coordinate systems to determine levels of variance amongst attributes. Cumulative variance plots are created that define the Principle Components; these new statistical representations of attributes replace the original attributes contained in the feature vector. PCA has been shown to have mixed results – in some cases enhancing the performance of the classification model while in other cases PCA produces poorer results.

5.1.4 Summary of Model Formulation Options

The process of developing and testing a mode classifier is complex and diverse. Several options exist in creating the feature vector, determining the classifying technique, and managing the data. To summarize, decisions to be made in mode identification include:

- Computing the Differentiating Power *(DP)* amongst attributes. Evidence exists that beyond a certain number of attributes, prediction accuracy begins to decrease due to conflicting signals from the input data. The approach taken here is to initially order all possible attributes based on their strength in differentiating amongst modes;

- Determining appropriate weights *(β)* when computing the weighted average amongst attributes' differentiating power (*ADP);*
- Addressing correlation amongst features in feature selection process (α): the data used to identify mode of travel – speeds, acceleration, and jerk – tend to be highly correlated. As a result, when creating the ordered *ADP* list above, the potential exists to rank an attribute very highly not because it independently has strong differentiating power, but it has strong correlation to a variable with strong differentiating power. As such, a method is necessary to manage correlations amongst variables, producing an adjusted attribute differentiating power variable, *AADP*.
- Creating the final Feature Vector *(FV)* with the appropriate, ordered list with the desired Number of Features *(NF).*
- Classification techniques (*CL*): NB and QDA have the positive attributes of relatively easy implementation and interpretable results. In many cases, NB has produced the best results. k-Nearest Neighbors is slightly more complex to implement, but is generally more robust, while still producing interpretable results.
- The representation of the data (*Disc*): the data can be either continuous variables which is represented by continuous statistical distributions or discrete variables. In some machine learning models, the nature of the data has a significant impact on the model's performance. For instance, *NB* performs better with discrete features; on the other hand, DA performs better when using continuous attributes.
- The use of Principle Components Analysis (*PCA*): in some cases, PCA methods are used to reduce the dimensionality of data. The results have been mixed, with PCA improving predictive power in some cases, but performing less well in other cases.

5.1.5 Optimizing the classification model formulation

Given the breadth of options in the classification process, and the lack of *a priori* knowledge of what may produce optimal results, the classification process was automated to iteratively test all permutations. The classification model's performance is influenced by varying:

1. α: the degree to which correlation in ranking Feature Vector attributes is considered. This parameter varies from no consideration to full consideration over six levels {0,0.2,0.4,0.6,0.8,1.0}; (6 Levels)

2. *NF*: the number of features used to classify modes. This set of features may contain {1,3,5,6,7,8,10,11} variables; (8 Levels)
3. *CL*: the classifier model: {1: NB, 2: k-Nearest Neighbor, 3: DA}; (3 Levels).
4. *Disc*: feature Discretization {0,1} binary variable (0→ Continuous, 1→ discrete); (2 Levels);
5. *PCA*: the use of PCA {0,1} binary variable (0→ when original data is used, 1→ when PCA is applied ; (2 Levels);

An experimental design method is adopted to determine the optimal value for each parameter. In order to have a better estimate for the variance of the error, the model is run with all permutations of the inputs listed above. In each model run, the classifier is trained using a subset of data – establishing ranges for variables that represent a known mode. Then, the trained model is used to classify unseen data on transportation mode segments. The model's performance is quantified by calculating the misclassification rate (*MCR*) – the number of transportation segments incorrectly labeled divided by the total number of segments. This technique produces multiple independent *MCR* for each permutation.

Therefore, a more robust vetting of the model can be accomplished using a technique known as stratified cross validation. The data are split into sections, *NK* folds, such that each section contains roughly the same proportion of different modes; then, iterations are performed with the folds as follows:

1. Train the model with folds 2 to *NK* and test the model using the data in the first fold. Then, the Misclassification Rate (MCR) is estimated.
2. Train the model with folds 1 and 3 to *NK* and test the model using the data in fold 2; then, calculate the *MCR*.
3. Iterate for *NK* and accumulate the *MCR* for all iterations.

This results in *NK* independent observations of *MCR* for each combination. Using the range of results, a linear regression model is calibrated. Essentially, *MCR* observations are regressed against binary variables representing each parameter level. This approach allows the combination of feature vector composition, parameter values, data transformation and statistical classification technique to be determined *a posteriori* that optimizes the performance of the model.

The experimental design method helps in investigating both main effects and interactions between the examined factors. In addition, it provides insight as to whether or not the examined factors are statistically significance. Therefore, this approach results in a significantly more flexible modeling method than previous work, and increases the model's performance.

5.1.6 Classification model training and optimization results

The input data for the model are user-supplied information about transportation mode segments (*TMS*) as provided previously in Table 4-2.

Feature Estimation and Selection

Next, statistical properties of attributes were computed. Average, maximum speed, minimum speed, acceleration and jerk were calculated; the standard deviations, the 98th percentiles and the difference between the 98th and 50th percentile values for these three parameters of motion were also computed. These attributes were ranked based on their *ADP* as shown in Equation (5-4).

The features with *ADP* less than 90% were eliminated, as the inclusion of weakly differentiating attributes actually diminished the classifiers' performance. For instance, the minimum speeds for all transportation modes were almost equal – nearly zero – which created a common attribute and precluded differentiation. Figure 5-8 illustrates the ranked features and their *ADP;* those shaded in grey with values less than 90 were eliminated from the model. As a result of this step, the optimization method considers only a range of between 1 and 11 attributes in the Feature Vector, with the remaining five being eliminated.

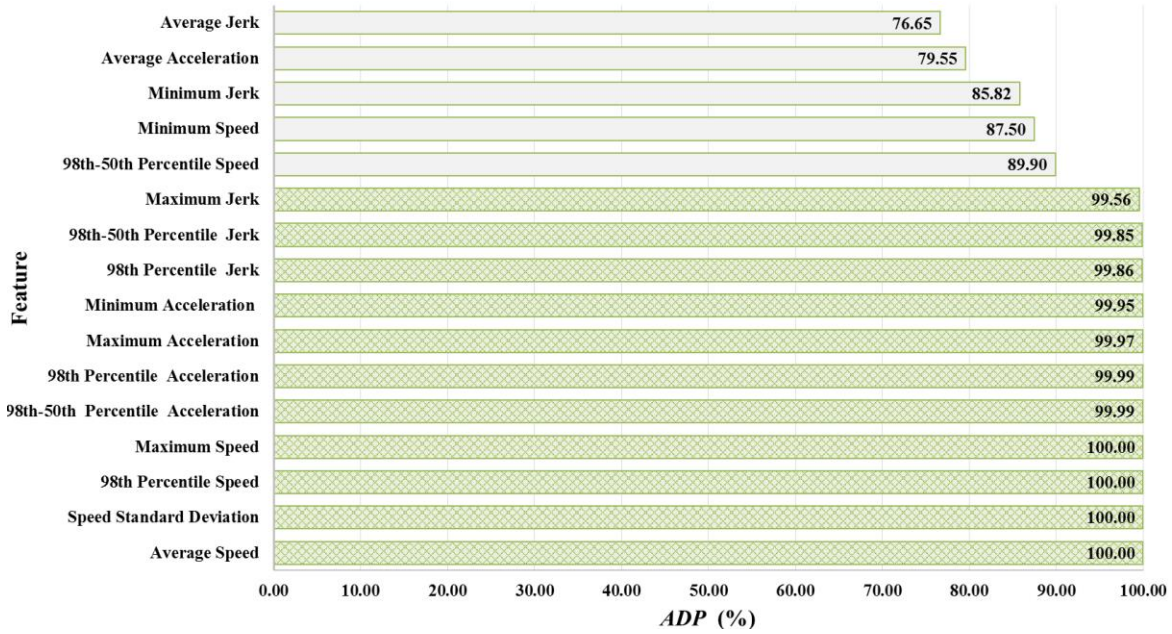


Figure 5-8 *ADP* values of Ranked Attributes

Feature vector and classification model optimization

Given this list of attributes to be included in the classifier model, the next step is to investigate the optimal form of the model: α (the degree of correlation to be applied in calculating *AADP*)*;* the data format; the classification technique and the parameters that define them; and the number of features (of the 11 brought forward from the previous step).

To this end, an iterative process was developed to calculate *MCR* for a given combination of model parameters. In fact, based on the size of the dataset, six independent values of *MCR* were calculated for each combination using the stratified cross validation technique above. Figure 5-9 shows the pseudo code for calculating the *MCR* under different combinations that varies all parameters.

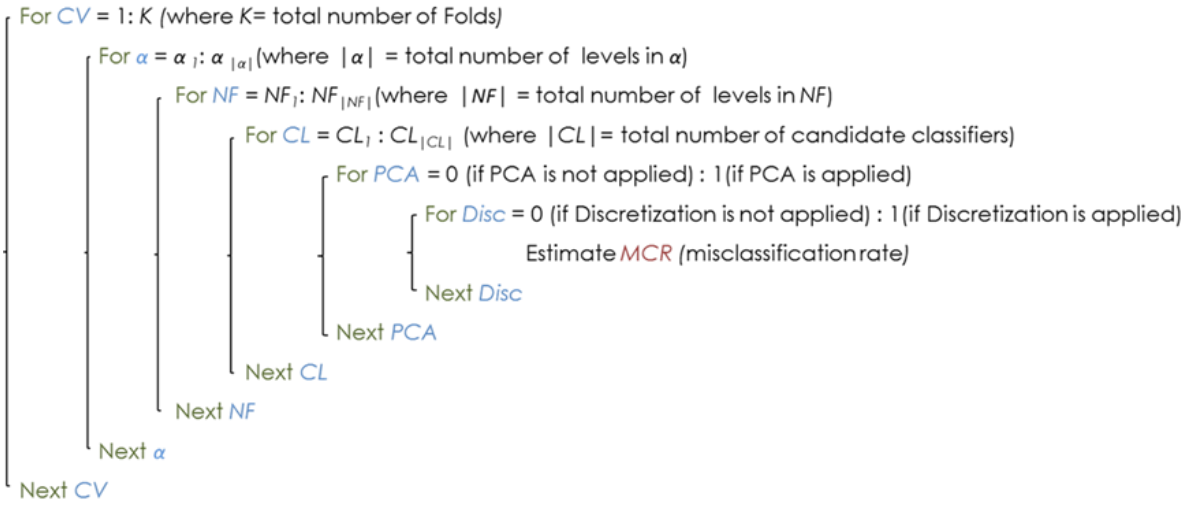


Figure 5-9 The pseudo code for calculating the MCR under different combinations

The number of combinations tested can be calculated as the product of all choice sets of input variables. These are: α = 6; *NF* = 8; *CL*= 3; *Disc* = 2; *PCA* = 2. Thus, the total is 576 combinations. Using the N-fold approach, in this case with N=6, produces six MCR results for each combination of inputs. It is possible to determine the "optimal" model performance simply by finding the minimum MCR amongst the total combinations. A more robust approach is to use a linear regression model to determine two outputs: the explanatory power of each combination and the interactions amongst a subset of the model's parameters. The approach taken is to assign a new, binary variable for each possible state of the input variables. For example,α_1=1 when the model is run with α=0, and α_1=0 otherwise. Mathematically, it can be written:

when α_1=1, $\alpha_i = 0 \; \forall i \neq 1$.

Similarly, CL_1 represents the state (type) of the classifier used in the model. When *NB* is used, CL_1=*1* and $CL_i = 0 \; \forall i \neq 1$. The result is 21 binary variables representing the states of each input. Table 5-6 presents the notation used for the binary variables.

Table 5-6 Notation of binary variables

	$x_{i,j}$	j ⟶	$\lvert x_i \rvert$
i ↓	α	{0:0.2:1}	6
	NF	{1,3,5,6,7,8,10,11}	8
	CL	{NB, k-NN, QDA}	3
	PCA	{0, 1}	2
	Disc	{0, 1}	2
		Total	21

Each variable is indexed by the subscript *i*; the individual values that each variable can take are indexed with the subscript *j*. The length of each variable, $|x_i|$, is the number of levels that are possible; for example, α may take on six values, so its length is six.

Using this notation, a generalized form of the regression equation can be written. Equation (5-11) shows the equation to be solved. Here, *Y* represents the independent variable, in this case the misclassification rate, *MCR*. The variable θ represents the regression constant and coefficients of each term. The superscripts on θ indicate the "level" of interaction: level 1 is the main effect of the binary variables; level 2 is the pairwise interactions amongst all binary variables; and level 3 is the three-way interaction amongst select binary variables. For the three-way interactions, the model considers interactions amongst:

- *α, NF, CL;*
- *NF, CL, PCA;*
- *NF, CL, Disc;*
- *CL, PCA, Disc.*

$$Y = \theta_0 + \sum_{i=1}^{L}\left(\sum_{j=1}^{|x_i|}\left(\theta_{i,j}^{(1)}\, x_{i,j} + \sum_{k=i+1}^{L}\left(\sum_{l=1}^{|x_k|}\left(\theta_{k,l}^{(2)}\, x_{i,j}\, x_{k,l} + \sum_{m=k+1}^{L}\left(\sum_{n=1}^{|x_m|}\left(\theta_{m,n}^{(3)}\, x_{i,j}\, x_{k,l}\, x_{m,n}\right)\right)\right)\right)\right)\right)$$

Main effects — Second-order interactions — Third-order interactions

(5-11)

The regression model is calibrated iteratively. In the first case, all variables are included. After the first calibration, those variables that are statistically significant are retained in the model; those without statistical significance are eliminated. The model is then re-calibrated. The output of the model solution is a matrix of values for θ. The optimal model solution can be identified by finding the sum of θ values that minimize the misclassification rate. Given the complexity present in the model outputs – both the pairwise and three-way interactions – it is difficult to find through inspection the best combination of variables. Therefore, a graphical illustration of the sums of the θ values, including interaction terms, is shown in Figure 5-10. The optimal model performance is shown in the top right diagram.

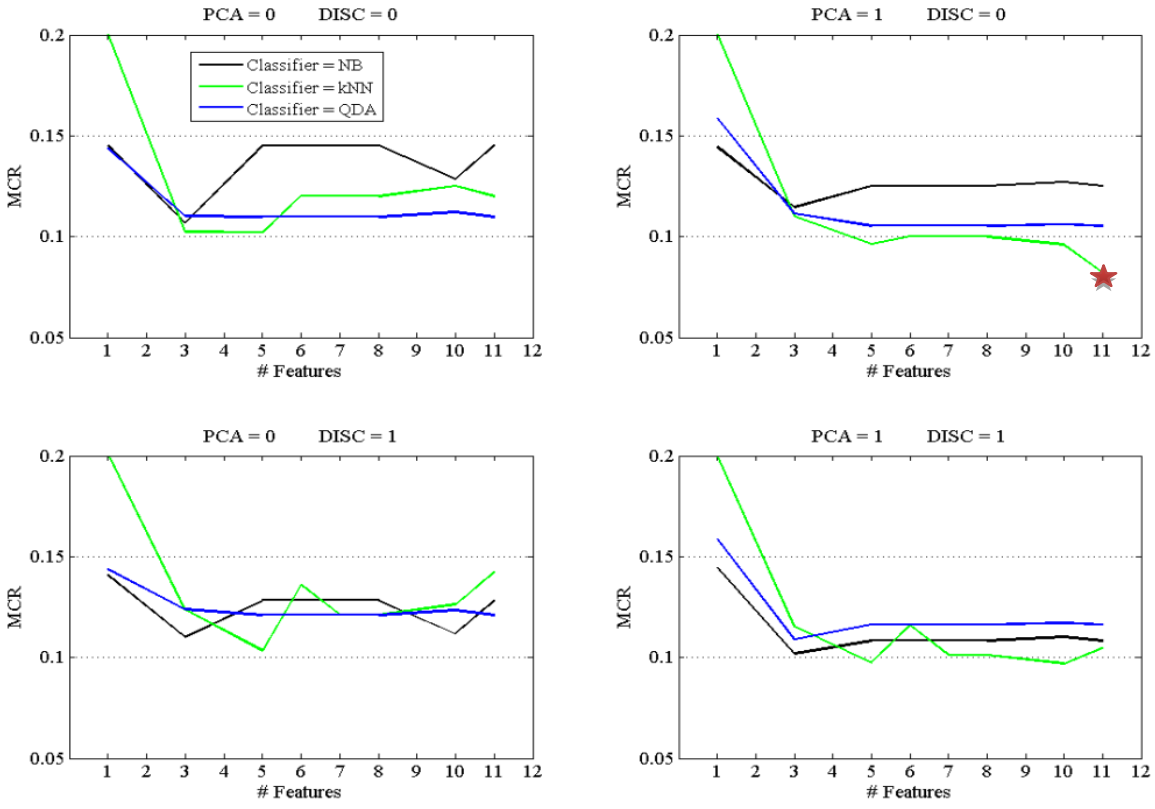


Figure 5-10 Results of the Relative Contribution of Optimizing Model's Parameters

A number of observations can be made on the basis of the results provided in Figure 5-10 and provided in Appendix A-1:

- The coefficients associated with α were not statically significant. Recall that the purpose of α is to reorder the attributes in the feature vector to account for correlation amongst these attributes. The interpretation of α being statistically insignificant is not that the variables were not reordered, but rather that the reordering did not influence the model's performance significantly. This result can be explained by the high correlation between the attributes considered in the feature vector (*FV*). The calculated correlation coefficients are included in Appendix A-2.
- The performance of the NB classification model improves when feature discretization is applied (*Disc*). This outcome is consistent with results found in the literature. However, feature discretization does not improve the performance of the QDA and k-NN models.
- When *PCA* is not applied, increasing the number of features in the feature vector either degrades or does not improve the models' performance. It is speculate that the lack of improved performance with the increased number of features is because of the high correlation between some of the included features.
- The best model performance was obtained when applying *PCA* on the continuous (non-discretized) attributes using the whole set of features (11 feature) and the k-NN classification model. Generally, *PCA* transforms the original correlated data into uncorrelated linear components, called principal components; therefore, it overcomes the problem of the highly correlated set of features.

Final Model Formulation

Based on the model structure determined in the previous step, the parameters of the classification model are now calibrated. The results of the regression produce an optimized model with the following characteristics:

α =0;

All 11 attributes are included in the model;

The best performing classification method is k-NN with k=11;

PCA is applied with 98% variation in the data; and

The data should be treated as continuous.

Although *PCA* is usually used for the purpose of dimensionality reduction, in this case *PCA* transforms the original data into linearly uncorrelated components which provides the best results perhaps as a result of the strong correlation amongst the attributes in the data set. As expected, producing a higher number of principal components captures an increasing amount of the variance in the data. Using a threshold of acquired retained variance of 98%, the first five ranked principal components have been selected. Figure 5-11 shows the accumulated variance; note that the first three principal components explain almost 95% of the total variance in the data.

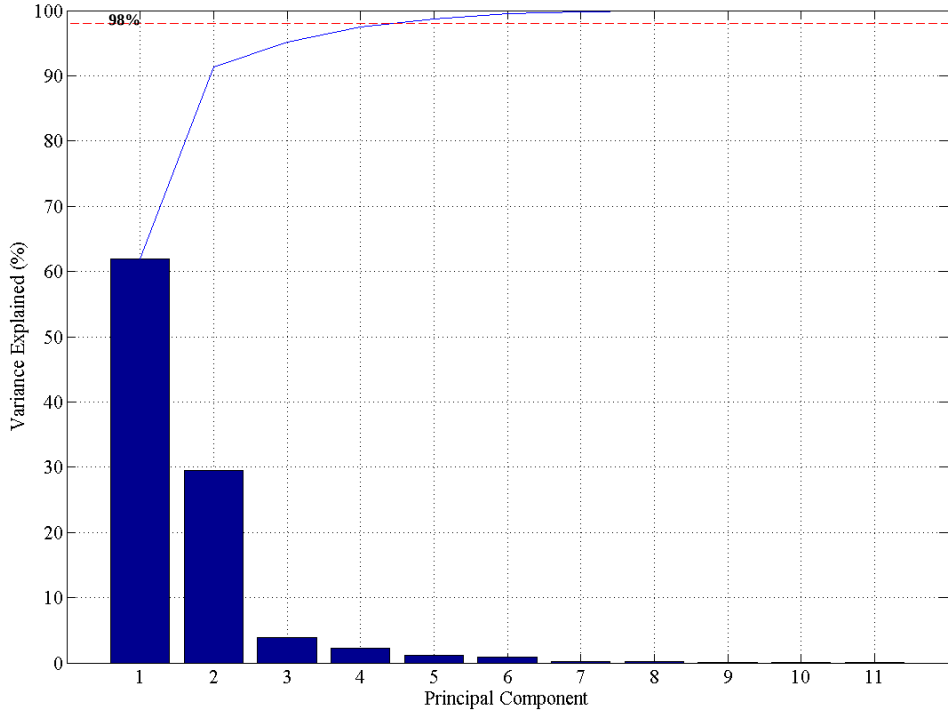


Figure 5-11 Cumulative and Individual Explained Variance by each Principal Component

As a preliminary test of the model formulation, the *MCR* was computed for the entire data set using cross validation with 10 folds. The average result of this training and testing produces a *MCR* of 9%.

5.2 Transportation mode classification model application

The previous steps have developed an optimized classifying model that specifies which Feature Vector attributes should be used to identify modes; the degree of correlation to be considered amongst these attributes; whether data should be considered as discrete or continuous; whether

the use of *PCA* improves the predictive power of the model; and whether k-NN, QDA or NB is the appropriate classification method.

To apply the trained classifier, the level of aggregation of GPS data is reset. More specifically, all transportation mode labels are assumed to be removed from the data set; the boundaries of transportation mode segments are also removed. The only information maintained from the travel diary is the identification of the beginning and end of one trip. Then, Module B2 is applied as shown in Figure 5-12.

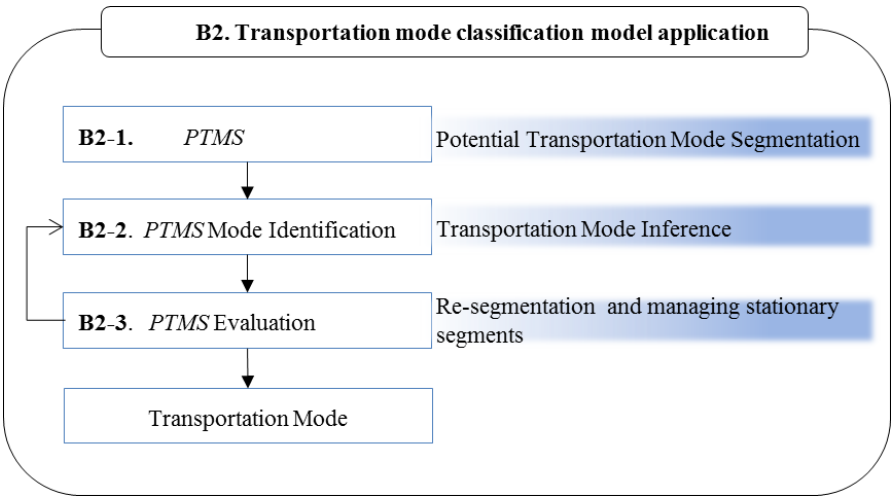


Figure 5-12 Transportation mode classification model application module (B2)

5.2.1 Potential Transportation Mode Segmentation

Because *TMSs* are not known, the first step is to identify moving sub-segments for which a transportation mode can be identified. Stationary sub-segments are also identified that can represent two possible situations: a normal transportation event, such as stopping at a traffic signal or the interval during when a traveler changes transportation modes (Potential Mode Transfer Segment). The criteria for identifying stationary sub-segments are: speed, v, must be less than a speed threshold, v_{th} for duration of time between five and 120 seconds. Those stationary sub-segments are labeled as potential mode transfer segment (*PMTS*) where the users are more likely to change their transportation mode. Figure 5-13 shows the first hypothetical trip as shown previously in

Figure 4-6. In this case, the diagram shows the boundaries of moving and stationary sub-segments.

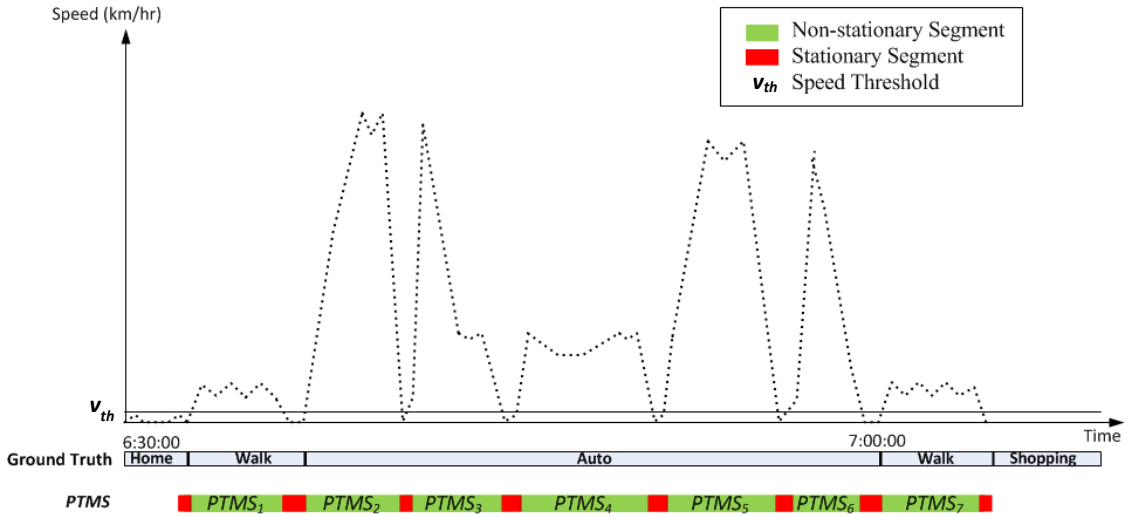


Figure 5-13 Dividing trips into moving and stationary sub-segments

5.2.2 *PTMS* Mode identification

The next step in the analysis is to apply the optimized k-NN classification model and identify the most likely transportation mode for each moving segment. As illustrated in Figure 5-14, the classification model has labeled moving sub-segments 1 and 7 as walking; moving sub-segments 2, 3, 4, 5 and 6 have all been labeled as auto.

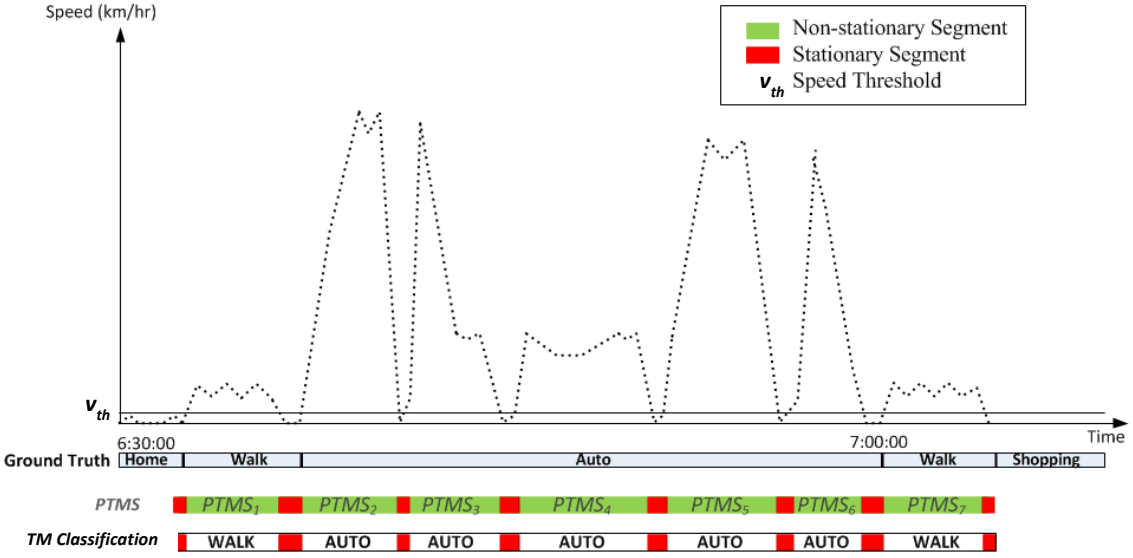


Figure 5-14 *PTMS* Classification results based on the optimized k-NN classifier

5.2.3 *PTMS* Evaluation (Managing Stationary Segments)

Initially, all stationary sub-segments are characterized as potential mode transfer points – where a traveler changes transportation modes. Now that the moving sub-segments have been assigned a mode, mode transfer points can be evaluated. In Figure 5-14, the k-NN model has assigned walking mode for $PTMS_1$ and auto mode for $PTMS_2$. As such, the stationary sub-segment between them will be considered a mode transfer segment. By definition, all points in the stationary sub-segment are added to the previous moving segment – part of the walking trip; the last point in the stationary sub-segment is the mode transfer point.

The stationary sub-segment between $PTMS_2$ and $PTMS_3$, in contrast, are bound by two auto sub-segments. Thus, this stationary sub-segment between these two moving sub-segments is not a mode transfer point (no change in modes occurred) and therefore the stationary sub-segment can be eliminated. $PTMS_2$ and $PTMS_3$ are merged to form a single, larger, *PTMS*. This same process of merging *PTMS* is continued so that a single *PTMS* is formed that includes the original $PTMS_2$ through $PTMS_6$.

The k-NN classification model is then re-run on the merged sub-segments: new features are calculated and a new mode is estimated for each *PTMS*. The process is repeated until no further merging can occur, a mode has been estimated for all *PTMS*, all stationary sub-segments have been assigned to the associated previous *PTMS* and the mode transfer points (*MTP*s) are identified (Figure 5-15).

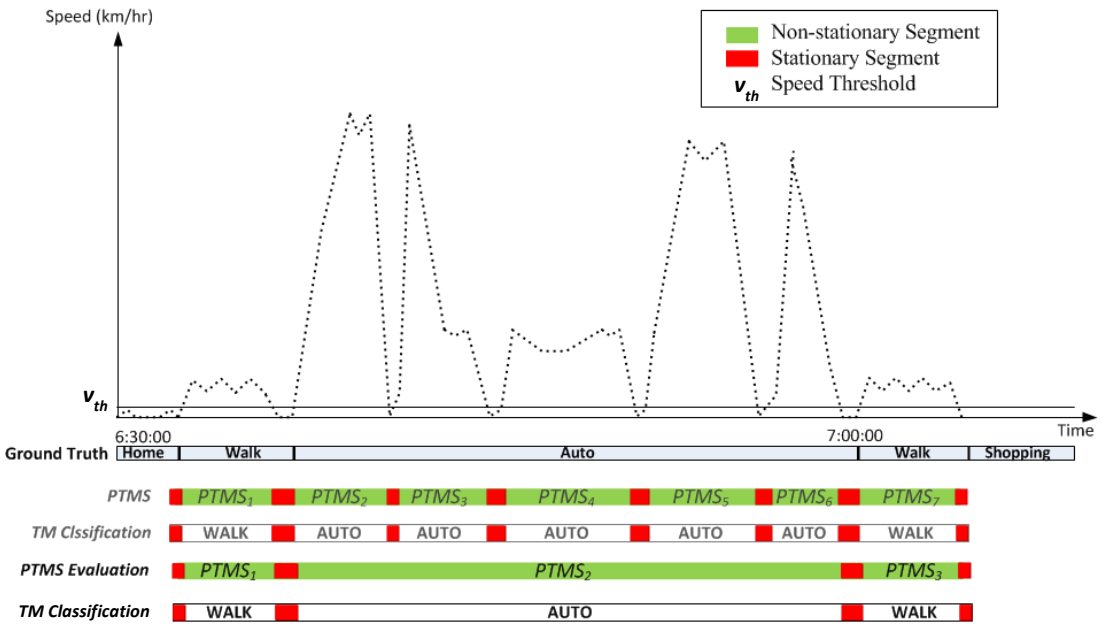


Figure 5-15 *PTMS* Evaluation and reclassifying results based on the optimized k-NN

The results from before and after *PTMS* evaluation are summarized in Table 5-7. The impacts of combining adjacent, similarly assigned *PTMS* can be seen by plotting the distribution of number of points contained in the *PTMS* before and after the evaluation process. Figure 5-16 shows these results. Within the data set, there exist 4075 moving sub-segments. Of these, 2028 contain fewer than 20 points; an additional 1366 contain between 21 and 40 points. These very short moving sub-segments represent 84% of all *PTMS*. After *PTMS* evaluation (merging) process, the total number of moving sub-segments is reduced to 2592. The merging process also produces a greater proportion of longer segments; after merging, sub-segments with fewer than 40 points constitute only 68% of the total.

Table 5-7 *PTMS* and *PTMS* evaluation results summary

	Before *PTMS*		**After *PTMS* Evaluation**	
	Segments	**Points**	**Segments**	**Points**
Moving	4075	92470	2592 (-36%)	99210 (+7%)
Stationary	2705	13168	1222 (-55%)	6428 (-51%)
Total	6780	105638	3814	105638

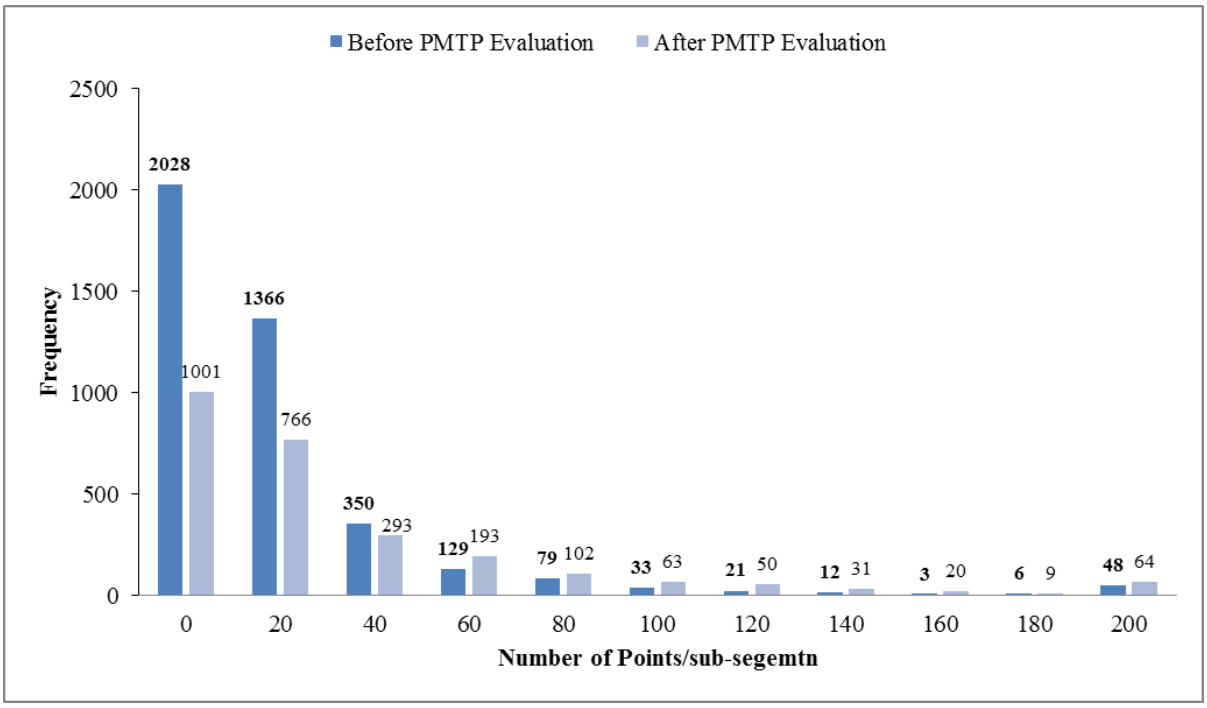


Figure 5-16 Distribution of number of points per *PTMS* before and after *PTMS* evaluation

The benefits of producing fewer, longer segments are that the attributes for the feature vector are calculated from a larger data set and, as a result, more likely to produce accurate results. Further, the merging process removes many stationary sub-segments – 1,483 sub-segments or 54.8% of the total – which decreases the likelihood of identifying incorrect mode transfer points. Finally, the merging process produces a distribution of segment lengths that more closely resemble the actual (user-labeled) trip patterns.

In terms of mode transfer points (*MTP*), the actual number of *MTP* between different transportation modes in our data set is 203. Before the *PMTP* evaluation process, 1009 *PMTP* were identified and only 172 of them were actually *MTP*. So, while the model initially correctly identified 85% of actual *MTP*, the model also categorized an additional 837 incorrect *PMTPs*. After the *PMTP* evaluation process, 178 *PMTP* were correctly identified (88%), but an additional 756 incorrect *PMTPs* remained.

These results indicate that the model suggests that a traveler changes modes much more frequently than in reality. It should be noted that every incorrect segment label produces two incorrect *PMTPs* – one at the beginning and end of the incorrect segment. For clarity, two possible cases exist for the sub-segment following an incorrect *PMTP*. The first is that the

incorrect *PMTP* results from a change from an incorrectly labeled sub-segment to a correctly labeled sub-segment. The second is the case with a change from a correctly or incorrectly labeled sub-segment to another mode that is incorrectly labeled. Figure 5-17 shows a hypothetical example of these two cases.

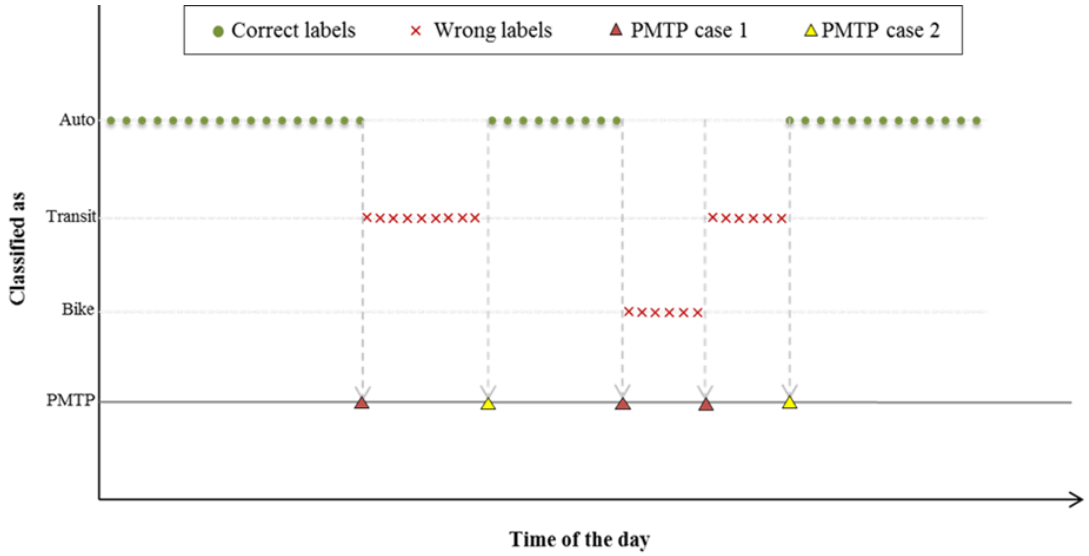


Figure 5-17 Labels for hypothetical example illustrating different cases for incorrect *PMTP*

To understand the cause of the incorrect *PMTPs*, further investigation was carried out. The initial expectation was that a large number of short segments were misclassified with wrong transportation modes, generating multiple, incorrect *PMTP*. To assess the impacts of short segments on *PMTP* errors, the sub-segments following the incorrect *PMTPs* were investigated. The distributions of the duration of the sub-segments following an incorrect *PMTP* before and after the *PMTP* evaluation are shown in Figure 5-18.

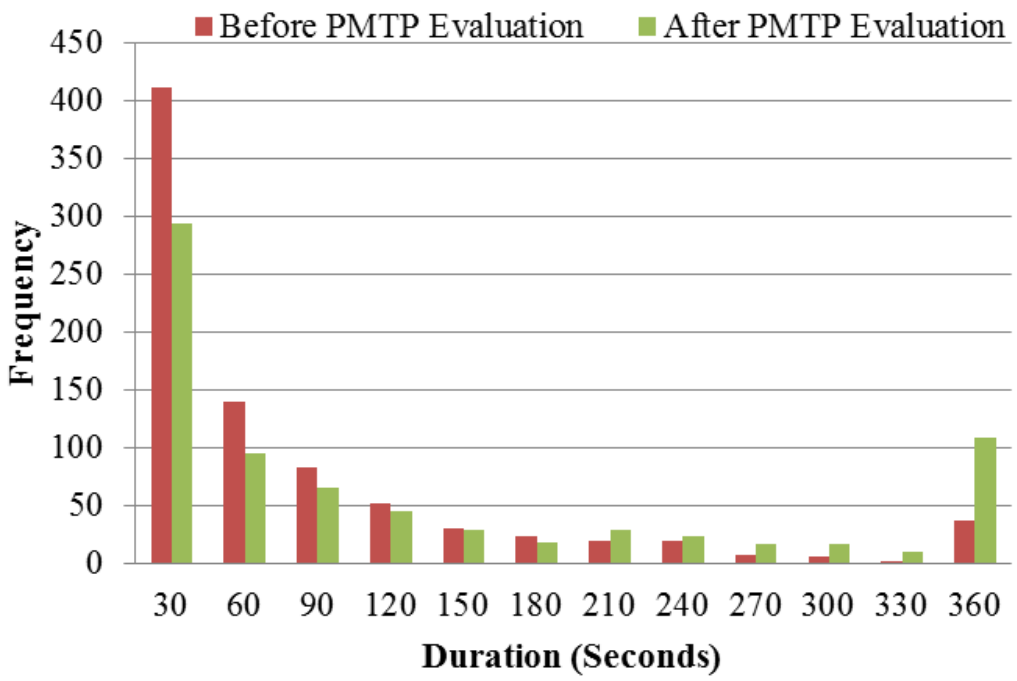


Figure 5-18 Distribution of sub-segments preceded by misclassified *PMTP*

It can be observed that the number of very short segments – less than one minute – is very high. Prior to the *PMTP* evaluation, these short sub-segments constitute 65% of the total incorrectly labeled sub-segments. After the *PMTP* evaluation process, the number of sub-segments has been reduced from 65% of all sub-segments to 50% of all sub-segments. This demonstrates that the *PMTP* evaluation process was only marginally successful in dealing with the challenge of very short sub-segments.

Interestingly, after *PMTP* evaluation, the number of long segments (greater than six minutes) following an incorrectly labeled *PMTP* increased significantly, to 14% of the sub-segments. An investigation of theses long segments found that these sub-segments were associated with the misclassified *PMTP* followed by a correct transportation mode label – described as case 1 in Figure 5-17 above.

To determine the relative importance of case 1 outcomes, Figure 5-18 is redrawn in Figure 5-19 with the case 1 outcomes eliminated. In other words, the distributions in Figure 5-19 are the duration of sub-segments that both follow an incorrect *PMTP* and are incorrectly labeled (misclassified). The number of the misclassified *PMTPs* has been reduced from 1009 and 938 before and after *PMTP* evaluation to 668 and 617 respectively. The numbers of incorrectly labeled *PMTPs* also decrease from 837 to 496 before *PMTP* evaluation

and from 756 to 439 after *PMTP*. The elimination of case 1 scenarios also increases the importance of very short segments. Short segments – less than one minute – account for 85% and 75% of all case 2 incorrect *PMTPs* before and after *PMTP* evaluation.

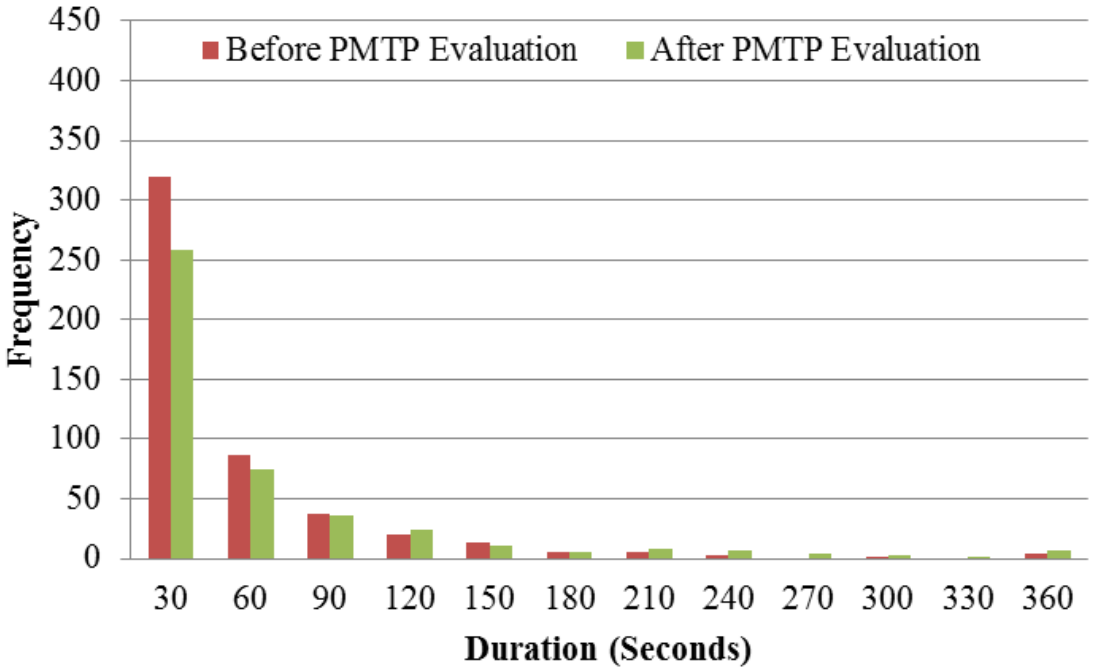


Figure 5-19 Distribution of misclassified sub-segments preceded by misclassified *PMTP*

5.3 Mode inference model results and performance evaluation

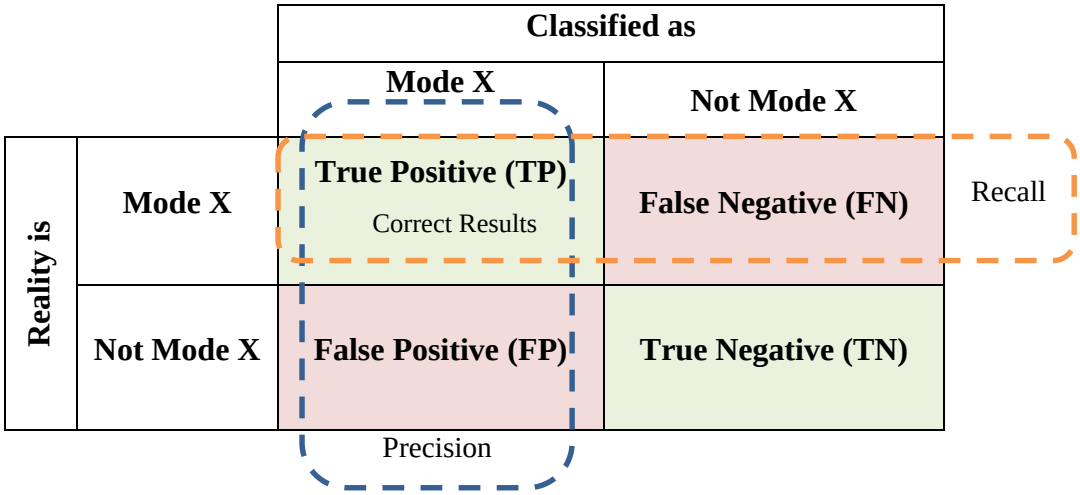
In order to analyze the performance and measure the quality of the classification model, two concepts, computed as ratios, are introduced: precision and recall. For precision, the numerator is the number of objects that were classified as mode X that are in reality Mode X; these results are called "True Positives." The denominator is the sum of the True Positives and "False Positives": objects labeled as Mode X that are in reality not mode X. Mathematically, the ratio is shown in Equation 5-10. The numerator for recall is also the True Positives. The denominator for recall is the sum of the True Positives and the False Negatives: objects which are classified as not Mode X but in reality are Mode X. The equation for recall is given in 5-11. The relationships for precision and recall are shown in Table 5-8.

Precision indicates the percentage of time that a segment that is classified as Mode X is truly Mode X; Recall measures the percentage of time that a segment that is in reality Mode X is classified as such.

$$precision = \frac{TP}{TP + FP} \tag{5-10}$$

$$recall = \frac{TP}{TP + FN} \tag{5-11}$$

Table 5-8 Precision and Recall estimation



Normally, the results of precision and recall calculations are shown in what is known as a confusion matrix – with reality in rows and model outputs in columns. This allows for quick assessments of model performance and areas of weakness.

The evaluation methodology is not done at the transportation mode segment level. Instead, the trained, calibrated model is evaluated on a sub-segment basis; each moving sub-segment in the full data set is classified as one of the four candidate modes. Given this approach, all sub-segments for which the classifier model predicts the correct transportation mode can be quantified. We also calculate recall at the point level. If a sub-segment is classified by the k-NN model as having been made by mode *m*, then all points contained in that sub-segment are assigned mode *m*. By computing the confusion matrix at both the sub-segment level and the point level, we are able to account for different sub-segment durations. Using this approach, we can measure the impact of misclassification on long sub-segments, with many points, and

short sub-segments with fewer points. The model's performance is also quantified at aggregate levels – motorized versus non-motorized modes – to provide a sense of how well the model can distinguish between fundamentally different transportation options.

The classifier assigned the correct aggregate label – motorized (transit or auto) and non-motorized (walking or cycling) – to 94% of sub-segments and 82.6% of points. The classifier assigned the correct mode to 68.71% of the sub-segments and 87.98% of the points in the data set. Table 5-9 shows the confusion matrix at the aggregate level for points.

Table 5-9 Aggregated Confusion Matrices at Sub-Segment and Point Levels– grey highlights represent correct inference

Table 5-9a. Sub-Segment Level

		Classified as		**Recall**
		Non-Motorized	**Motorized**	
Reality	**Non-Motorized**	1132	22	98.09%
	Motorized	429	1009	70.17%
Precision		72.52%	97.87%	***82.60%***

Table 5-9b. Point Level

		Classified as		**Recall**
		Non-Motorized	**Motorized**	
Reality	**Non-Motorized**	33699	722	97.90%
	Motorized	5163	66054	92.75%
Precision		86.71%	98.92%	***94.43%***

Table 5-10 shows the classifier's performance in terms of the confusion matrix for all four modes at the sub-segment and point levels.

Table 5-10 Disaggregated Confusion Matrices at Sub-Segment and Point Levels– grey highlights represent correct inference

Table 5-10a. Sub-Segment Level

		Classified as				Total	Proportion	Recall
		Walk	**Bike**	**Transit**	**Auto**			
Reality	**Walk**	776 (97.86%) **(67.01%)**	7 (0.88%) **(1.74%)**	6 (0.76%) **(9.52%)**	4 (0.50%) **(0.41%)**	793	0.75%	97.86%
	Bike	155 (42.94%) **(13.39%)**	194 (53.74%) **(48.14%)**	0 (0.00%) **(0.00%)**	12 (3.32%) **(1.24%)**	361	0.34%	53.74%
	Transit	69 (19.55%) **(5.96%)**	47 (13.31%) **(11.66%)**	48 (13.60%) **(76.19%)**	189 (53.54%) **(19.52%)**	353	0.33%	13.60%
	Auto	158 (14.56%) **(13.64%)**	155 (14.29%) **(38.46%)**	9 (0.83%) **(14.29%)**	763 (70.32%) **(78.82%)**	1085	1.03%	70.32%
Total		1158	403	63	968	***Overall Accuracy 68.71%***		
Proportion		44.68%	15.55%	2.43%	37.35%			
Precision		67.01%	48.14%	76.19%	78.82%			

Table 5-10b. Point Level

		Classified as				Total	Proportion	Recall
		Walk	**Bike**	**Transit**	**Auto**			
Reality	**Walk**	19581 (97.77%) (**80.48%**)	78 (0.39%) (**0.54%**)	171 (0.85%) (**5.96%**)	198 (0.99%) (**0.31%**)	20028	18.96%	97.77%
	Bike	2027 (14.08%) (**8.33%**)	12013 (83.46%) (**82.67%**)	0 (0.00%) (**0.00%**)	353 (2.45%) (**0.55%**)	14393	13.62%	83.46%
	Transit	1497 (16.65%) (**6.15%**)	558 (6.20%) (**3.84%**)	2462 (27.38%) (**85.81%**)	4476 (49.77%) (**7.00%**)	8993	8.51%	27.38%
	Auto	1226 (1.97%) (**5.04%**)	1882 (3.02%) (**12.95%**)	236 (0.38%) (**8.23%**)	58880 (94.63%) (**92.13%**)	62224	58.90%	94.63%
Total		24331	14531	2869	63907	***Overall Accuracy 87.98%***		
Proportion		23.03%	13.76%	2.72%	60.50%			
Precision		80.48%	82.67%	85.81%	92.13%			

(xx) = Recall = number of points classified as mode *n* divided by total # of actual points of mode *m* times 100%.
(**yy**) = Precision = number of mode *m* points divided by the total number of points classified as mode *n*× 100%.

The classifier performs very well in identifying walking sub-segments, with nearly 98% recall. The model was able to correctly classify 53% of all bike sub-segments; interestingly, the model's performance was substantially better for biking at the point level (83.5%). This implies that errors in bicycle classification are a result of many short sub-segments that are misclassified. A similar observation can be made for auto sub-segments. The recall improves from 70% at the sub-segment level to 94.6% at the point level. Further work is necessary to improve the classifier's performance on these short sub-segments.

Significant challenges were experienced in the classification of transit segments. The attributes of transit vehicle movements –buses operating in mixed traffic – are particularly difficult to distinguish from auto movements and, in some congested cases, walking segments. As with the recall for bike and auto modes, transit recall improved at the point level relative to the sub-segment level.

5.4 Assessment of inference model's performance

In order to provide additional insights into the performance of the classification model, two trips have been chosen for a visual investigation using the developed GUI. Figure 5-20 and Figure 5-21 illustrate three different diagrams for each trip as follows:

1. Speed-time diagram showing the speed threshold v_{th} with the horizontal red line.
2. Cumulative distance-time diagram
3. The actual label versus the labels from the classifier.

The first trip was performed using an auto. The classification model correctly classified 98.5% of the points as Auto with only few points misclassified as Bike. The trip consists of 10 moving and 9 stationary sub-segments. It can be observed in the speed-time diagram that the consistency of the speed profile in the moving sub-segments during most of the trip allowed the model to merge all 19 sub-segments into only two sub-segments. It can be noted that the misclassified points as Bike is justified by the drop in speed for the last sub-segment before performing the activity. The misclassification of this sub-segment caused wrong indication of *PMTP* as shown on the speed-time diagram with a red arrow where in reality there is no mode transfer point (*MTP*).

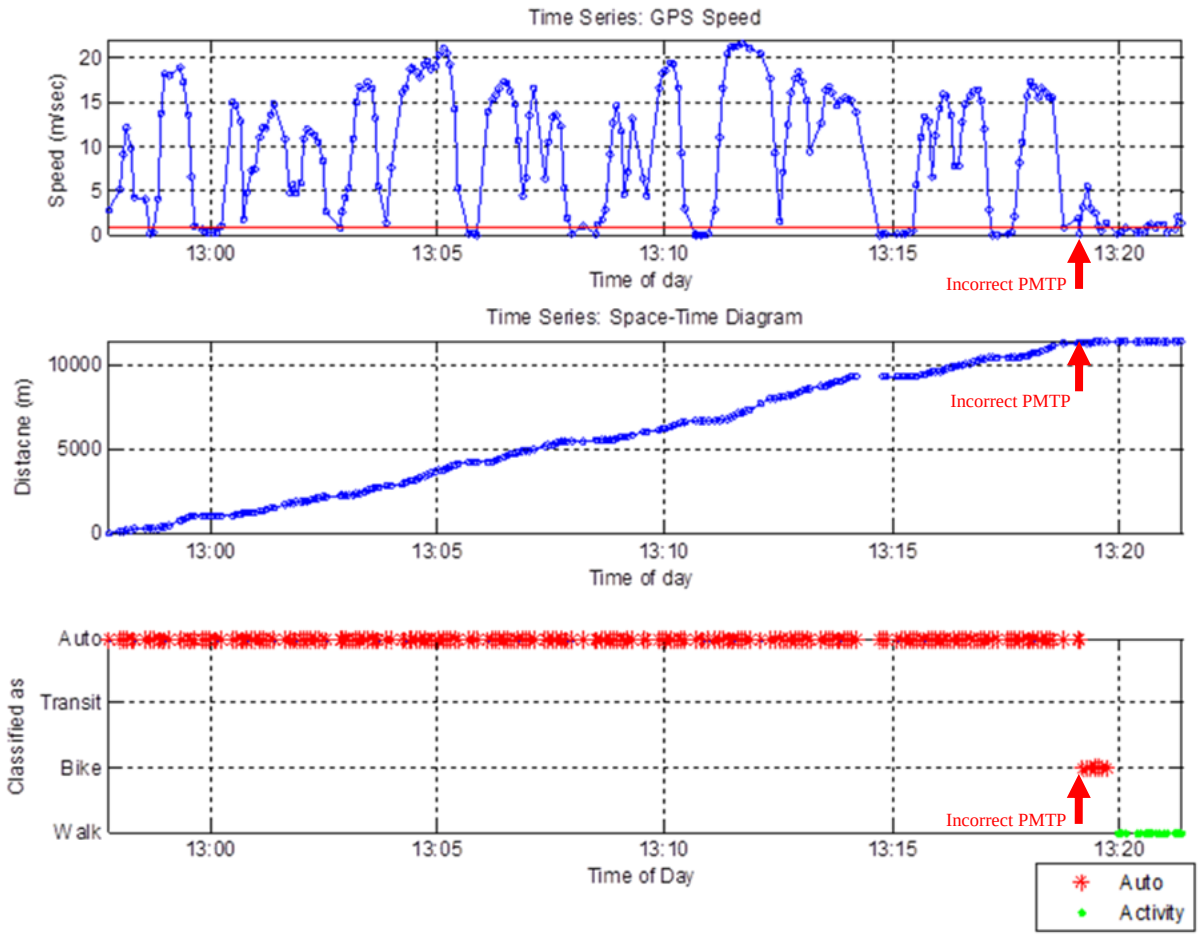


Figure 5-20 GUI diagram for a private auto trip (Trip # Trip 495111)

The second investigated trip actually includes two MTPs and three transportation mode segments as follows: Walk-Transit-Walk (Figure 5-21). The model was only able to classify the Walk *TMS* correctly and misclassified many segments that should have been identified as a single transit *TMS*. Unlike the first trip, significant inconsistency is observed in the speed profile during the transit *TMS*. As a result, the model misclassified the moving sub-segments with multiple transportation modes i.e. Auto, Bike, and Walk. The model's labels are shown in the classification label-time diagram. The large number of misclassified segments also produced 14 wrong PTMP while only correctly labeling one of the two actual *PMTPs*. The green arrow in diagram show this points.

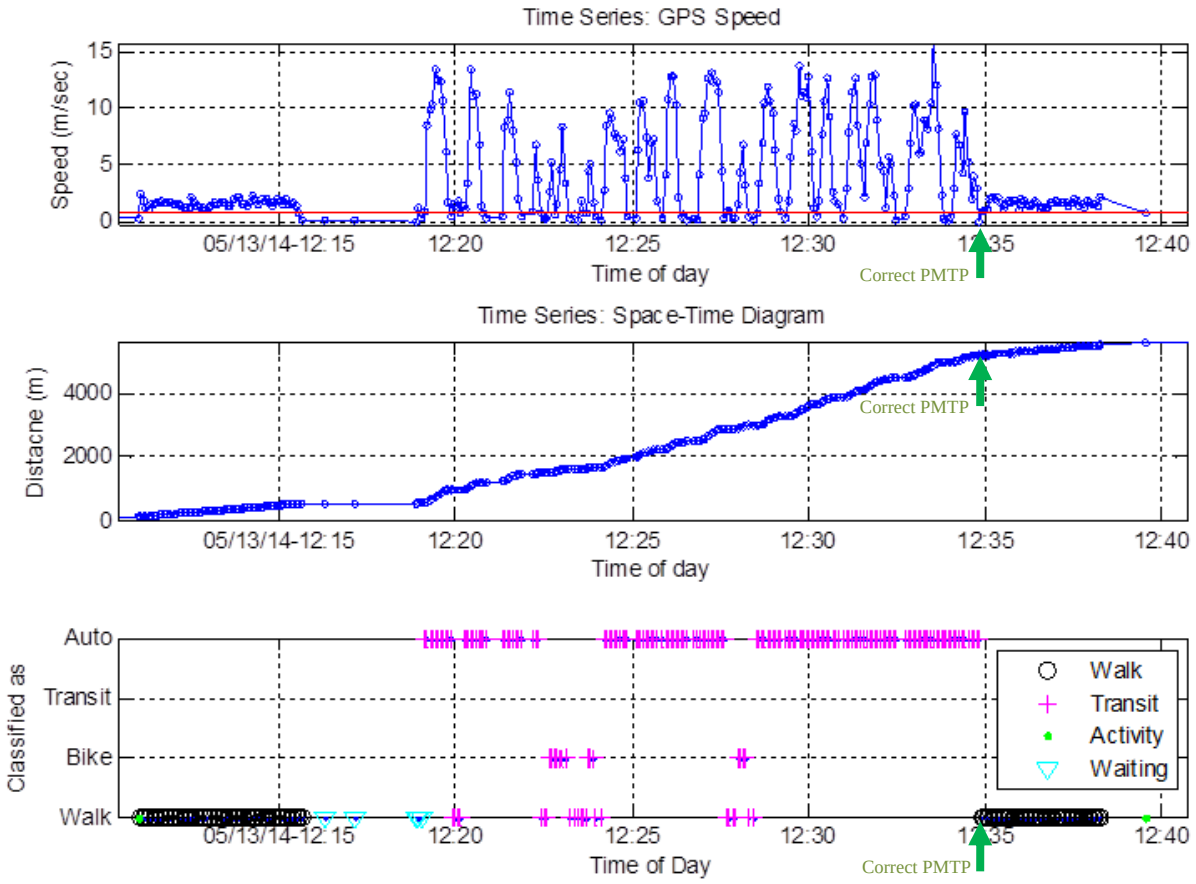


Figure 5-21 GUI diagram for a private transit trip (Trip # 1522)

5.5 Model comparison to previous work

The performance of the mode classification, as quantified in the previous section, is not as good as the performance of some models reported in the literature. For example, Chung and Shalaby (2005) reported model accuracy of 91.7% based on 60 *TMS*. However, the differences between the performance of the model presented here compared to results previously published by other researchers is largely a result of:

- The frequency of the data collection. In this work data are collected every five seconds, while others have gathered data at one second intervals. It should be noted that in practice, battery power management on smart phones is a significant concern and frequent position acquisitions using the onboard GPS tends to deplete the smart phone battery rather quickly (Taghipour, 2012). Previous work suggests that

acquiring GPS data at a frequency greater than one observation every 5 seconds is likely not practical.

- The quality of the data. Previous researchers gathered GPS data from GPS loggers, as opposed to GPS enabled smart phones.
- The size of the data set. The data set in this research is larger than what has been typically used in previous research. As such, the challenge of training a model was much greater in the present research. As noted, model performance is typically measured (using recall, precision or overall accuracy) as percentages. Naturally, achieving similar performances in terms of percentages on larger data sets indicates greater number of actual segments correctly identified by the model.
- The diversity of the data from which the mode inference was made. It is not uncommon in previous work for the researchers to have given specific instructions to survey participants on the order and duration of travel by a given mode. In this research, participants were asked to record their data during normal travel activity, and to label these data. No specific instructions were given on the patterns of travel – i.e. participants were not required to sequence their modes to inform the model development.
- The constructs around mode transfers. Previous researchers provided specific instructions to indicate mode transfers by a specific activity – for example at least one minute of walking time when changing modes. Also, the beginning and ends of trips were defined by walking segments of at least 60 seconds. No such limitations were placed on participants who provided data for this research.
- The last but most important factor explaining the differences in reported model performance is the level of aggregation in evaluating the model. Previous researchers trained their model using transportation mode segments with known labels. They then tested the performance of their model at the same level of aggregation as in training – identifying the mode of known transportation mode segments. In this research, the training was complete at the transportation mode segment level, but the evaluation method assumed no knowledge of the limits of transportation mode segments.

5.6 Summary

This chapter has addressed the challenge of identifying the transportation modes used for trips (or portions of trips) solely on the basis of the GPS data. The work presented here has made several contributions. First, a novel method by which attributes are either included or excluded from the classification model has been introduced and implemented. Using statistical techniques, an attribute's differentiating power relative to other candidate attributes is quantified and the Feature Vector is populated with attributes in descending order of differentiating power. Second, the approach taken can be used to objectively optimize the selection of the type of classification model, and whether or not to apply transformation techniques (*e.g.*, discrete versus continuous and PCA) to the feature vector attributes. Third, the proposed model requires fewer and less restrictive assumptions about the trip structure than most existing models and unlike many previous studies, the proposed model was evaluated under the more realistic conditions as the segmentation of trip data is part of the transportation mode estimation problem.

The results indicate that overall the proposed model performs quite well. One of the main challenges in differentiating between the modes at aggregated level, motorized and non-motorized, is when the network is highly congested. This problem is common as some of the attributes between different classes overlap which make it difficult for the model to identify the correct class. Many of the misclassification errors are associated with short sub-segments and with distinguishing between transit and auto modes. The next chapter introduces and demonstrates a method to improve upon the model's performance for transit trips.

Chapter 6
Enhanced Model Integrating GIS

In the previous chapter, a novel method was developed that optimizes a transportation mode classification model with respect to the following: (1) the number of attributes to consider in the model based on the ability of those attributes to distinguish between the available transportation modes; (2) the type of data processing employed; and (3) the model parameters. This method was employed using a set of labeled GPS data to develop an optimized mode inference model (denoted as the optimized k-NN model) which was shown to perform well in classifying modes, but performs best when differentiating at an aggregate level (i.e. motorized versus non-motorized modes). One problem that remains very difficult is the correct identification of transit travel. The problem stems from the fact that transportation modes within each aggregate category (*e.g.*, motorized or non-motorized modes) exhibit very similar distributions of basic attributes (*e.g.*, speed and acceleration).

In this chapter, the optimized model from the previous sections is supplemented to improve the identification of transit trips. The method employed seeks to extract additional information about modal attributes by utilizing the spatial information from the GPS data and combining this with other commonly available spatial data (*e.g.*, road network topology, etc.) within a Geographic Information Systems (GIS). The combination of these two data sources generates new classification features that have the potential to vastly improve the accuracy of developed transportation mode inference model. This step is the last step in mode framework proposed in this dissertation as shown in Figure 6-1.

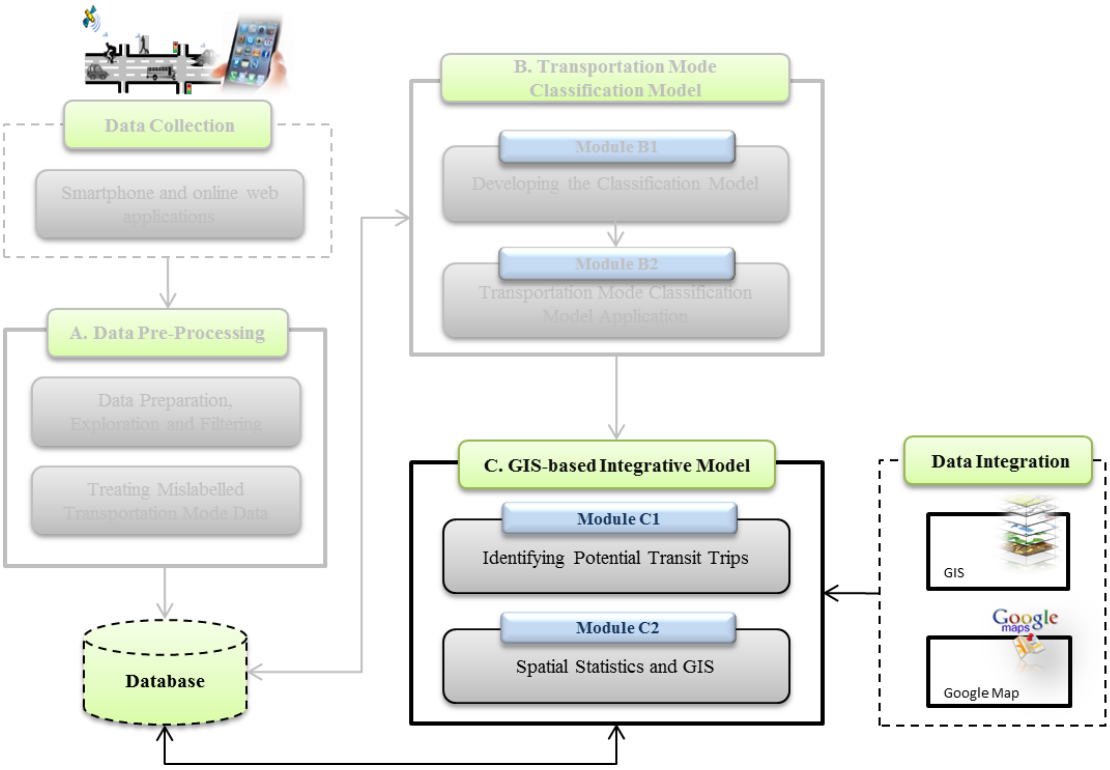


Figure 6-1 Proposed Transportation Mode Classification Framework
(C. GIS-based integrative model)

6.1 Model Enhancement with GIS Integration

The approach builds upon the optimized transportation mode classification model from the previous chapter with two additional steps. Instead of reviewing all trips in the dataset, the model starts by identifying initial potential transit trips (*IPTT*) based on the results from the optimized k-NN model. Next, an integrated, second-stage methodology is proposed that compares the locations and patterns of stops from the GPS data, to locations of transit stops and signalized intersections from GIS data. An additional algorithm is developed to detect the actual location (time and space) at which travellers change between transit and other modes.

6.1.1 Identifying Potential Transit Trips

From the optimized k-NN model, all transportation mode sub-segments within each trip were classified into one of the four possible transportation modes (walk, bike, transit, or private-auto). As mentioned previously, the main model provides remarkably high accuracy at an aggregated level (*e.g.*, motorized and non-motorized). Given this accuracy, the primary interest

is dealing with transit segments which are incorrectly labeled as auto mode, those trips for which all sub-segments were labeled as non-motorized can be eliminated from the data set. This process vastly reduces the overall size of the data set and allows for faster processing of the remaining trips.

6.1.2 Spatial Statistics and GIS attributes

To improve the classifier's performance, the approach is to try and define a set of characteristics that best identify transit trips. Generally, transit trips have two distinctive traits: they must begin and end at spatial locations that coincide with transit stop locations and the stopping pattern tends to be distinct from other motorized travel. Based on these two observations, and the ubiquitously available and very rich GIS data set, the following sections describe the methods used to identify transit trips. The proposed module includes three steps as shown in Figure 6-2

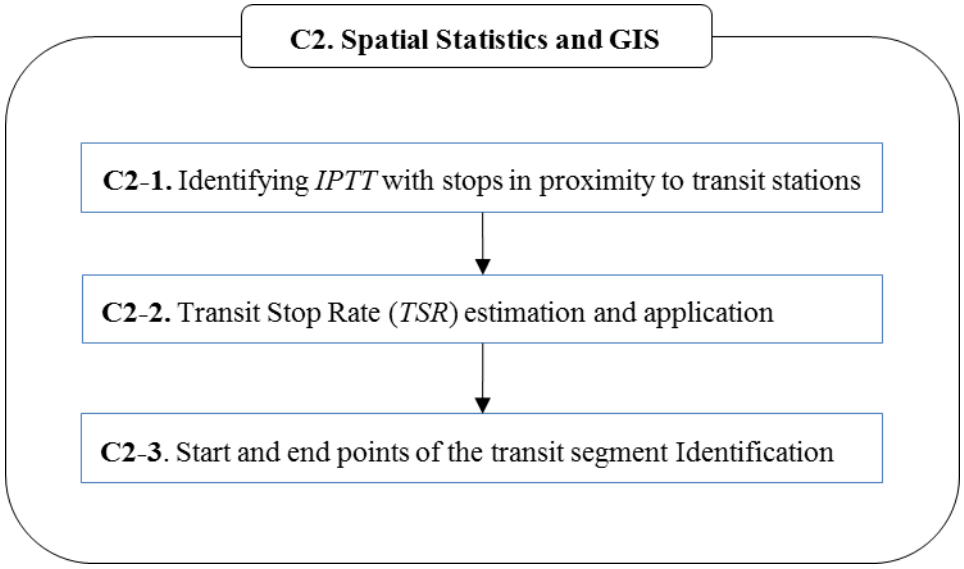


Figure 6-2 Spatial Statistics and GIS module

6.1.2.1 Identifying *IPTT* with stops in proximity to transit stations

When a traveler uses transit to perform a specific trip, there should be at least two locations where the traveler boarded into and alighted from the bus (i.e. origin and destination transit stations). Therefore, the set of *IPTT* identified earlier is reassessed to include only trips that

contain two or more stationary sub-segments (a series of points with velocity less than a threshold) that are within a specified distance of a transit station (TD_{th}), see Figure 6-3.

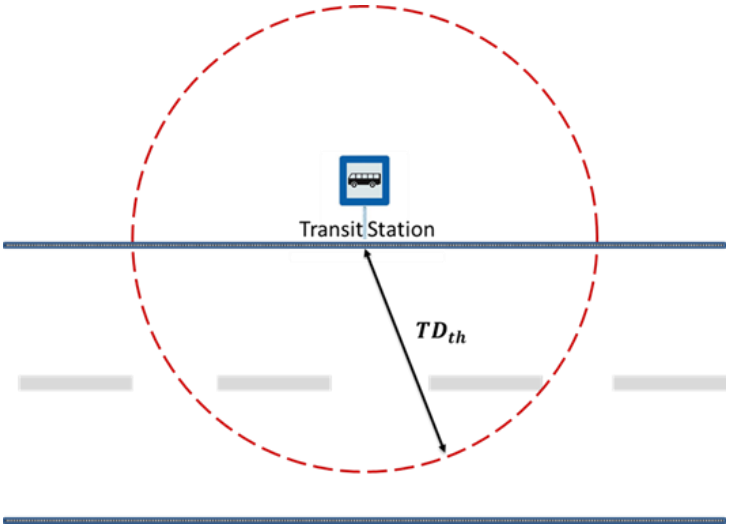


Figure 6-3 A hypothetical transit station illustrates the distance threshold

Essentially, trips are identified that contain at least two sets of points, SS_1 and SS_2 for which:

$$v_{1,i} < v_{th} \; for \; all \; i$$

$$v_{2,j} < v_{th} \; for \; all \; j$$

$$\text{and } TD_1 and \; TD_2 < TD_{th}$$

where:

v is the speed for observation i or j

v_{th} is the speed threshold (0.75 m/sec)

TD is the Euclidian distance between the transit station and the centroid of the positions of the stationary sub-segment points

TD_{th} is the proximity threshold (30m).

Thirty meters was empirically chosen as a threshold value for TD_{th} to allow for error in the GPS data, the range of actual stopping points for transit vehicles in the vicinity of the GIS point identifying the transit stop, and the length of the transit vehicle. In this application, 30 meters balanced the conflicting problems of failing to recognize proximate stops and identifying stops beyond the area of interest. This value should be established based on local conditions for different applications.

6.1.2.2 Transit Stop Rate (*TSR*) estimation and application

The data set is refined a third time using information about the frequency and location of stops (i.e. locations at which the traveler is stationary). For transit trips, the stopping pattern is expected to be more frequent and have a stronger correlation (spatially) with the location of transit stations than for private vehicles. Figure 6-4 illustrates a space-time diagram for a two hypothetical trips; the first trip presents a private auto mode and the second trip presents a transit mode. However, one complicating factor arises. Many transit stations are located in the vicinity of signalized intersections. Both transit vehicles and private vehicles are impacted by the traffic signals and, therefore, at these locations very similar stopping patterns can be observed between private cars and transit vehicles. To control for this phenomenon, the transit stations that are proximate to signalized intersections were eliminated from the analysis; a value of 50m was chosen as the proximity threshold from the signalized intersection (SD_{th}) based on the physical properties of the intersections in the data set. This value is dependent on the network characteristics and therefore should be calibrated for local conditions.

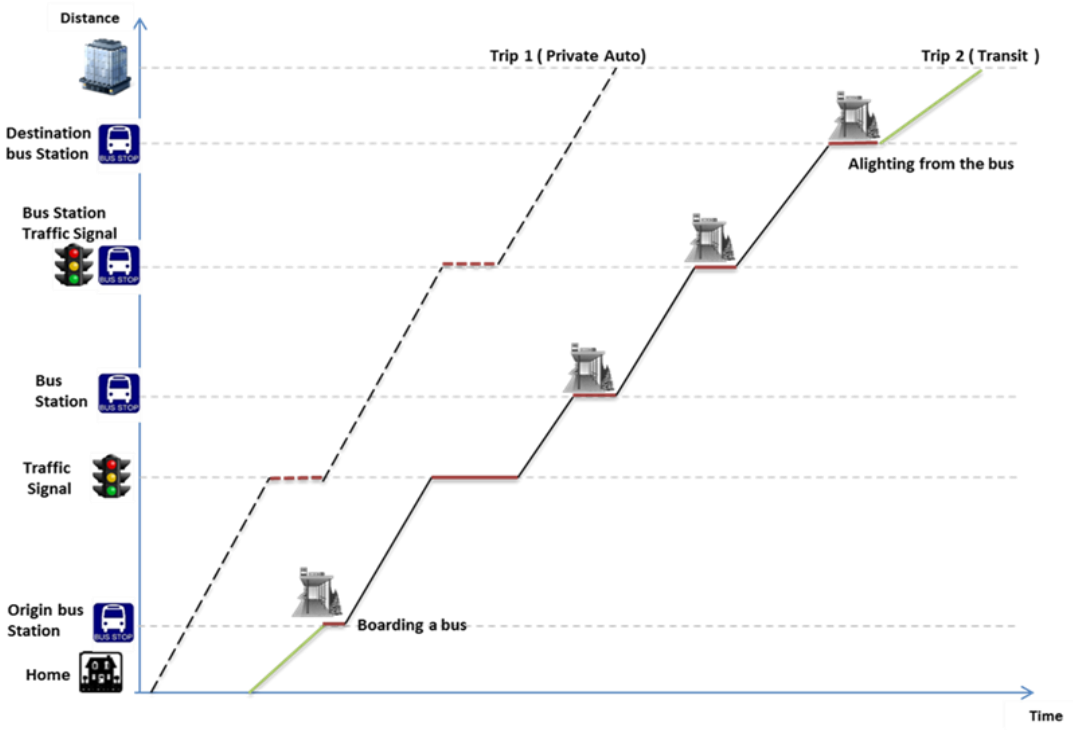


Figure 6-4 Space-time diagram for a hypothetical example for two trips by private-auto and transit

As discussed previously, other researchers have proposed comparing the number of stops per trip as a way to distinguish between auto and transit modes. In this research, a similar approach was adopted; however, this metric was improved by calculating the number of stops per distance traveled which presents a more robust differentiating factor, as the following example demonstrates.

Suppose two trips are made, one is 10km the other is 3km long. In the first trip, the GPS data indicate four stops in proximity to transit station of which three are in proximity to a signalized intersection and one is away from a traffic signal. For the second trip, the GPS data indicate three stops, again with only one occurring away from a traffic signal. Both trips have one stop unrelated to a traffic signal and therefore we must conclude that both trips are equally likely to have been made using transit.

A better metric can be calculated as follows. First, the number of stops occurring at transit stations but away from traffic signals is computed. Next, the ratio of these stops to length is calculated. The output of this calculation, shown in equation (6-1), is defined as the transit stop rate (*TSR*).

$$TSR = \frac{NStop_{TS}}{D} \tag{6-1}$$

where

$NStop_{TS}$ is the number of stops in proximity to transit stop and not in proximity to signalized intersections

D is the cumulative distance for the trip

Returning to the previous example, the estimated *TSR* for trip 1 is 0.1 stops per km while for trip 2 the *TSR* is 0.33. From these results, trip 2 is more likely to have been made by transit than trip 1. To apply this technique to the full data set, the *TSR* is calculated for every potential transit trip. Then, the distribution of these *TSR* values is plotted and a threshold value, TSR_{th} is established; TSR rates that exceed this threshold identify trips for which further analysis should be considered. The appropriate threshold is quantified by comparing the cumulative distribution functions for *TSR* for known transit and non-transit trips.

Figure 6-5 shows the cumulative distribution functions for *TSR* generated from the data for transit trips and private automobile trips. About 80% of private automobile trips have a *TSR*

less than 0.3 stops per km. Only 25% of transit trips have a *TSR* less than this value. As a result, all trips from the data set with a *TSR* less than 0.3 stops/km are eliminated.

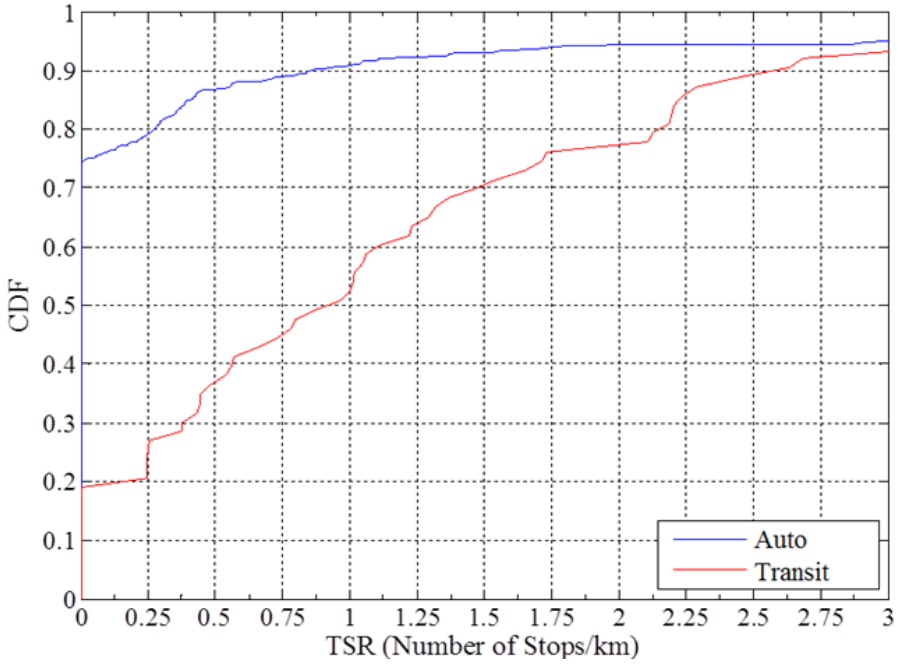


Figure 6-5 TSR for transit and private auto trips

6.1.2.3 Start and end points of the transit segment identification

Up to this point, a subset of potential transit trips was identified for which some level of confidence exists that at least one sub-segment was conducted by transit. In the next step of the analysis, the boundaries of the transit sub-segment are identified. The start and end points of the transit component for each of these trips is identified as follows. The Potential Transit Starting Point (*PTSP*) is defined as the first point in a non-stationary sub-segment that satisfies the following conditions:

- The sub-segment to which the point belongs has a maximum speed exceeding a threshold (VM_{th}); the threshold was chosen to be the 90th percentile of the non-motorized maximum speed. In other words, the start point of the transit trip must belong to a segment with speeds that reflect motorized travel.
- The point is located within the proximity threshold of a transit station;
- The preceding moving sub-segment was labeled by the original classifier as non-motorized (i.e. walk or bike to the transit stop);

The Potential Transit End Point (*PTEP*) is defined as the last point in a non-stationary sub-segment that satisfies the following conditions:

- The sub-segment to which the point belongs has a maximum speed exceeding a threshold (VM_{th});
- The point is located within the proximity threshold of a transit station;
- The following sub-segment is classified by the original classifier as a non-motorized segment.

To determine an appropriate threshold value for maximum speed (VM_{th}), the cumulative distribution functions for motorized and non-motorized trips were computed from the labelled data (Figure 6-6).

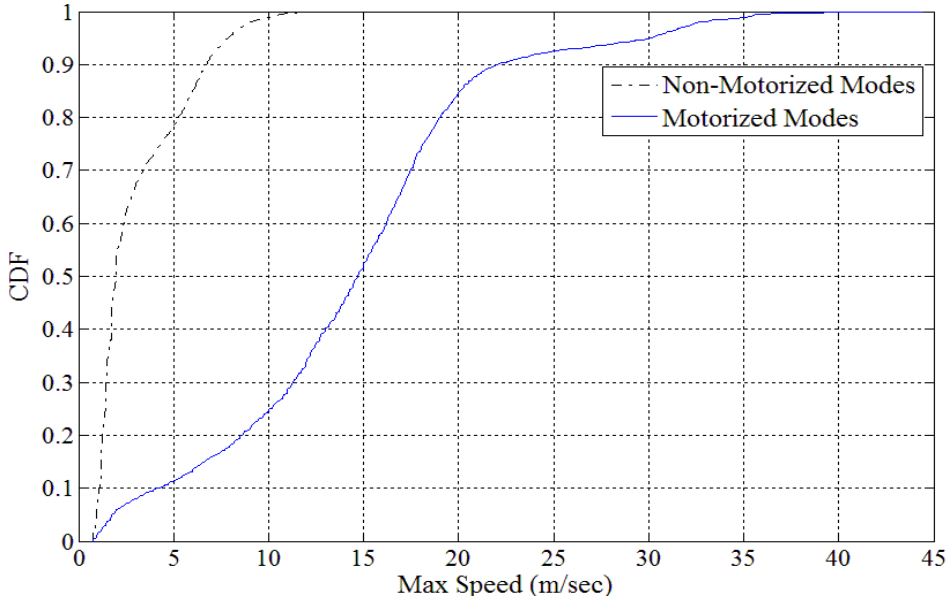


Figure 6-6 Cumulative distribution of maximum segment speed for motorized versus non-motorized modes

In this case, it was observed that 90% of non-motorized segments have a maximum speed that is less than 6.75m/s (24 km/h). Less than 18% of motorized segments exhibit maximum speeds lower than this value. Therefore, this value was selected for the threshold (VM_{th}) implying that the sub-segment containing a potential transit start point must have a maximum speed > VM_{th} = 6.75 m/s.

Based on these definitions, the algorithm employs a stepwise (forward and backward) approach to determine limits of a transit mode segment. Suppose there are *k* stops on a trip.

The algorithm begins at stop 1 and assesses whether this stop meets the criteria for a *PTSP*. If so, the stop is labeled as the beginning point of a transit segment. If not, then the algorithm iterates to stop 2 and the process continues until either a transit starting point is found at stop *i* or all *k* points have been evaluated and no transit start point has been identified. In the latter case, the trip is then eliminated from further consideration and the segment labels remain as determined by the original classification model.

In the former case, the algorithm then seeks a transit segment end point beginning at point *k.* If point *k* satisfies the requirements, then the transit segment is bounded by points *i* and *k.* If not, the algorithm iterates to *k-1* and tests for suitability. This iterative process continues for *n* iterations such that $k-n>i$. If no suitable end point is found, the trip is eliminated; if a suitable end point is found in iteration *n*, then the transit segment is bounded by points *i and k-n.*

6.2 Results

In order to evaluate the developed model, the GPS data described earlier and the results from the previous chapter have been used. The total number of trips is 665, containing over 105,000 points representing transportation modes – i.e. not engaging in an activity (shopping, etc.). As presented in Table 6-1, the dataset contained 8993 points for which the mode of travel was transit. The optimal classifier determined in the previous chapter did not make use of spatiotemporal data and correctly identified only 27.8% of these transit points. The purpose of this model is to improve these results.

Table 6-1 The optimized k-NN classification model results

		Classified as				Total	Proportion	Recall
		Walk	**Bike**	**Transit**	**Auto**			
Reality	**Walk**	19581 (97.77%) (**80.48%**)	78 (0.39%) (**0.54%**)	171 (0.85%) (**5.96%**)	198 (0.99%) (**0.31%**)	20028	18.96%	97.77%
	Bike	2027 (14.08%) (**8.33%**)	12013 (83.46%) (**82.67%**)	0 (0.00%) (**0.00%**)	353 (2.45%) (**0.55%**)	14393	13.62%	83.46%
	Transit	1497 (16.65%) (**6.15%**)	558 (6.20%) (**3.84%**)	2462 (27.38%) (**85.81%**)	4476 (49.77%) (**7.00%**)	8993	8.51%	27.38%
	Auto	1226 (1.97%) (**5.04%**)	1882 (3.02%) (**12.95%**)	236 (0.38%) (**8.23%**)	58880 (94.63%) (**92.13%**)	62224	58.90%	94.63%
Total		24331	14531	2869	63907	***Overall Accuracy 87.98%***		
Proportion		23.03%	13.76%	2.72%	60.50%			
Precision		80.48%	82.67%	85.81%	92.13%			

(xx) = Recall = number of points classified as mode *n* divided by total # of actual points of mode *m* times 100%.
(**yy**) = Precision = number of mode *m* points divided by the total number of points classified as mode *n*× 100%.

The process described in Section 6.1.1 – to identify Initial Potential Transit Trips – was applied to eliminate all trips for which all non-stationary segments have been labelled as non-motorized. The result of this step is the retention of only 501 trips labeled as initial potential transit trips (*IPTT*).

The spatial statistics and GIS information then have been deployed following the same order in the methodology. First, the trips in *IPTT* that have no stops (i.e. non-stationary segment or point) in proximity to transit station locations were eliminated. This further reduces the number of trips to 323.

Next, the transit stopping rate (*TSR*) filter was applied. The first step was estimating *TSR* for each of the different motorized modes. As discussed earlier, based on the TSR_{th}, all trips from the data set with a *TSR* less than TSR_{th} (0.3 stops/km) were eliminated. As a result, this filter identified an additional 236 trips that do not meet the *TSR* requirement, leaving 87 potential transit trips. For these trips, the stepwise algorithm to identify the start and end points of the transit trip is applied.

The integrated GIS model identified 87 trips which had been classified originally as being made by auto, but which have not been estimated to have been made by transit. Of these 87 trips, 57 actually contain at least one transit segment; the remaining 30 did not contain any transit segments. On the other hand, six trips that actually contain at least one transit segment were eliminated by the integrated GIS model – i.e. classified as not containing any transit segments. A further analysis of these trips suggests that the conservative value for the stopping rates as well as unusual GPS errors lead to the misclassification. Table 6-2 provides the confusion matrix for the disaggregated results at the point level when the proposed method for including the spatiotemporal information is applied. The impact of the proposed method can be determined by comparing the results in Table 6-2 with those from the optimized k-NN model (Table 6-1).

Table 6-2 Confusion Matrix after applying GIS integrative model

		Classified as				**Total**	**Proportion**	**Recall**
		Walk	**Bike**	**Transit**	**Auto**			
Reality	**Walk**	19528 (97.50%) **(84.24%)**	91 (0.45%) **(0.67%)**	309 (1.54%) **(3.40%)**	100 (0.50%) **(0.17%)**	20028	18.96%	97.50%
	Bike	2167 (15.06%) **(9.35%)**	11911 (82.76%) **(88.28%)**	5 (0.03%) **(0.06%)**	310 (2.15%) **(0.52%)**	14393	13.62%	82.76%
	Transit	368 (4.09%) **(1.59%)**	45 (0.50%) **(0.33%)**	7698 (85.60%) **(84.73%)**	882 (9.81%) **(1.47%)**	8993	8.51%	85.60%
	Auto	1119 (1.80%) **(4.83%)**	1445 (2.32%) **(10.71%)**	1073 (1.72%) **(11.81%)**	58587 (94.15%) **(97.84%)**	62224	58.90%	94.15%
Total		23182	13492	9085	59879	***Overall Accuracy 92.51%***		
Proportion		21.94%	12.77%	8.60%	56.68%			
Precision		84.24%	88.28%	84.73%	97.84%			

(xx) = Recall = number of points classified as mode *n* divided by total # of actual points of mode *m* times 100%.
(**yy**) = Precision = number of mode *m* points divided by the total number of points classified as mode *n*× 100%.

Recall that the objective was to improve the classification of transit points. The results in Table 6-2 demonstrate that this goal was achieved. The change in the classification performance (recall and precision) are summarized and presented in Figure 6-7 (positive values indicate improvement). We have improved the recall of transit from 27.8% to 85.6% (an

increase in recall of 58% compared to the original classifier). We also note that there is a small increase in the number of walk and auto points which are incorrectly labelled as transit. However, the overall precision for each transportation mode is still improved.

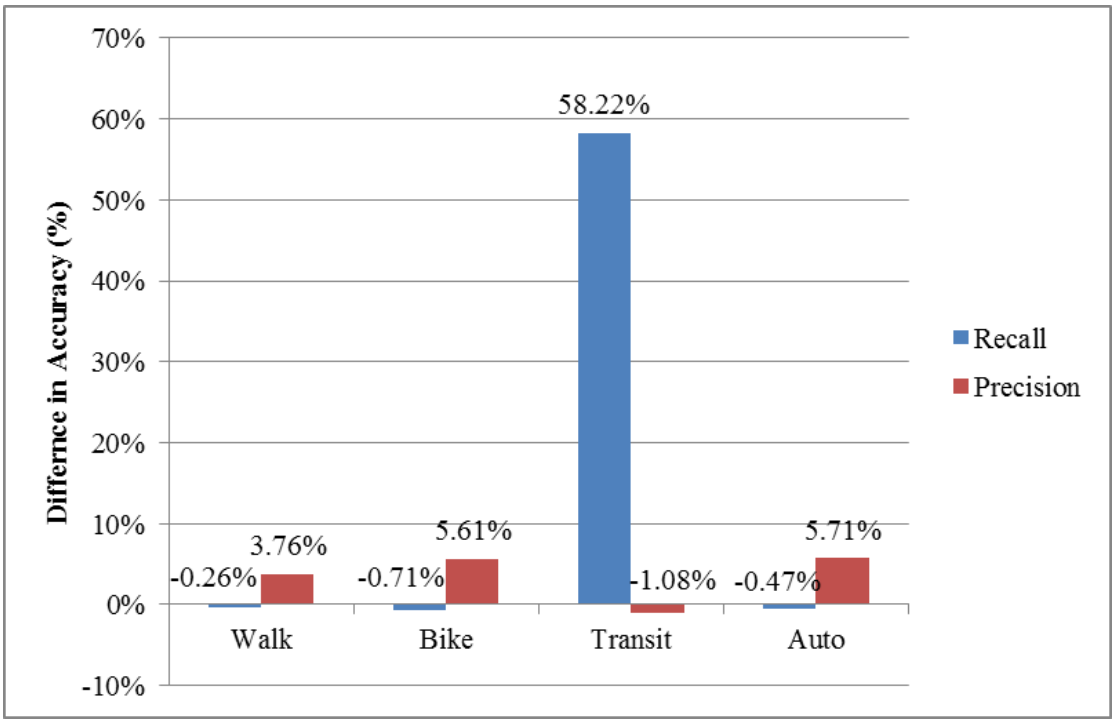


Figure 6-7 Impact of proposed method for including spatiotemporal data on classification results

6.3 GIS model assessment

In order to assess the performance of the GIS model in identifying transit mode, the same trip presented at the end of the previous chapter and shown in Figure 5-21 was investigated through the developed GUI (see Figure 6-8). As mentioned before, the trip consists of two *MTP* and three *TMS* as follows: Walk-Transit-Walk. The GIS model correctly identified the location of start and end of transit segment and relabeled all the points in between correctly as transit achieving 100% accuracy in identifying the correct labels. Consequently, the results of *PMTP* were correctly located as shown on the diagrams with the green arrows. Overall, the introduction of the GIS model reduced the number of case 1 incorrect *PMTPs* in the model by 21% (i.e. from 439 to 347).

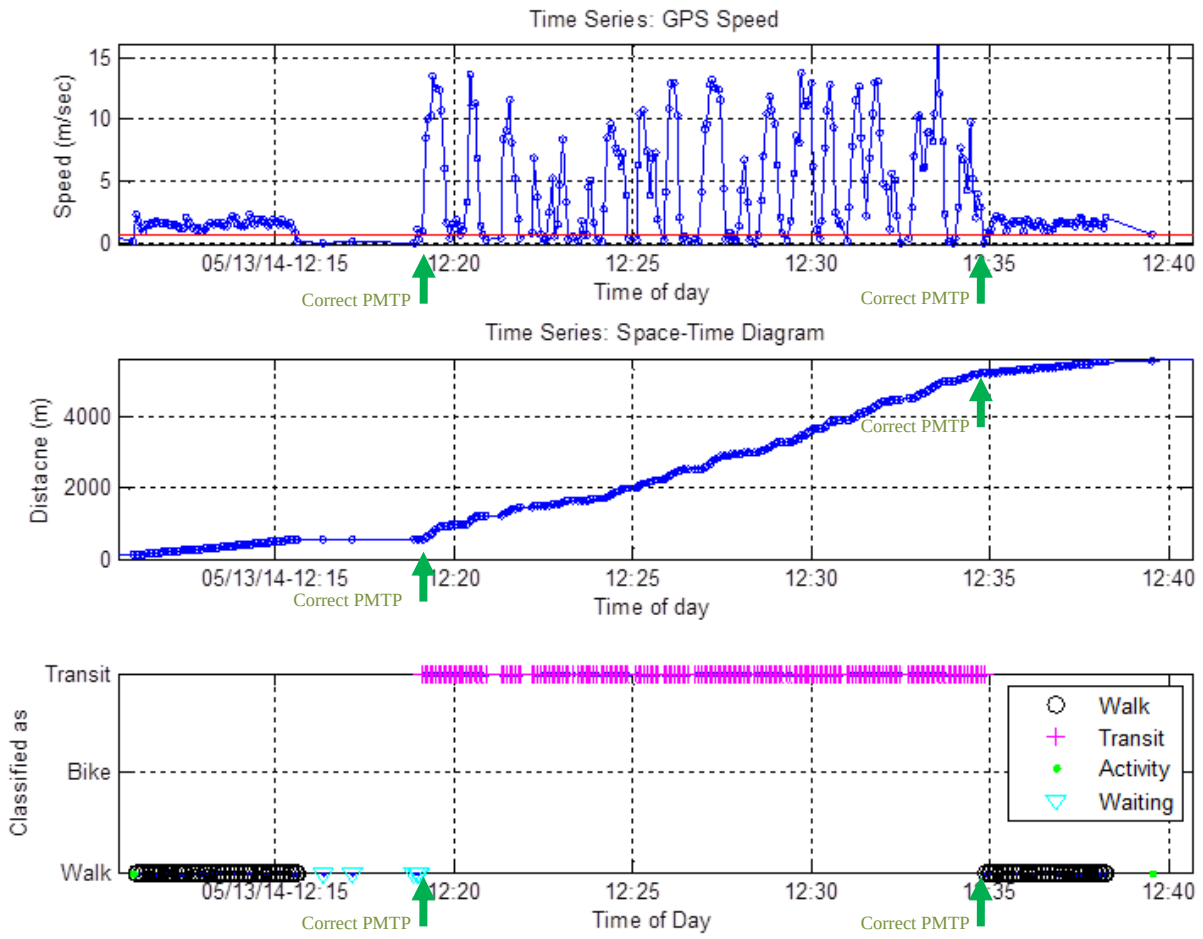


Figure 6-8 GUI diagram for a transit trip

6.4 Summary

In this chapter, a proposed model has built upon the optimized model from the previous chapter to classify travel modes from GPS data. The optimized model was successful in identifying trips (at the point and segment levels) as motorized and non-motorized. The model performed less satisfactorily in differentiating between auto and bus transit modes, primarily due to the fact that bus transit in the study area operates in mixed traffic and therefore has properties which are very difficult to distinguish from private auto. In lieu of further complicating the original model, a complementary procedure was built that improves the combined models' performance in correctly identifying transit trips.

The approach presented here integrated the classifications from the previous model with additional information derived from both the GPS data and GIS information. More

specifically, the analysis was limited to *IPTTs*, trips containing at least one motorized segment. Then, the *IPTT* dataset was further refined to contain only trips with stationary segments that spatially coincided with the location of transit stations. Next, the dataset was filtered again to eliminate those trips for which the transit stop rate – a metric that isolates the stopping pattern at transit stations as opposed to traffic signals – failed to exceed an empirically derived threshold. Then, a stepwise algorithm was applied that cycled through stop locations to find logical transit starting and end points. All transit segments bounded by the start and end point were labeled as transit. If a start or end point were not found, the trip was eliminated from the analysis set.

The application of the proposed method resulted in a vast improvement in the classification of transit trips, with only minor degradation in the classification of other modes – walking particularly. With the application of the spatiotemporal methods described here, the transit recall was improved from 27% to 85%, an increase of nearly 60%. The proposed method can be applied to the results of the k-NN model proposed in the previous chapter. In terms of model transferability, the model parameters need to be calibrated to local conditions using a small sample of labelled data as well as network characteristics.

Chapter 7
Conclusions and Recommendations

7.1 Conclusions

Transportation planning and engineering rely on valid transportation data – origins, destinations, departure times, modes and paths. Previously, these data were gathered through manual processes, often with limited accuracy. A growing field in transportation research is the collection, analysis and interpretation of automatically generated travel data from Global Positioning Systems (GPS). With GPS-enabled smartphones becoming nearly ubiquitous, transportation researchers now possess a very rich data set from which travel behavior can be gathered and be used to inform transportation infrastructure investments as well as policy analysis. Recently, GPS data have been utilized widely for collecting travel survey data, largely because of the lower burden on survey respondents and the higher accuracy as compared to more traditional trip diary surveys.

The main challenge of utilizing passively collected raw GPS data is inferring travel behavior and trip characteristics. One common and important question addressed in the literature is how to use passively collected GPS data to determine the transportation mode of travel. The development and implementation of a method to improve upon previous solutions to this problem is the focus of this research.

In order to achieve the main goal of this research, the quality of the collected and stored data should be assured. GPS data are subject to different sources of error. In addition, labelled data were collected to permit model calibration and evaluation. Survey participants were asked to label their transportation modes when they performed their trips using the smart phone application. However, these labels were prone to errors due to the subjectivity of the decisions of where and when the traveller chose to change the transportation mode label. It was necessary to develop and apply methods for identifying and correcting these errors so that a reliable data set was available for model development.

Developing an accurate transportation mode inference model is a non- trivial task, mainly due to the complex nature of mixed-traffic that causes high variability and overlap in the different transportation modes' attributes in this environment. Mode classification on the basis

of GPS data requires the following three steps: (1) selecting a type of classification model to be used; (2) selecting attributes to be used in the feature vector; and (3) calibrating model parameters.

In this thesis we addressed these challenges and proposed an integrated framework for a robust self-optimizing model that can convert raw GPS data, passively collected through smartphones, to a set of trips with highly accurate transportation mode labels. The main contributions of this research and the direction for further studies are described in the remainder of this chapter.

7.2 Major Contributions

The work described in this thesis has resulted in the following five major contributions: (1) Development of a custom software tool that can be used to investigate and review the smart phone data in a dataset and to screen these data for data anomalies or errors; (2) Development of an algorithm to automatically correct mislabelled mode transfer points (*MTPs*); (3) Development and implementation of a quantitative method to rank attributes on the basis of their differentiating power as an aid for selecting attributes to be included in the feature vector; (4) Development and evaluation of an objectively optimized classification model by which transportation mode can be identified automatically; (5) Development and evaluation of a GIS-based method to enhance the performance of the transportation mode classifier. The following sections describe each of these contributions in more detail.

1. *Development of a custom software tool to investigate and review the trips in the dataset to ensure the quality of saved trips*: A custom software tool, which permits simultaneous quantitative reporting and data visualization, was developed. The tool permits the spatiotemporal attributes of trips, including space-time diagrams, speed-time diagrams, and actual versus model's labels to be viewed and / or manipulated. Using the tool, a user can easily accomplish the following:
    a. understand specific type of behavior associated with each of the different transportation modes;
    b. comprehend the behavior of the developed model and algorithms;

c. help to identify and justify the chosen criteria in the developed criterion algorithms.

d. investigate the identified suspicious mislabelled observations and interactively switch between plots and the data source table in order to correct them.

2. *Development of an algorithm to automatically correct mislabelled mode transfer points* (*MTPs*): This research introduced a method to systematically identify errors in the labelled *MTPs* and to automatically correct these errors. The algorithm begins with defining a set of criteria for describing the circumstances associated with Mode Transfer Points, the points where users change from one to another transportation mode, to ensure that the participants' indicated the change in transportation mode at logically correct times and locations. This method aims to correct MTPs by shifting the mode transfer label forward or backward along the time axis. Correctly identifying *MTPs* is essential for accurately reporting the performance of the classification model.

3. *Development and implementation of a quantitative method to rank attributes on the basis of their differentiating power as an aid for selecting attributes to be included in the feature vector*: The common practice in previous research is to select a set of attributes to be contained in the feature vector based on either assumptions or general methods. The novel ideas in this research are to add and order attributes to the feature vector based on their differentiating power calculated through the *p values* of pairwise comparisons. The proposed method then translates the pairwise comparisons to a single attribute differentiating power (ADP_a) statistic for each attribute using a different weight (calibrated endogenously in the model) for some pairwise comparisons that are more important than others, specifically those attributes that differentiate between motorized and non-motorized modes.

4. *Develop and evaluate an objectively optimized classification model by which transportation mode can be identified automatically:* In this research, an objective optimization of the proposed classification model has been completed following an efficient full factorial analysis which allowed us to investigate the significance of the model parameters (both main effects and interactions).The parameters in the proposed

optimization method include: the selection of the type of classification model, the number of feature vector attributes, the type of attributes to include in the feature vector, and whether or not to apply transformation techniques (*e.g.*, discrete versus continuous and PCA) to the feature vector attributes. The resulted optimized inference model requires fewer and less restrictive assumptions about the trip structure than most existing models and unlike many previous studies, the model was evaluated under the more realistic conditions that the segmentation of trip data is part of the transportation mode estimation problem. It must be noted, that these methods must be configured so that they do not require the trip maker to record or identify trip attributes.

5. *Development and evaluation of a GIS-based method to boost the transportation mode classifier*: The main goal of the GIS based inference model is improving the accuracy of identifying transit modes using the additional information about modal attributes obtained by integrating the GPS data with Geographic Information Systems (GIS). Due to the fact that bus transit in the study area operates in mixed traffic and therefore has properties which are very difficult to distinguish from private auto, spatial attributes have been estimated based on the transportation infrastructure information (i.e. transit stations and signalized intersections) that resulted in a vast improvement in the classification of transit. In addition, a stepwise algorithm was developed that cycled through stop locations to find logical transit starting and end points to specify the correct boundary of transit segments.

7.3 Future Research

The recommended future research is divided into two categories, namely methodology and application. For the methodological part, the following extensions to the existing research are recommended:

1. In order to test the robustness and transferability of the mode, the developed framework should be tested on a different set of smartphone data. Of particular interest may be to test the model with data generated from a more heterogeneous set of travelers, traveling over a larger spatial area, in more varied conditions (i.e. levels of congestion).

2. In this research, activities are assumed to be known. Very short activities – those lasting about 5 minutes – are not differentiated from major activities – events lasting several hours or more. The performance of the model may be enhanced by attempting to recognize activities of different durations. Consider the example where two, 10 minute auto trips surround a five minute activity. Merging these events such that the 20 minutes of moving segments inform the model classification will likely improve the model's performance compared to considering the two 10 minute trips in isolation.

3. In this research, four transportation modes (i.e. Walk, Bike, Auto, and Bus Transit) were used to test the performance of the developed framework. It is recommended that the developed framework to be applied on a data-set with a large variety of transportation modes including LRT, Train, etc.

4. The proposed model does not make any assumptions about the likelihood of sequencing of transportation modes (*e.g.*, probability of transferring from one mode to another mode). While this means the proposed model is also more flexible than models that rely on mode transfer probabilities, the model results suggest that improved mode performance might be achieved by imposing some constraints on the mode transfers. Therefore, it is recommended that the model to be extended to use more features to assess the logical sequence of the identified transportation mode within a trip.

5. An automated robust model is recommended to calibrate the parameters used in the GIS model (i.e. the distance from transit station and signalized intersection) as a function of the transportation infrastructure.

6. In this research, GIS model aims to specify the start and end of transit segment within the trips and label all the points as Transit. However, the current model does not consider the possibility of transfers between transit routes. Therefore, a more sophisticated model is required to capture those transfers between different routes.

Regarding the application part, this research can be extended by using the obtained results in different applications as follows:

1. As mentioned previously in chapter 1, the accuracy of many transportation models or assessments depends heavily on the quality of the user information provided. Typically, this information has been gathered through surveys, with mixed results. Therefore, the

outcomes from this research can represent a significant improvement in the quantity and quality of traveler behavior data that are input into these kinds of models. More specifically, activity-based models are increasingly being developed and applied. These models rely heavily on accurate data around trip-chaining and trip tours. Passively gathered, GPS data that contain correct mode inferences have the potential to inform these activity based models at very low cost and with high accuracy.

2. Individualized marketing (IM) can be one of the most significant applications using the results from this research. IM in transportation is a customer oriented technique that promotes the use of more sustainable transportation modes such as public transportation, walking, and biking. The technique is based on a targeted, personalized, and customized marketing approach that encourages travelers to change their travel behaviour in mode choice. It establishes direct contact with travelers at an individual level to identify their trips' needs and requirements and then provides information back for more sustainable transportation modes alternatives that match their trips' needs. The IM technique requires establishing existing travel behaviour, providing customized information to interested participants, and then measuring changes in travel behaviour. Therefore, through the results obtained from this research, trips profile can be easily generated in order to identify the most appropriate alternatives available at individual level.

Appendix A

A-1 Linear regression model results (response variable = *MCR*)

Variable Name	Description	Coefficient	Confidence Interval	
			Lower	Upper
Constant		0.1451	0.1376	0.1526
θ_{NF2}	*NF2=3*	-0.0383	-0.0505	-0.0262
θ_{NF7}	*NF7=10*	-0.0165	-0.0286	-0.0043
θ_{CL2}	*CL2=k-NN*	-0.0251	-0.0334	-0.0168
θ_{CL3}	*CL3=QDA*	-0.0353	-0.0436	-0.0269
θ_{PCA}	*PCA*	-0.0199	-0.0278	-0.0120
θ_{Disc}	Discretization	-0.0168	-0.0248	-0.0089
$\theta_{NF1,CL2}$	*NF1=1* and *CL2=k-NN*	0.0809	0.0653	0.0965
$\theta_{NF2,CL2}$	*NF2= 3* and *CL2=k-NN*	0.0207	0.0051	0.0363
$\theta_{NF3,CL2}$	*NF2= 5* and *CL2=k-NN*	-0.0177	-0.0333	-0.0021
$\theta_{NF7,CL2}$	*NF7= 10* and *CL2=k-NN*	0.0215	0.0059	0.0372
$\theta_{NF1,CL3}$	*NF1= 1* and *CL3=QDA*	0.0340	0.0184	0.0496
$\theta_{NF2,CL3}$	*NF2= 3* and *CL3=QDA*	0.0387	0.0231	0.0543
$\theta_{NF7,CL3}$	*NF7= 1*0 and *CL3=QDA*	0.0188	0.0032	0.0345
$\theta_{NF1,PCA}$	*NF1=1* and *PCA*	0.0193	0.0084	0.0301
$\theta_{NF2,PCA}$	*NF2=3* and *PCA*	0.0277	0.0168	0.0386
$\theta_{NF7,PCA}$	*NF7= 10* and *PCA*	0.0183	0.0075	0.0292
$\theta_{NF1,Disc}$	*NF1=1* and *Disc*	0.0127	0.0018	0.0236
$\theta_{NF2,Disc}$	*NF2=3* and *Dsic*	0.0201	0.0093	0.0310
$\theta_{CL3,PCA}$	*CL3=QDA* and *PCA*	0.0154	0.0054	0.0254
$\theta_{CL2,Disc}$	*CL2=k-NN* and *Disc*	0.0180	0.0080	0.0280
$\theta_{CL3,Disc}$	*CL3=QDA* and *Disc*	0.0279	0.0179	0.0379
$\theta_{NF3,CL2,PCA}$	*NF3= 5, CL2=k-NN,* and *PCA*	0.0139	0.0006	0.0272
$\theta_{NF7,CL2,PCA}$	*NF7= 10, CL2=k-NN,* and *PCA*	-0.0276	-0.0409	-0.0142
$\theta_{NF8,CL2,PCA}$	*NF8= 11, CL2=k-NN,* and *PCA*	-0.0180	-0.0313	-0.0047
$\theta_{NF2,CL3,PCA}$	*NF2= 3, CL3=QDA,* and *PCA*	-0.0220	-0.0353	-0.0087
$\theta_{NF7,CL3,PCA}$	*NF7= 10, CL3=QDA,* and *PCA*	-0.0200	-0.0333	-0.0067
$\theta_{NF1,CL2,Disc}$	*NF1=1, CL2=k-NN,* and *Disc*	0.0635	0.0502	0.0768
$\theta_{NF4,CL2,Disc}$	*NF4=6, CL2=k-NN,* and *Disc*	0.0146	0.0013	0.0279
$\theta_{NF8,CL2,Disc}$	*NF8=11, CL2=k-NN,* and *Disc*	0.0213	0.0080	0.0346
$\theta_{NF1,CL3,Disc}$	*NF1=1, CL3=QDA,* and *Disc*	-0.0253	-0.0386	-0.0120
$\theta_{NF2,CL3,Disc}$	*NF1=1, CL3=QDA,* and *Disc*	-0.0175	-0.0308	-0.0042
$\theta_{NF2,PCA,Disc}$	*NF2=3, PCA,* and *Disc*	-0.0160	-0.0269	-0.0052
Adjusted R^2*=0.6658*	F_{stat}*=34.04*	*p*-value = 0.0004		

A-2 Correlation Coefficients

		Speed				Acceleration				Jerk		
		Average	St. Dev.	Max	98th Perc	Max	Min	98th Per	98-50th Per	Max	98th Per	98-50th Per
Speed	Average	**1.00** (1.00)	0.82 ***(0.00)***	0.92 ***(0.00)***	0.93 ***(0.00)***	0.27 ***(0.00)***	-0.35 ***(0.00)***	0.27 ***(0.00)***	0.27 ***(0.00)***	0.07 ***(0.03)***	0.05 (0.10)	0.06 (0.09)
	St. Dev.	0.82 ***(0.00)***	**1.00** (1.00)	0.94 ***(0.00)***	0.96 ***(0.00)***	0.36 ***(0.00)***	-0.48 ***(0.00)***	0.46 ***(0.00)***	0.46 ***(0.00)***	0.08 ***(0.01)***	0.09 ***(0.01)***	0.09 ***(0.01)***
	Max	0.92 ***(0.00)***	0.94 ***(0.00)***	**1.00** (1.00)	0.99 ***(0.00)***	0.45 ***(0.00)***	-0.55 ***(0.00)***	0.48 ***(0.00)***	0.48 ***(0.00)***	0.20 ***(0.00)***	0.19 ***(0.00)***	0.19 ***(0.00)***
	98th Perc	0.93 ***(0.00)***	0.96 ***(0.00)***	0.99 ***(0.00)***	**1.00** (1.00)	0.37 ***(0.00)***	-0.47 ***(0.00)***	0.43 ***(0.00)***	0.43 ***(0.00)***	0.11 ***(0.00)***	0.11 ***(0.00)***	0.11 ***(0.00)***
Acceleration	Max	0.27 ***(0.00)***	0.36 ***(0.00)***	0.45 ***(0.00)***	0.37 ***(0.00)***	**1.00** (1.00)	-0.86 ***(0.00)***	0.88 ***(0.00)***	0.88 ***(0.00)***	0.90 (3.10)	0.85 ***(0.00)***	0.85 ***(0.00)***
	Min	-0.35 ***(0.00)***	-0.48 ***(0.00)***	-0.55 ***(0.00)***	-0.47 ***(0.00)***	-0.86 ***(0.00)***	**1.00** (1.00)	-0.81 ***(0.00)***	-0.81 ***(0.00)***	-0.77 ***(0.00)***	-0.73 ***(0.00)***	-0.73 ***(0.00)***
	98th Per	0.27 ***(0.00)***	0.46 ***(0.00)***	0.48 ***(0.00)***	0.43 ***(0.00)***	0.88 ***(0.00)***	-0.81 ***(0.00)***	**1.00** (1.00)	1.00 ***(0.00)***	0.75 ***(0.00)***	0.80 ***(0.00)***	0.80 ***(0.00)***
	98-50th Per	0.27 ***(0.00)***	0.46 ***(0.00)***	0.48 ***(0.00)***	0.43 ***(0.00)***	0.88 ***(0.00)***	-0.81 ***(0.00)***	1.00 ***(0.00)***	**1.00** (1.00)	0.75 ***(0.00)***	0.80 ***(0.00)***	0.80 ***(0.00)***
Jerk	Max	0.07 ***(0.03)***	0.08 ***(0.01)***	0.20 ***(0.00)***	0.11 ***(0.00)***	0.90 (3.10)	-0.77 ***(0.00)***	0.75 ***(0.00)***	0.75 ***(0.00)***	**1.00** (1.00)	0.95 ***(0.00)***	0.95 ***(0.00)***
	98th Per	0.05 (0.10)	0.09 ***(0.01)***	0.19 ***(0.00)***	0.11 ***(0.00)***	0.85 ***(0.00)***	-0.73 ***(0.00)***	0.80 ***(0.00)***	0.80 ***(0.00)***	0.95 ***(0.00)***	**1.00** (1.00)	1.00 ***(0.00)***
	98-50th Per	0.06 (0.09)	0.09 ***(0.01)***	0.19 ***(0.00)***	0.11 ***(0.00)***	0.85 ***(0.00)***	-0.73 ***(0.00)***	0.80 ***(0.00)***	0.80 ***(0.00)***	0.95 ***(0.00)***	1.00 ***(0.00)***	**1.00** (1.00)

(*p-value*) is the probability of getting a correlation as large as the observed value; if p-value <0.05, then the correlation is significant.

A-3 Student t-test

Mathematically , student *t*-test (Case 3: two samples with unequal sizes and unknown variances) to conduct pairwise comparisons between the distribution of the selected features for different transportation modes to determine if the two sets are significantly different from each other has been done as follows:

1. Compute t_{stat}

$$t_{stat} = \frac{\Delta(\mu_{TMP})}{SE} \tag{1}$$

where

$\Delta(\mu_{TMP})$ is the difference between sample means ($\mu_{m1} - \mu_{m2}$) of the feature for the transportation mode pair *TMP*

SE is the standard error of the difference between the two means

2. Compute SE (Standard Error)

$$SE = \sqrt{\frac{S_{m1}^2}{O_{m1}} + \frac{S_{m2}^2}{O_{m2}}} \tag{2}$$

where

S^2 is the unbiased estimator of the variance of the two samples

O is the number of observations in each sample

3. Calculate *df* (Degree of Freedom)

$$df = \frac{\left(\frac{S_{m1}^2}{O_{m1}} + \frac{S_{m2}^2}{O_{m2}}\right)^2}{\frac{\left(S_{m1}^2 / O_{m1}\right)^2}{O_{m1} - 1} + \frac{\left(S_{m2}^2 / O_{m2}\right)^2}{O_{m2} - 1}} \tag{3}$$

4. Compare t_{stat} with t_{crit} from t-table

If $|t_{stat}| > |t_{crit}|$, the test rejects the null hypothesis (H_0: the means of both samples come from populations with equal means) at $\alpha = 5\%$ confidence level. If not, the test fails to reject the null hypothesis. , which means the selected feature is a suitable for differentiating transportation modes.

A-4 Pairwise differentiating power (*DP*) results

$$DP_a^{m1,m2} = \left(1 - P_a^{m1,m2}\right)$$

		Speed				Acceleration				Jerk			
		walk	Bike	Transit	Auto	walk	Bike	Transit	Auto	walk	Bike	Transit	Auto
Average	walk	0.0000	1.0000	1.0000	1.0000	0.0000	0.9656	0.9680	0.7971	0.0000	0.9953	0.4365	0.9982
	Bike	1.0000	0.0000	1.0000	1.0000	0.9656	0.0000	0.9931	0.2419	0.9953	0.0000	0.7482	0.7599
	Transit	1.0000	1.0000	0.0000	1.0000	0.9680	0.9931	0.0000	0.9890	0.4365	0.7482	0.0000	0.7844
	Auto	1.0000	1.0000	1.0000	0.0000	0.7971	0.2419	0.9890	0.0000	0.9982	0.7599	0.7844	0.0000
Maximum	walk	0.0000	1.0000	1.0000	1.0000	0.0000	0.9900	1.0000	1.0000	0.0000	0.9892	0.9985	0.9869
	Bike	1.0000	0.0000	1.0000	1.0000	0.9900	0.0000	1.0000	1.0000	0.9892	0.0000	0.9994	1.0000
	Transit	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	0.9900	0.9985	0.9994	0.0000	0.9968
	Auto	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	0.9900	0.0000	0.9869	1.0000	0.9968	0.0000
Minimum	walk	0.0000	0.9991	0.9974	0.8655	0.0000	1.0000	1.0000	1.0000	0.0000	0.9932	0.9947	0.3051
	Bike	0.9991	0.0000	0.5171	0.9978	1.0000	0.0000	1.0000	1.0000	0.9932	0.0000	0.9977	1.0000
	Transit	0.9974	0.5171	0.0000	0.9949	1.0000	1.0000	0.0000	0.9946	0.9947	0.9977	0.0000	0.9942
	Auto	0.8655	0.9978	0.9949	0.0000	1.0000	1.0000	0.9946	0.0000	0.3051	1.0000	0.9942	0.0000
98th Percentile	walk	0.0000	1.0000	1.0000	1.0000	0.0000	0.9999	1.0000	1.0000	0.0000	0.9923	0.9990	0.9999
	Bike	1.0000	0.0000	1.0000	1.0000	0.9999	0.0000	1.0000	1.0000	0.9923	0.0000	0.9995	1.0000
	Transit	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	0.9992	0.9990	0.9995	0.0000	0.9971
	Auto	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	0.9992	0.0000	0.9999	1.0000	0.9971	0.0000
98th-50th Percentile	walk	0.0000	1.0000	1.0000	1.0000	0.0000	0.9999	1.0000	1.0000	0.0000	0.9912	0.9990	1.0000
	Bike	1.0000	0.0000	1.0000	1.0000	0.9999	0.0000	1.0000	1.0000	0.9912	0.0000	0.9996	1.0000
	Transit	1.0000	1.0000	0.0000	0.7758	1.0000	1.0000	0.0000	0.9992	0.9990	0.9996	0.0000	0.9971
	Auto	1.0000	1.0000	0.7758	0.0000	1.0000	1.0000	0.9992	0.0000	1.0000	1.0000	0.9971	0.0000
98th-50th Percentile	walk	0.0000	1.0000	1.0000	1.0000								
	Bike	1.0000	0.0000	1.0000	1.0000								
	Transit	1.0000	1.0000	0.0000	1.0000								
	Auto	1.0000	1.0000	1.0000	0.0000								

A-5 Attribute differentiating power (*ADP*) results

$$ADP_a = \frac{\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\left(DP_a^{m1,m2}\cdot\beta^{m_1,m_2}\right)}{\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}(\beta^{m_1,m_2})}$$

Feature	*DP* (different group)	*DP* (same group)	*ADP*
Average Speed	4.00	2.000	100.00
Speed Standard Deviation	4.00	2.000	100.00
98th Percentile Speed	4.00	2.000	100.00
Maximum Speed	4.00	2.000	100.00
98th-50th Percentile Acceleration	4.00	1.999	99.99
98th Percentile Acceleration	4.00	1.999	99.99
Maximum Acceleration	4.00	1.997	99.97
Minimum Acceleration	4.00	1.995	99.95
98th Percentile Jerk	4.00	1.989	99.86
98th-50th Percentile Jerk	4.00	1.988	99.85
Maximum Jerk	3.98	1.986	99.56
98th-50th Percentile Speed	3.80	1.390	89.90
Minimum Speed	3.38	1.994	87.50
Minimum Jerk	3.30	1.987	85.82
Average Acceleration	3.00	1.955	79.55
Average Jerk	2.94	1.780	76.65

References

Axhausen, K. W., Schonfelder, S., Wolf, J., Oliveria, M., Samaga, U. (2004) "Eighty Weeks of GPS Traces, Approaches to Enriching Trip Information", Transportation Research Board 83rd Annual Meeting Pre-print CD-ROM.

Asakura, Y., & Iryo, T. (2007). Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. Transportation Research Part A: Policy and Practice, 41(7), 684-690.

Barbeau, S., Labrador, M., Georggi, N., Winters, P., & Perez, R. (2009). TRAC-IT: A software architecture supporting simultaneous travel behaviour data collection and real-time location-based services for GPS-enabled mobile phones. Proceedings of the National Academy of Sciences' Transportation Research Board 88th Annual Meeting, Paper, pp. 09-3175.

Brög, W. (1998). Individualized marketing: Implications for transportation demand management. Transportation Research Record: Journal of the Transportation Research Board, 1618(-1), 116-121.

Brög, W., Erl, E., & Mense, N. (2002). Individualised marketing, changing travel behaviour for a better environment. OECD Workshop: Environmentally Sustainable Transport Berlin, , 5. pp. 06-12.

Casas, J., & Arce, C. (1999). Trip reporting in household travel diaries: A comparison to GPS-collected data. 78th Annual Meeting of the Transportation Research Board, Washington, DC,

Casello, J., Nour, A., & Hellinga, B. (2009). Quantifying impacts of transit reliability on user costs. Transportation Research Record: Journal of the Transportation Research Board, 2112(-1), 136-141.

Casello, J., Nour, A. O., Rewa, K. C., & Hill, J. (2011). Analysis of stated-preference and GPS data for bicycle travel forecasting. Transportation Research Board 90th Annual Meeting, (11-3478)

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

Dasarathy, B.V. (1990). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA

De Jong, R., & Mensonides, W. (2003). Wearable GPS device as a data collection method for travel research. Institute of Transport Studies Working Paper, (ITS-WP-03-02)

Defense Mapping Agency (1987). "Department of Defense World Geodetic System1984-lts definition and relationships with local geodetic systems." DMA Technical Report 8350.2, Defense Mapping Agency, Washington, D.C

Dill, J., & Mohr, C. (2010). Long term evaluation of individualized marketing programs for travel demand management No. OTREC-RR-10-08). Portland, Oregon: Oregon Transportation Research and Education Consortium.

Doherty, S., Noël, N., Gosselin, M., Sirois, C., & Ueno, M. (2001). Moving beyond observed outcomes: Integrating global positioning systems and interactive computer-based travel behaviour surveys. Transportation Research Circular, , 449-466.

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(-1), 147-153.

Du, J., & Aultman-Hall, L. (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. Transportation Research Part A: Policy and Practice, 41(3), 220-232.

Duda, R. O. & P. E. Hart (1973). Pattern Classification and Scene Analysis. New York: John Wiley & Sons.

Ettema, D., Timmermans, H., & van Veghel, L. (1996). Effects of data collection methods in travel and activity research.

Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning. Proc 13th International Joint Conference on Artificial Intelligence. Vol. 2, San Mateo, CA: Morgan Kaufmann; 1993. p 1022]1027.

Fukunaga, K., and L. Hostetler. 1973. Optimization of k nearest neighbor density estimates. Information Theory, IEEE Transactions on 19 (3): 320-6.

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N. L., *et al.* (2008). Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. 15th World Congress on Intelligent Transportation Systems.

iTrans. (2006). Cost of non-recurrent congestion in canada No. TP 14664E)Transport Canada.

Jun, J., Guensler, R., & Ogle, J. H. (2006). Smoothing methods to minimize impact of global positioning system random error on travel distance, speed, and acceleration profile estimates. Transportation Research Record: Journal of the Transportation Research Board, 1972(-1), 141-150.

Kumar, C., Basu, D., & Maitra, B. (2004). Modeling generalized cost of travel for rural bus users: A case study. Journal of Public Transportation, 7, 59-72.

Lachenbruch P. A and Mickey M.A. (1968) Estimation of error rates in discriminant analysis Technometrics 10, 1-10

Langley, P. & S. Sage (1994). Induction of selective Bayesian classifiers. In R. L´opez de Mantar´as & D. Poole(Eds.), Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann. (pp. 399–406)

Lee-Gosselin, M. (2002). Some reflections on GPS-supported travel survey methods in an increasingly ICT-rich environment. Presentation at Workshop on ICT, Innovation and the Transport System, Arlington,

Lucas, K., Blumenberg, E., & Weinberger, R. (2011). Auto motives: Understanding car use behaviours Emerald Group Pub Ltd.

Mahalanobis, P.(1936). "On the generalized distance in statistics". Proceedings of the National Institute of Sciences of India 2 (1): 49–55

McFadden, D., Talvitie, A., Cosslett, S., Hasan, I., Johnson, M., Reid, F., et al. (1977). Demand model estimation and validation Institute of Transportation Studies.

Moiseeva, A., Timmermans, H., & Jessurun, J. (2010). Semi-automatic imputation of long-term activity-travel diaries using GPS traces: Personal versus aggregate histories. 12th WCTR, Lisbon, Portugal.

Murakami, E., & Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting? Transportation Research Part C: Emerging Technologies, 7(2-3), 149-165.

Murakami, E., & Wagner, D. (1997). Comparison between computer-assisted self-interviewing using GPS with retrospective trip reporting using telephone interviews. Washington DC: US Department of Transportation,

Murphy, K. (1998). Dynamic Bayesian Networks: Representation, Inference and Learning. Doctoral Dissertation, University of California. Berkeley, CA, USA.

Nour, A. (2009). Quantifying the Impacts of Transit Reliability on Users' Costs—A Simulation Based Approach. MASc thesis. Department of Civil and Environmental Engineering, University of Waterloo, ON, Canada

Ogle, J., Guensler, R., Bachman, W., Koutsak, M., & Wolf, J. (2002). Accuracy of global positioning system for determining driver performance parameters. Transportation Research Record: Journal of the Transportation Research Board, 1818(-1), 12-24.

Ohmori, N., Nakazato, M., & Harata, N. (2005). GPS mobile phone-based activity diary survey. Proceedings of the Eastern Asia Society for Transportation Studies, , 5. pp. 1104-1115.

Ortúzar, J. de D. and L. G. Willumsen (2001). Modelling transport, 3rd Edition Chichester: Wiley

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA, USA: Morgan Kaufmann Publishers.

Rasmussen, T., Ingvardson, J. B., Halldórsdóttir, K., & Nielsen, O. A. (2013). Using wearable GPS devices in travel surveys: A case study in the Greater Copenhagen area. Transport conference at Aalborg University. ISSN 1603–9696.

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

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN), 6(2), 13.

Schönfelder, S., Li, H., Guensler, R., Ogle, J., Axhausen, K. W., Axhausen, K. W., et al. (2006). Analysis of commute atlanta instrumented vehicle GPS data: Destination choice behaviour and activity spaces ETH, Eidgenössische Technische Hochschule Zürich, IVT, Institut für Verkehrsplanung und Transportsysteme.

Schönfelder, S. and U. Samaga (2003). Where do you want to go today? - more observations on daily mobility, paper presented at the 3th Swiss Transport Research Conference, Ascona, March 2003.

Schüssler, N., Axhausen, K. W., Axhausen, K. W., & Axhausen, K. W. (2008). Identifying trips and activities and their characteristics from GPS raw data without further information ETH, Eidgenössische Technische Hochschule Zürich, IVT.

Shepard, Donald (1968). "A two-dimensional interpolation function for irregularly-spaced data". Proceedings of the 1968 ACM National Conference. pp. 517–524.

Stopher, P.R., P. Bullock and F. Horst (2002). "Exploring the Use of Passive GPS Devices to Measure Travel," in Proceedings of the 7th International Conference on Applications of Advanced Technologies to Transportation, edited by K.C.P. Wang, S. Medanat, S. Nambisan, and G. Spring, ASCE, Reston VA, pp. 959-967.

Stopher, P., & Collins, A. (2005). Conducting a GPS prompted recall survey over the internet. Proceedings of the 84th Annual Meeting of the Transportation Research Board,

Stopher, P., Jiang, Q., & FitzGerald, C. (2005). Processing GPS data from travel surveys. 2nd International Colloqium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, Toronto,

Stopher, P., Zhang, Y., Zhang, J., & Halling, B. (2009). Results of an evaluation of TravelSmart in south australia. Paper Accepted for Presentation to the 32nd Australasian Transport Research Forum, Auckland, New Zealand.

Tsui, S., & Shalaby, A. S. (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(-1), 38-45.

Wagner, D. (1997). Lexington area travel data collection test: GPS for personal travel surveys. Final Report, Office of Highway Policy Information and Office of Technology Applications, Federal Highway Administration, Battelle Transport Division, Columbus,

Witte, T., & Wilson, A. (2004). Accuracy of non-differential GPS for the determination of speed over ground. Journal of Biomechanics, 37(12), 1891-1898.

Wolf, J. (2004). Applications of new technologies in travel surveys. 7th International Conference on Travel Survey Methods, Costa Rica,

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(-1), 125-134.

Yalamanchili, L., Pendyala, R., Prabaharan, N., & Chakravarthy, P. (1999). Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behaviour. Transportation Research Record: Journal of the Transportation Research Board, 1660(-1), 58-65.

General Symbol List

AADP	The Adjusted Attribute Differentiating Power considering the correlation between the attributes in *FV*
ADP_a	Attribute differentiating power *(ADP)* statistic for attribute *a*
α	The parameter to determine the sensitivity to correlation within *AADP*
β^{m_1,m_2}	The weighting associated with the comparison between mode m_1 and m_2 in computing the weighted average of the *ADP* statistic
CL	The type of classifier *e.g.*, Naïve Bayes, k-NN, etc.
D	The cumulative distance for the trip
Disc	The binary variable associated with feature discretization process
$DP_a^{m1,m2}$	Differentiating power of attribute *a* between mode m_1 and mode m_2
EOT	End of trip
FV^m	Feature vector for mode *m*
IPTT	Initial potential transit trips
MCR	The classifier's misclassification rate
MD_k	Mahalanobis distance between a training case and testing case
MTP	Mode Transfer Points - the points where users change from one to another transportation mode.
NF	The number of features in the classification model
NSS	Non-stationary Sub-segment - a series of consecutive points with speeds higher than a certain threshold (v_{th})

$NStop_{TS}$	The number of stops in proximity to transit stop and not in proximity to signalized intersections
PCA	Principal Component Analysis
PMTP	Potential Mode Transfer Points - the points where the model predicts that a change from one to another transportation mode has been occurred.
PTEP	The Potential Transit End Point
PTMS	Potential Transportation mode segments - portions of a trip that contain sequential data points classified as having been completed by the same mode
PTSP	The Potential Transit Starting Point
ρ	The average correlation between each candidate feature and the features in the chosen set
SD_{th}	The proximity threshold from the signalized intersection
SOT	Start of trip
SS	Stationary Sub-segment - a series of consecutive points with speeds below a certain threshold (v_{th})
$t_a^{m1,m2}$	*t* statistics for the means of attribute *a* for modes m_1 and m_2
TD	The Euclidian distance between the transit station and the centroid of the positions of the stationary sub-segment points
TD_{th}	The proximity threshold from a transit station
TMS	Transportation mode segments - portions of a trip that contain sequential data points labeled as having been completed by the same mode
TSR	Transit Stop Rate - the ratio of the number of stops occurring at transit stations but away from traffic signals to cumulative distance.

TSR_{th} Transit Stop Rate threshold

v_{th} The speed threshold (0.75 m/sec) for stationary points

VM_{th} Maximum speed threshold - The 90^{th} percentile of the non-motorized maximum speed