

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Human Activity Recognition: A Data-driven Approach

Permalink

<https://escholarship.org/uc/item/4w98w1zd>

Author

Luong, Thi Bich Thuy

Publication Date

2015

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
IRVINE

Human Activity Recognition: A Data-driven Approach

DISSERTATION

submitted in particular satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Transportation Science

by

Thi Bich Thuy Luong

Dissertation Committee:
Professor Douglas Houston, Co-Chair
Professor Will Recker, Co-Chair
Professor Michael McNally

2015

DEDICATION

To

My husband
for your love and encouragement to keep me stay active
but focused on my research.

My parents and parents-in-law
for their support and patience throughout many years in school
to finish this dissertation.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	viii
ACKNOWLEDGEMENTS	x
CURRICULUM VITAE	xi
ABSTRACT OF THE DISSERTATION	xii
CHAPTER 1: INTRODUCTION	1
1.1 Background and motivation for research	1
1.1.1 <i>Changes in transportation system</i>	1
1.1.2 <i>Changes in travel demand modeling</i>	2
1.1.3 <i>Routine behavior</i>	3
1.1.4 <i>Data requirements</i>	4
1.2 Research objectives	7
1.3 Outline of dissertation	9
CHAPTER 2: DATA DESCRIPTION	10
2.1 The Expo Line Study	10
2.2 The Boyle Heights study	13
2.3 Existing techniques techniques for post-processing GPS data	15
2.4 Terminology	19
CHAPTER 3: TRAVEL MODE DETECTION	21
3.1 Introduction	21
3.2 Literature Review	21
3.2.1 <i>Segmenting GPS trajectories</i>	21
3.2.2 <i>Multiple changepoints detection</i>	24
3.2.3 <i>Travel mode detection using GPS data</i>	25
3.2.4 <i>Summary</i>	29
3.3 Methodology	30
3.3.1 <i>Study framework</i>	30
3.3.2 <i>Data Cleaning</i>	32
3.3.3 <i>Segmentation algorithm</i>	37
3.3.4 <i>Classification and prediction algorithm</i>	39
3.4 Framework of experiments	44
3.4.1 <i>Segmentation results</i>	45
3.4.2 <i>List of predictors</i>	46
3.4.3 <i>Classification Results</i>	47
3.4.4 <i>Improved mode classification model</i>	51
3.4.5 <i>Model validation</i>	54
3.5 Conclusion	58
CHAPTER 4: TRIP/ACTIVITY DETECTION AND CLASSIFICATION	62

4.1	Introduction	62
4.2	Data Preparation	65
4.2.1	<i>Data selection</i>	65
4.2.2	<i>Data imputation</i>	66
4.3	Discussion of results	67
CHAPTER 5: INFERRING TRIP PURPOSE		72
5.1	Introduction	72
5.2	Literature review	73
5.3	Methodology	79
5.3.1	<i>Naïve Bayes model</i>	81
5.3.2	<i>Multinomial Logistic Regression</i>	84
5.3.3	<i>List of predictors:</i>	86
5.4	Home location identification rules	95
5.5	Classification models	96
5.5.1	<i>Results from Random Forest model</i>	96
5.5.2	<i>Results from Naïve Bayes model</i>	102
5.5.3	<i>Results from Multinomial Logistic model</i>	106
5.5.4	<i>Predictive performance comparison</i>	110
5.6	Conclusion	114
CHAPTER 6: COMPARISON OF ACTIVITY PATTERNS OVER TIME		116
6.1	Introduction	116
6.2	Background	116
6.2.1	<i>Activity-based modeling</i>	116
6.2.2	<i>Time use research</i>	118
6.2.3	<i>Psychological impacts</i>	120
6.3	Methodology	124
6.4	Data	126
6.5	Results	129
6.5.1	<i>MANOVA results</i>	129
6.5.2	<i>Multilevel model results</i>	131
6.6	Conclusion	133
CHAPTER 7: PUBLIC OPINIONS AND RAIL TRANSIT SERVICE IN LOS ANGELES; AN ANALYSIS USING TWITTER DATA		135
7.1	Introduction	135
7.2	Literature review	137
7.3	Methodology	140
7.3.1	<i>Data collection</i>	140
7.3.2	<i>Data processing</i>	143
7.3.3	<i>Theme identification</i>	145
7.3.4	<i>Sentiment estimation</i>	147
7.3.5	<i>Network analysis: "who retweet whom"</i>	148
7.3.6	<i>Topic modeling</i>	149
7.4	Results	151
7.4.1	<i>Flow of information through networks</i>	151
7.4.2	<i>Sentiment analysis</i>	156
7.4.3	<i>Topic modeling</i>	159

7.5 Discussion and conclusion	165
CHAPTER 8: CONCLUSION	167
REFERENCES	171

LIST OF TABLES

Table 2-1: Summary of selected demographic variables	15
Table 3-1: Classification matrix for binary classification problem.....	43
Table 3-2: Sample of data after segmentation.....	45
Table 3-3: Speed pattern of each mode.....	48
Table 3-4: Classify moving points model – baseline model – confusion matrix – training set	49
Table 3-5: Classify travel mode model – baseline model – confusion matrix – training set	49
Table 3-6: Classify moving points model with transit distance – confusion matrix – training set.....	51
Table 3-7: Classify travel mode model with transit distance – confusion matrix – training set	51
Table 3-8: Final result– confusion matrix – testing set	55
Table 3-9: Decision tree model - confusion matrix - testing set	56
Table 4-1: Periods of signal loss (N = 1,125 days, 119 participants)	66
Table 4-2: Aggregate GPS trip counts by dwell time and mode (N= 892 days, 116 participants).....	68
Table 4-3: Mean daily discrepancies between GPS-derived measures and self-reported trip counts.....	70
Table 5-1: Trip purpose definition.....	90
Table 5-2: Land use type vs purpose - BH data	91
Table 5-3: Distribution of purposes - BH data.....	94
Table 5-4: Trip purpose model - Confusion matrix – RF – training set	97
Table 5-5: Trip purpose model - Confusion matrix - RF - testing set.....	98
Table 5-6: Trip purpose model with selected variables - Confusion matrix – RF – Testing set	101
Table 5-7: Trip purpose model without demographic variables - Confusion matrix – RF – Testing set.....	102
Table 5-8: 10-fold cross validation for NB	103
Table 5-9: Prior probability	104
Table 5-10: Conditional probability	105
Table 5-11: Trip purpose model - Confusion matrix – NB – testing set.....	106
Table 5-12: 10-fold cross validation for MNL.....	107
Table 5-13: Coefficient for MNL - trip purpose	108
Table 5-14: Trip purpose model - Confusion matrix – MNL – testing set.....	110
Table 5-15: Error rate comparison	111
Table 6-1: Comparison of Trip-based and Activity-based models.....	117
Table 6-2: Data description.....	127
Table 6-3: Land use type vs purpose – Expo data	128
Table 6-4: MANOVA results	129
Table 6-5: Sample mean for duration and number of visits	130

Table 6-6: Multilevel model's components	132
Table 7-1: Theme of tweets	146
Table 7-2: Distribution of Twitter accounts	147
Table 7-3: Regression results	153
Table 7-4: Topic model result for the Blue line	164

LIST OF FIGURES

Figure 2-1: The Expo study area	12
Figure 2-2: The Boyle Heights study area	14
Figure 2-3: Previous “manual” GPS classification framework.....	17
Figure 2-4: Proposed data processing framework	18
Figure 3-1 Travel Mode Detection Framework	31
Figure 3-2: Kalman Filter scheme	33
Figure 3-3: Data cleaning and smoothing result	37
Figure 3-3: Experiment on Expo data.....	44
Figure 3-5: Changepoints in speed, penalty value = $\log(n)$	46
Figure 3-6: Important variables in status model.....	50
Figure 3-7: Speed pattern of each mode (first 1000 points).....	53
Figure 3-8: Important variables in mode model	54
Figure 3-9: Sensitivity	57
Figure 3-10: Sample movement patterns	58
Figure 4-1: Expo Study trip log format.....	63
Figure 4-2: Trips detection process.....	64
Figure 4-3: Compare dwell time results.....	68
Figure 4-4: Compare number of GPS-derived trips with self-report trips	71
Figure 5-1: Process to predict trip purpose	80
Figure 5-2: Generative method v.s discriminative method.....	81
Figure 5-3: Example of Kenel density estimation	87
Figure 5-4: Box plot for trip purposes - BH data	92
Figure 5-5: Association among arrive time, daily frequency, age, and purpose - BH data.....	93
Figure 5-6: Distribution of activity duration by travel mode and purpose - BH data.....	94
Figure 5-7: Distribution of trip purpose - BH data	95
Figure 5-8: Error rate for trip purpose model	98
Figure 5-9: Trip purpose model - Variable importance ranking	99
Figure 5-10: Trip purpose model - Error rate when adding variables	100
Figure 5-11: Box plot of MNL's fitted value	109
Figure 5-12: Sensitivity of classifying trip purposes.....	112
Figure 6-1: Theory of Reasoned Action and Theory of Planned Behavior	121
Figure 6-2: Framework of Theory of the Interpersonal Behavior	122
Figure 6-5: Imputing trip purpose process for EXPO data	126
Figure 6-6: Activity duration: travel mode vs activity type.....	131
Figure 7-1: LA light rail network.....	141
Figure 7-2: Geographical distribution of tweets	142
Figure 7-3: Scheme of Twitter data analysis	143
Figure 7-4: Word count distribution	146
Figure 7-5: Retweets distribution, by rail line category and days of week status.....	152
Figure 7-6: Who retweet whom	154
Figure 7-7: Who retweet whom by Theme.....	155

Figure 7-8: Average sentiment score by rail transit line	157
Figure 7-9: Sentiment by time of day	158
Figure 7-10: Sentiment by day of week	159
Figure 7-11: Word clouds.....	161
Figure 7-12: Network analysis of term for the Blue Line	162

ACKNOWLEDGEMENTS

I would like to take this very special opportunity to acknowledge the importance of my family in the completion of this dissertation. In particular, I am grateful to my husband, my soul mate, Ming. This dissertation would not have been achieved without the unlimited support from you. You are the one who has been taking care of me 24/7, sharing with me every sadness and happiness in my life. Thank you for always being there, giving me your strong and warm shoulder, and encouraging me during my PhD journey.

I am also grateful for the support I received from my beloved families in Vietnam and Taiwan, my parents, and parents in law, my brother, and brother-in-law for all their love and their boundless faith in my strength and spirit. You are the ones who let me choose my own way in life and always by my side in all my pursuits despite the distance.

I would like to express my deepest appreciation to my advisor, Dr. Will Recker, who continually and convincingly conveyed a spirit of adventure in regard to research. He inspired me to start my research by just one sentence “Let’s do it”, gave me time to try and explore the ocean of knowledge. Your intelligence and way of thinking made a very strong impression on me.

I also do appreciate my co-advisor, Dr. Douglas Houston, for your trust in my capability and passion, giving me an opportunity to work on amazing projects that extend my knowledge. Doug, I also want to thank you not only as a professor, but also as a friend. Without your guidance and persistent help, this dissertation would not have been possible.

I appreciate all of my supportive and encouraging friends in Institute of Transportation Studies, Computer Science department, as well as in Sociology and Social Ecology. Their warmth and friendship have made UC Irvine a wonderful place to be!

The Expo Line Study was supported by the California Air Resources Board, the Haynes Foundation, the Lincoln Institute of Land Policy, the San Jose State Mineta Transportation Institute, the Southern California Association of Governments, the University of California Transportation Center, the University of California Multi-Campus Research Program on Sustainable Transportation, and the University of Southern California Lusk Center for Real Estate. The Boyle Heights Activity and Exposure Study was supported by the University of California Transportation Center. I am grateful to the GPS study participants and those who assisted with data collection and processing.

CURRICULUM VITAE

Thi Bich Thuy Luong

- 2015 Ph.D. in Transportation Science, University of California, Irvine
Dissertation: Human activity recognition: a data-driven approach
- 2008 M.B.A., Southern Taiwan University of Technology, Taiwan
Thesis: Vietnamese consumers' brand perception of luxury scooters:
comparison between Italian brands and Japanese brands
- 2006 B.A. in Business Administration, Vietnam National University, Hanoi

FIELD OF STUDY

Activity-based Model, Travel Behavior, Travel Survey, Social Media, Consumer Behavior

PUBLICATIONS

Luong T.B.T., Houston, D. (2015). Public opinions of light rail service in Los Angeles, an analysis using Twitter data. In Proc. iConference 2015.

Houston, D., **Luong, T.B.T.**, & Boarnet, M. G. (2014). Tracking daily travel; Assessing discrepancies between GPS-derived and self-reported travel patterns. Transportation Research Part C, 48, 97–108.

Yang, M., **Luong T.B.T.**, and W.W. Recker (2013). *Extracting traffic patterns from loop detector data using multiple changepoints detection*. Transportation Research Board, 93rd Annual Meeting 2014.

PRESENTATIONS

Luong T.B.T. and W.W. Recker (2015). *Analyzing how socio-demographic characteristic associate public transit usage - a comparison between southern and northern California*. Transportation Research Board, 94th Annual Meeting 2015.

Luong T.B.T. and W.W. Recker (2014). *Analyzing the differences in driving behavior in a dilemma zone while being distracted*. The 20th Annual UC Transportation Research Conference, Pomona.

Yang M., Regué R., **Luong T.B.T.**, and Recker W., 2013. *Visualizing and Understanding Lane Changing Maneuvers in the Vicinity of Freeway-to-Freeway Connectors*, 7th

International Symposium on Visualization in Transportation

Luong T.B.T. (2013). *Analyzing Children's Activity Patterns in Southern California by Using Random Forest*. The 19th Annual UC Transportation Research Conference, Los Angeles.

RESEARCH EXPERIENCE

2010-2014 Graduate Student Researcher, University of California, Irvine

Human activity recognition and analysis (The Expo Light Rail Line Study)

- Used a change detection algorithm and Random Forest to classify transportation modes (car, bus, train, bike, walk) from GPS data
- Applied multiple machine learning techniques to predict travel trip purposes from GPS data
- Analyzed changes in travel patterns before and after operating a new light rail line
- Developed R programs to monitor collecting and processing GPS data project

Freeway traffic patterns analysis

- Applied a change detection algorithm to capture freeway congestion patterns from sensor data
- Identified hotspots based on the congestion periods
- Analyzed lane-changing behavior from freeway-to-freeway

Mining Twitter data about metro rail service in Los Angeles

- Applied data-mining techniques to analyze Twitter data about commuters' perceptions, attitudes, and behavior.
- Built an interactive web-application using R-shiny

ABSTRACT OF THE DISSERTATION

Human Activity Recognition: A Data-driven Approach

By

Thi Bich Thuy Luong

Doctor of Philosophy in Transportation Science

University of California, Irvine, 2015

Professor Douglas Houston, Co-Chair

Professor Wilfred W. Recker, Co-Chair

In this research, we propose a series of models designed to take advantage of the availability of data—both structured and unstructured—from a variety of sources ranging including passive data, questionnaires, and social media data to analyze underlying patterns and trends in travel and activity behavior, and to provide results that support enhancements both in transportation planning and the application of programming to support such efforts.

First, we introduce a framework for automatically inferring the travel modes and trip purposes of human movement when tracked by a GPS device. We utilize a multiple changepoints algorithm to divide trajectories into segments using only speed data. Then, Random Forest is used to classify segments into moving and not-moving types. For moving segments, travel mode (car, bus, train, walk, and bike) is predicted. Next, multiple machine learning algorithms are employed, validated, and tested to identify the most suitable model for inferring trip purposes. The overall accuracy for prediction

is over 80% on the testing set, both with and without data on socio-demographic variables. The model also predicts “shop” trips with an accuracy of 92.1%, while its accuracy for “go home” and “studying” trips reaches 100%.

Additionally, we utilize the classification results in the first stage of research to compare households’ travel patterns from before a new light rail transit line began service to two periods of time after service began. Our results indicate that, although the average of activity duration varies significantly over days of week and waves, the random effect of these two factors on activity duration was minor; time of day contributed over one third of the total variance in the duration.

Finally, the dissertation demonstrates uses of Twitter data as a potentially important data source to understand comments, criticisms, and responses about light rail in Los Angeles. The results of information flow analysis, sentiment estimation, topic modeling, and its application can be useful for exploring trends among commuters and how their sentiment changed according to the light rail line they used, the time of day, and the day of the week.

CHAPTER 1: INTRODUCTION

1.1 Background and motivation for research

1.1.1 Changes in transportation system

The central purpose of transportation planning and management is to match transportation supply, in the form of transportation infrastructure and its associated operation, with travel demand. A thorough understanding of existing travel patterns is necessary for identifying and analyzing traffic-related problems. Hence, modeling and predicting travel behavior has been an important field of transportation research. The importance of demand-related strategies was significantly increased after the Clean Air Act Amendments of 1990, the Inter-modal Surface Transportation Efficiency Act of 1991, and the California Global Warming Solutions Act of 2006 were adopted. There have been changes in transportation planning in order to make urban transportation systems more sustainable. Government and non-government agencies have searched for strategies and plans to reduce private automobile use, including policies to support a shift from personal vehicle driving to public transit. Public transit includes such services as taxi, carpools, buses, rails, ferries and their variations. Efficient transit systems can stimulate transit-oriented development, multi-modal neighborhoods where residents tend to own fewer vehicles, drive less, use more travel alternative modes, and have more active living.

Significant light rail transit developments have occurred in Los Angeles during the last decade. The city will open six new rail transit lines before 2020 in order to move the city further into its rail development with optimistic claims that the extended urban rail systems would

increase transit ridership, attract people out of their car, ease congestion, and improve environment (Los Angeles County Metropolitan Transportation Authority, 2009). An evaluation program is crucial for policy makers, planners, and public to understand how efficient and effective these plans are. However, different evaluation methods can lead to different conclusions concerning the value of a transit project. A method based solely on economic evaluation could overlook the benefit to travelers if it focuses limitedly on operating and maintaining cost. Alternatively, an evaluation method that incorporates travel demand provides how much travelers' behavior change by comparing the travel demand before and after the service is opened, especially how the new system affects people who do not currently use transit services.

1.1.2 Changes in travel demand modeling

Following criticism, mainly from academics, there has been a shift from the traditional four-step transportation demand model to an activity-based model structure as an improvement in methodology to predict travel demand. These models aim at simulating individuals' decision-making behavior across many dimensions simultaneously, such as which activities are conducted where, when, for how long, with whom, and the transport mode involved. Most activity-based models can be divided into two parts: an activity generator and activity scheduler (Nurul Habib & Miller, 2006). Despite significant advances, a number of fundamental issues still need to be addressed within activity-based models. Included among these are improvements to the activity generation process. Most of the current models are not fundamentally applicable on a dynamic scale due to the lack of a dynamic activity generation scheme. For example, current models assume that activity-travel decisions are uniform from

one day to the next. Modeling dynamic activity generation remains a challenge to improve prediction capabilities of activity-based models.

1.1.3 Routine behavior

There has been a trend in the literature toward researching day-to-day, week-to-week variability in activity-travel. Previous studies (Arentze & Timmermans, 2009; Hirsh, Prashkea, & Ben-Akiva, 1986; Pas, 1986) suggested that disregarding a weekly effect on daily activity patterns may lead to biased predictions. In the other words, we should pay attention not only to the variability in activity-travel behavior across individuals/households on the same days (interpersonal – variability), but also to the variability in travel behavior for a given individual/household over a span of time (intrapersonal – variability). The most obvious measurement for intrapersonal-variability of activity is activity frequency and sequence. Ignoring the existence of these factors would give an inappropriate travel behavior description, which may lead to a misguided strategy for policy of transportation planning and infrastructure management (Tarigan, Fujii, & Kitamura, 2012).

As highlighted by leading travel-behavior researchers, it has become increasingly evident that psychology is an important predictor of travel patterns. Explanatory studies show that mode and route choice can be better explained when psychological factors are taken into account. Recently, attention has been given to attitudes as an important determining factor in decisions (Otto, 2010). Kitamura et al. (Kitamura, Mokhtarian, & Laidet, 1997) showed that there are three determinant variable categories that explain travel behavior: socio-economic, land-use, and attitudinal variables. They found that the highest proportion of the variation in the data could be explained by attitudinal variables. Attitudes may result from reactions to such

aspects as travel speed, comfort and flexibility. However, the influence of habits and routines on decisions has to be considered to a much greater extent because research shows that attitudes alone cannot predict behavior, especially routine behavior such as daily commuting (Otto, 2010). Habits and routines are deeply rooted into individual's behavior as the rhythm of daily life.

There is a large body of literature in psychology applying the Theory of Planned Behavior (Fishbein & Ajzen, 1975) and the Theory of Interpersonal Behaviour (Triandis, 1977), which use attitudes and habit to explain behavior. Attitudes generate intention as behavior's immediate psychological antecedent. Intention derives from a person's overall evaluation of performing a behavior (attitude toward it), perceived social pressure surrounding it (subjective norm), and perceived control over factors that may facilitate or inhibit performance (perceived behavioral control). However, the application of these theories in transportation has been just limited to proving a significant effect of psychology to choice decision. Therefore, there remains room for bringing the theories into a real model to make prediction of travel behavior, especially in the case of upgrading the transportation system such as extending public transit network.

1.1.4 Data requirements

1.1.4.1 GPS data

In order to observe the routine of daily-life activities, detailed, and multi-day travel data are needed. Traditionally, travel survey data are collected by asking participants to describe their travel times, trip origins and destinations, travel modes and other different travel details during the survey period. There are a number of methods in conducting travel surveys,

including face-to-face, telephone, mail, Internet-based surveys, or some combination of these. However, these methods rely on the accuracy of self-reported information. It has been shown that data collected using these methods deviate systematically from actual behavior (Stopher, 2004). Participants may neglect to report trips, especially short trips with increasing duration of the survey (Golob & Meurs, 1986) and non-home-based trips (Stopher, 1992) because of the burden of taking notes for extended periods. Participants may also complete the survey at the end of each day or survey period, which heightens the probability of forgetting partial or whole trips. They might also fail to report their detail activities, or trips they considered unimportant to report or for which they have privacy concerns (Wolf, Loechl, Thompson, & Arce, 2001b). Another reason potential problem is that participants could misunderstand travel survey instructions.

To overcome these problems, over the past decade, the number of studies using Global Position System (GPS) for collecting individual humans' daily activities has increased. GPS is a satellite-based system that can be used to locate positions anywhere on the earth, which is operated by the U.S. Department of Defense, Navigation Satellite Timing and Ranging. GPS provides continuous (24 hours/day), real-time, 3-dimensional positioning, navigation and timing data worldwide. Recent developments in the technology include more accurate and better recording of positional accuracy, especially in areas of high densities of buildings and/or trees, within urban canyons along roadways with tall buildings, and in trains or buses where satellite reception may be poor. Although GPS devices can collect the latitude, longitude, and the time of each location point, they do not directly classify in the data when a participant moves or stops, mode of transportation, and trip purpose. Supplemental post-processing is needed to impute this information from raw GPS data and/or other data sets. In most of existing studies,

these tasks have been done manually or/and with some level of automation (Wu, Jiang, Houston, Baker, & Delfino, 2011). Even so, researchers typically still need to validate, correct, or add missing trips through a supplemental Internet recall survey, mailing, or telephone, which could be burdensome for the participants and the result relies on the participants' cooperation and memory. The field still needs more improvement in automated classification methods in order to perform the tasks faster, more efficiently, and more accurately.

1.1.4.2 Social media data

Smartphone and social media have changed the world and the way in which we communicate with each other and spread information. The potential transformation of how we view and analyze happenings in a city is being influenced by the availability of an incredible amount of open source data. Social media platforms can provide public feedback and opinion in a real-time manner that is tied to geographical information when using devices with GPS functionality. An increasing number of commuters are using social media such as Twitter to communicate or report accidents, road congestion, and special events. Many government organizations also use Twitter to communicate with the public and commuters. The contents of popular social media contain abundant information on transportation-related issues, which could potentially be a good source of data for transportation planners and researchers to better understand commuters and public opinion.

Using this kind of data, cities and planning organizations can identify potential scenarios for particular planning and management processes, for example detecting events and emergency reaction. Although use of Twitter data for opinion mining is popular in many fields, its use in transportation operations sector remains limited. Therefore, there is opportunity to

better understand how to collect and analyze these data to improve the travelers' experience, both for individual commuters and transportation agencies.

1.2 Research objectives

The objectives of this research are to expand the existing literature on the use of GPS technology in household travel surveys and social media data to understand public opinion about transportation services.

First, this research develops a system that creates probabilistic models of users' activity patterns drawn from GPS data using data segmentation, machine learning, and other statistical techniques. The tasks in building the models include the following: Cleaning and smoothing data, detection of change point, travel mode identification, trip/activity detection, and trip purpose imputation. The models accomplish the following:

- Separates GPS data tracks into location and movement periods
- Infers travel mode (walking, biking, to travel by car, bus, or train)
- Estimates the location types a person visits in a day (home, work, shopping, etc)
- Identifies activities a person participates over a week
- Classifies personal activity data without required for manual processing

The models are then applied to classify GPS data collected as part of the Expo Line evaluation study of the impact of a new light rail transit service in Los Angeles, California, which demonstrated the usability of the GPS data for analyzing the impacts of infrastructure improvements. The research team collected data in a largely low-income and non-white community in multiple waves in south Los Angeles, California from 2011 to 2013. The current

research validated the automated classification model results by comparing these classifications to those derived in the Expo Study based on a manual review of GPS traces using GIS and supplemental data about the roadway and built environment.

Second, using the GPS data processed with the classification models developed, this dissertation also conducted an in-depth exploration of routine/non routine activities focusing on habitual behavior, in order to estimate the impact of change in the transportation system to routine travel behavior. A wide range of factors including socio-economics, land use, and transportation network variables will be investigated in order to understand how people change routine travel/activity. The results will help policy makers and planners implement sustainable transportation strategies that have the greatest impact both on land use and the transportation system.

Third, this research applied text-mining techniques to process Twitter social media data to provide differentiated insights based on the comments, criticisms, responses, and shared information about rail transit service in Los Angeles by transit users and transit providers. The analysis includes the following steps: Flow analysis, sentiment analysis, and topic modeling. The resulting application provides a practical and applicable tool for measuring commuters' emotions and opinions that can be used to understand trends among commuters and how their emotions changed according to the light rail line they used, the time of day, and the day of the week. The methods developed in this study can also be applied to analyze other areas of service operations, to help government agencies manage information and facilitate interaction in meaningful ways, and to achieve broader public participation than is possible through normal

channels. Analysis of social media data can provide access to segments of users that have not participated in traditional travel and activity surveys.

1.3 Outline of dissertation

The organization of this dissertation is as follows. Chapter 2 describes the data set, how it was collected and pre-processed. There are two data sets used, which were collected for different studies, at different time periods, but both of which share similar characteristics in socio-demographics. To address the limitations of existing data processing frameworks, a new scheme is proposed. Each part of this framework will be detailed in the following chapters.

Chapter 3 introduces the travel mode detection methodology and results. Chapter 4 introduces methods for identifying activity periods and trip segments in GPS data. Chapter 5 presents methods for trip purpose imputation. Based on data generated in these previous chapters, Chapter 6 analyzes how travel behavior changes in relation to the introduction of a new rail transit service. Chapter 7 demonstrates methods to process and classify Twitter data as a potentially important data source to understand comments, criticisms, and responses about rail transit in Los Angeles. The final chapter summarizes the results and explains the contributions, conclusion, and future work of the dissertation.

CHAPTER 2: DATA DESCRIPTION

In this chapter, the data sets used in this dissertation are described, including how they were collected and the processing framework. Basing on the limitations of existing methodologies, a new research framework is proposed and implemented in this dissertation.

2.1 The Expo Line Study

Significant rail transit development has occurred in Southern California over the last decade; for example, in south Los Angeles, the 13.2-km (8.2-mi) Expo Phase 1 line opened in April 2012. Work is advancing on the 10.8- km (6.7-mi) Phase 2 segment to complete the route to Santa Monica by 2015 (Los Angeles County Metropolitan Transportation Authority, 2009). The first data set used in this dissertation, which is used for calibrating the mode/trip/activity detection models and analyzing travel patterns, is was collected as part of the Expo Line Study, an evaluation of Phase 1 of this line. The Expo Line study area includes about 12 square miles (31 square kilometers) along the Exposition and Crenshaw corridors (Fig. 1). This is a longitudinal dataset included GPS data on the personal behavior of participants during during three time intervals: September 2011-February 2012 (before the line started service in April 2012), September 2012-February 2013 (after Expo service began), and September 2013 – April 2014. This multi-wave data from 116 households was designed to capture neighborhood-level travel change associated with the opening of the Expo Line light rail line in South Los Angeles in order to identify changes in intra-personal travel behavior associated with the new service. Participants were asked to complete a survey that included socio-economic, attitudinal, and

crime-related questions. Overall, residents of the study area were about 9% Non-Hispanic White, 41% Hispanic, and 43% African-American; about a fifth of residents lived in households with income below the federal poverty level, about a third were foreign-born, and about a quarter had an educational attainment of a Bachelor's degree or higher (US Census Bureau. 2005-2009 American Community Survey; US Census Bureau. 2010 Decennial Census Summary File 1).

Since the overall Expo Study used an experimental-control study design, the study area was divided into an “experimental” area within 0.5 miles of an Expo station corresponding to the area within which we hypothesize participants would likely walk to access the Expo service, and control areas farther than 0.5 miles of an Expo station with similar demographic, land use, and transit service patterns which we hypothesize are far enough from an Expo station to be impacted by the “treatment” (the new light rail service).

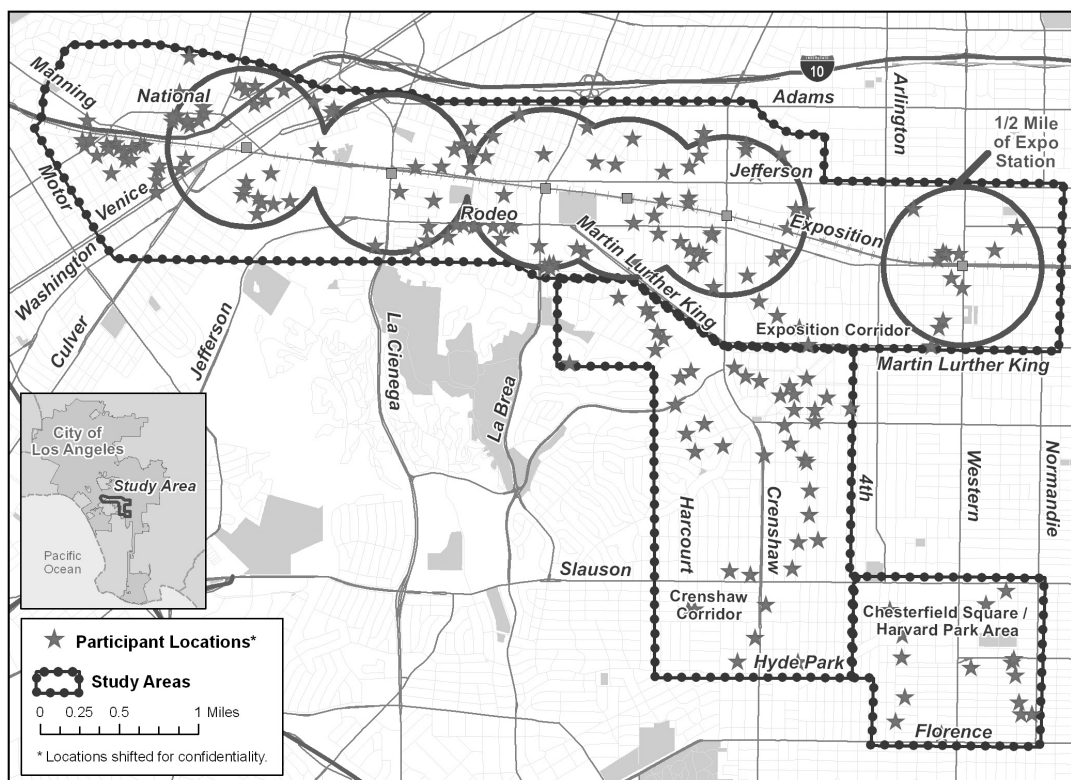


Figure 2-1: The Expo study area

To identify potential participants in the first phase of data collection, a mailing list of all household addresses within the study area (27,275) was purchased from InfoUSA, a marketing information firm. Each household was mailed an invitation letter and all households who indicated they were interested in participating (651) were invited into the study. Participants in the mobile tracking group received a grocery gift card worth \$30 as an incentive for completing phase 1 of the study, and this incentive was increased in value to \$75 to encourage the same households to participate again in phase 2. Participants met with a project staff member before his or her data collection period for training on the use of equipment and after his or her data collection period to retrieve materials and to ensure compliance to study procedures. Participant locations were recorded at a 15-second interval for the seven-day data collection

period using a portable QSTAR GPS logger device (Model QT-1000x) which has a high level of spatial precision and sufficient battery life for 24-hr daily participant monitoring between charges (Duncan et al., 2013). The data from this study hereafter are referred to as “Expo” data.

2.2 The Boyle Heights study

The second data set, which is used for trip purpose imputation modeling, is from the Boyle Heights Activity and Exposure Study (BH study) which utilized GPS-based location tracking combined with portable air pollution monitoring to provide a more complete, diurnal characterization of air pollution exposures for residents of Boyle Heights, a neighborhood near downtown Los Angeles (Houston, Wu, Yang, & Jaimes, 2013). Participant activity logs, GPS tracking, and follow-up surveys of 32 residents over 247 days were collected in this study. The GPS devices—the same type and model as devices used in the Expo study—tracked their locations and activities for 5-7 days during February to May, 2011 at a 15-second interval. The data from this study hereafter are referred to as BH data.

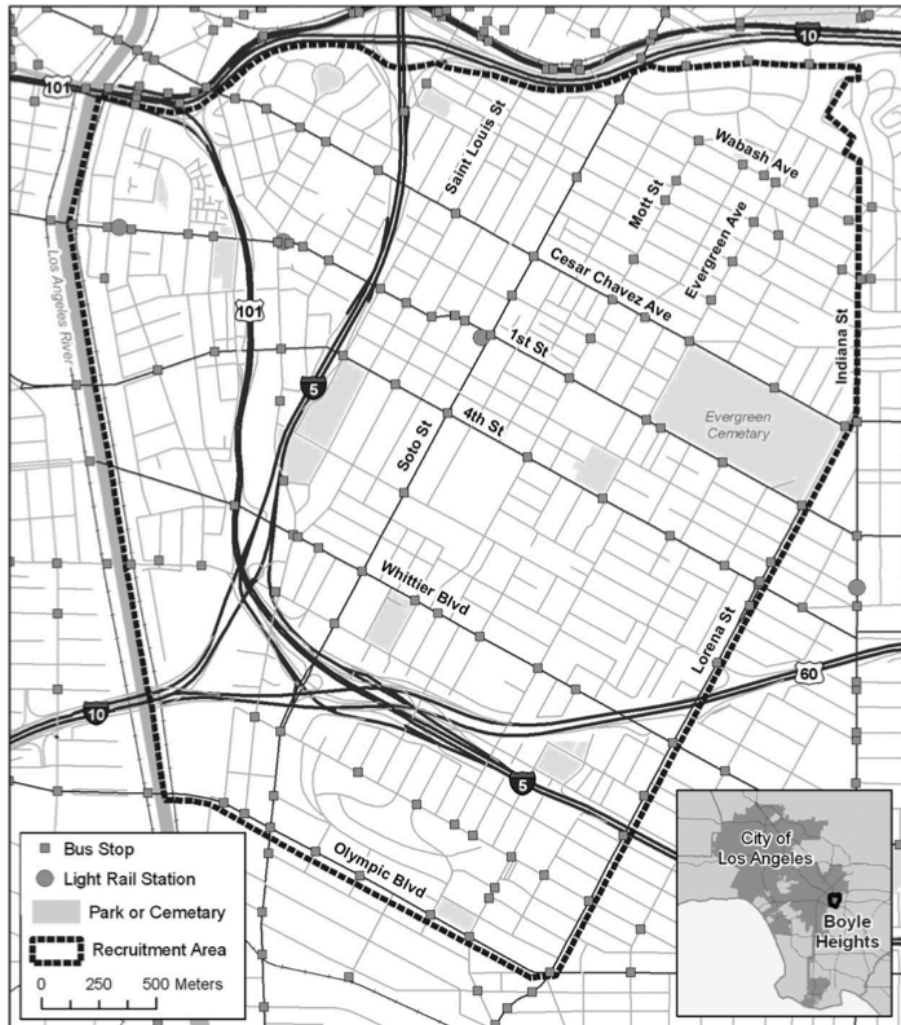


Figure 2-2: The Boyle Heights study area

Source: (Houston et al., 2013)

In an initial baseline survey, BH study participants were asked for information of about five to nine locations they visit frequently and activity type (*“What is typically your primary purpose or activity at this location?”*) they often conduct at each location. After processing all GPS data (using same “manual” classification process as in Expo study), the researchers identified all locations for which the activity data was not provided in database. Participant activities at locations then were clarified in a follow-up survey. During follow-up interviews,

participants were also asked to clarify discrepancies or unclear patterns in their location-exposure database. Researchers gathered additional information about activity locations and travel patterns using a visual prompt.

These data were used to create a model predicting trip purpose. This model will be subsequently applied to Expo study GPS data to predict purpose for all trips. Table 2-1 shows a summary of demographic characteristics of the participants in Expo study and BH study

Table 2-1: Summary of selected demographic variables

	Boyle Heights		Expo wave 1		Expo wave 2	
	Number	%	Number	%	Number	%
Gender						
Female	17	53.12	72	67.92	57	66.28
Male	15	46.88	34	32.08	29	33.72
Age						
20_29	13	40.62	11	10.38	8	9.3
30_39	9	28.12	15	14.15	9	10.47
+ 40	10	31.25	80	75.47	69	80.23
Employment						
Full time	3	9.38	41	38.68	33	38.37
Not employed	13	40.62	47	44.34	31	36.05
Part time	16	50	18	16.98	22	25.58
Number of cars						
0	12	37.5	15	14.15	14	16.28
1	13	40.62	44	41.51	38	44.19
+ 2	7	21.88	47	44.34	34	39.53
Number of HH	32		106		86	

2.3 Existing techniques techniques for post-processing GPS data

This section describes (1) the “manual” classification process, which was previously used to obtain location and travel mode status labels for GPS points in the Expo and BH data sets, and (2) the alternative, automated GPS classification process demonstrated in this dissertation.

The Expo and BH GPS data include location and travel mode status labels of GPS points in the selected Expo Study data set derived using a “manual” classification process. The

procedure used to derive these status labels has been previously described (Houston et al., 2013; Wu et al., 2011), and was used to identify periods during which participants were at a location (home, work, school, grocery, etc.) and periods when they were traveling between locations. This classification process depicted in Figure 2-3 involved point cluster analysis and extensive and careful inspection of all GPS points in sequence using map overlays.

Stationary periods were identified as sequential GPS locations which were clustered for longer than 2 minutes within 20 meters around a given location—a distance which corresponds to the range of positional errors of the GPS device used (Duncan et al., 2013). Periods not at such locations were classified as vehicle travel if the speed between sequential GPS points was 6 mph (10.7 kph). All points with an initial classification and non-classified points were reviewed to correct classifications using ArcGIS software (ESRI, Redlands, CA) to examine the proximity of sequential GPS locations to roadways, sidewalks, transit route/stops, and parcel boundaries and land use designations. Points recorded immediately after a vehicle travel period were also classified as an in-vehicle travel period if they corresponded with a brief in-vehicle stop along a roadway or parking area during a longer travel sequence. Vehicle travel periods were re-classified as public transit travel periods if a trip began at a transit stop, followed a transit route, and ended at a transit stop. Finally, GPS locations not classified in earlier steps as either being at a location or as part of a vehicle or transit travel period were reviewed to confirm if they corresponded with a walking travel period based on whether they occurred in sequence between locations, were at a typical walking pace (2-6 mph), and followed a reasonable path along roadways, sidewalks, or open space areas. GPS data for airplane travel is omitted since the focus of this dissertation is to improve the accuracy of travel

survey for inroad transportation systems. These “manual” GPS classifications are used in the current research to validate the automated models developed in the current research.

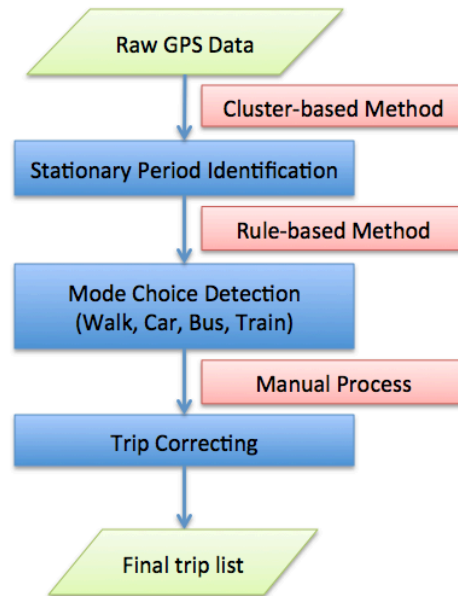


Figure 2-3: Previous “manual” GPS classification framework

This “manual” classification scheme has several limitations. The cluster-based method used set thresholds of 20 meters radius and 2 minutes duration. The selection of these thresholds were based on the literature, on experience, or on the accuracy of collecting information of GPS device, and is therefore arbitrarily set. The protocol also has the disadvantage of Rule-base method when detecting travel mode. Finally, correction of initial classifications requires a manual process, which is costly both in time and in labor.

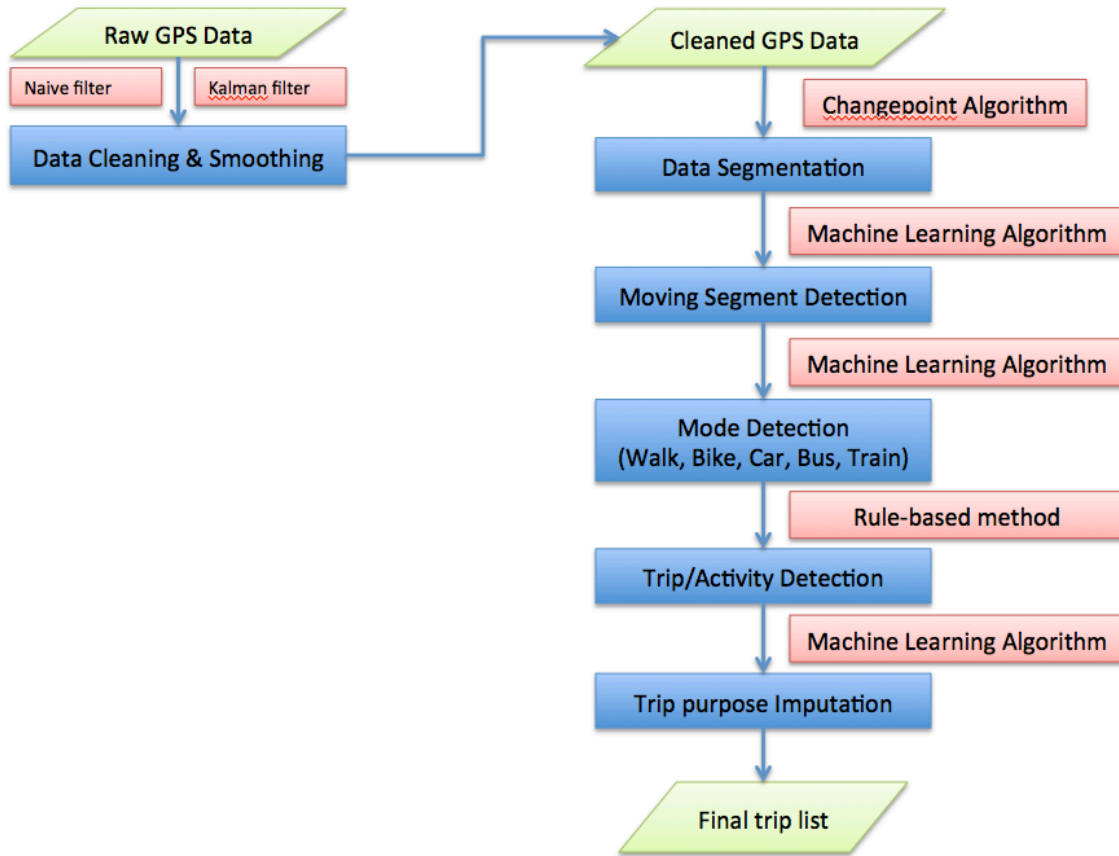


Figure 2-4: Proposed data processing framework

The scheme for data processing proposed in this dissertation (Figure 2-4) addresses these limitations the “manual” classification procedure previously used to post-process the Expo and BH GPS data. This methodology and results applies automated, fast, and efficient techniques to generate travel survey-like trip, mode and activity data based on GPS. The proposed method is designed specifically to identify bike trips, short walk trips, and transit trips, and to reduce the amount of time needed to manually classify GPS data.. This approach is expected to achieve high temporal accuracy and reasonable spatial accuracy.

2.4 Terminology

GPS trajectory (track)

A GPS trajectory contains continuous GPS points forming a participant's path in: 1) a day, 2) less than a day (in case of missing data), or 3) several days. A GPS trajectory includes stationary (stop) points when a participant stays at a location and moving points when the participant travels from a location to another location.

Trip

In general, a trip is defined as the movement of a participant from an origin to a destination location. A location could be either an origin or destination of a trip or both. One GPS track may consist of more than one trip. A trip could be conducted by using several transportation modes. For example, if a person walked to parking lot from an origin, took a car and drove to a parking lot, then walked to a destination, stayed there for a while before going to another location. In this case, the trip is understood as a trip chain.

Tour

A tour is a closed chain of trips starting and ending at a base location.

Segment

Typically, a segment is defined as a part of a trip that is conducted with a specific travel mode. Because this dissertation proposes using a segmentation method for detecting stationary points, here a segment specifically refers to part of GPS trajectory. Several sequential segments create a period that is related to a specific mode.

Trip purpose

The purpose of a trip is the activity type, which is conducted at a destination. For example, if a participant's origin was home and destination was her office, she went there to work, so the purpose of this trip is classified as working.

CHAPTER 3: TRAVEL MODE DETECTION

3.1 Introduction

In this chapter, transportation modes are inferred from raw GPS data, including the task of separating GPS tracks into move/non-move segments and classifying non-move segments into walking, biking, driving, and taking public transit. There have been a number of papers that have proposed methodologies for these tasks. However, most of the studies focused on solving travel mode detection from non-move segments but neglected the prior task of how to distinguish a GPS point from moving or not moving status. There remains room for improvement in the research of dividing a GPS trajectory into segments. Therefore, this chapter introduces a new method, using a multiple changepoints detection algorithm to segment GPS trajectories, and then infer travel mode for each segment.

3.2 Literature Review

3.2.1 Segmenting GPS trajectories

The study of trajectories has been basic to many fields such as traffic, biology, and ecology, etc. Trajectory is considered as a function that represents the changing of the location of an object moving in space during a given time interval (Celli, 2009). Today, with mobile wearable tracking devices, researchers can collect such mobility data as GPS or accelerometer data that were heretofore generally unavailable. Despite the fact that these data can provide temporal and spatial information of every location point or time point, transportation mode and activity cannot be detected directly and the task of such determination is relatively complicated. In most cases, the raw data drawn from the GPS trajectories need to be separated

into segments, which include moving segments—moves—and non-moving segments—stops. Between two consecutive stops is a move. A stop here is defined as a place where the person can conduct an activity, and not simply a physical stop, such as a stop at traffic light. A trajectory contains a sequence of stops. A stop is associated with a set of points of interest (POI).

To segment a trajectory into stops and moves is not new in literature. A number of research studies have focused on methods to divide a trajectory into segments and detect to which category a segment belongs (Palma, Bogorny, Kuijpers, & Alvares, 2008; Roider, Rudloff, & Ray, 2012; Wu et al., 2011), but still there is room for improvement. There are two main approaches that have been developed: intersection-based and clustering-based approaches. The first method generates stops based on the intersection of trajectory sample points with relevant geographic object types, mostly based on a time threshold for a stop duration, such as how long a person stops at a parking lot or a museum (Alvares et al., 2007). The second method is based on the variation of the trajectory in speed, acceleration, etc. In the literature, the second approach has been widely used. To detect indoor, outdoor static, outdoor walking, and in-vehicle travel period, Wu et al. (Wu et al., 2011) proposed a methodology based on the rule that if all of the points within a minimum of one minute had speed lower than 3 km/h, these points were treated as a static cluster. Then, periods of movement within a static clusters were detected by a line detection process (the presence of linear alignment of sequential points) until three continuous points that did not form a line are detected.

Roider (Roider et al., 2012) used a self-organizing network algorithm to reduce the GPS positions in a collected set of routes to a set of spatial nodes together with a topology of edges that represent routes between the calculated nodes. Then a clustering algorithm was applied to

reduce the number of nodes by their mean position. The rest of the points were assigned to the nearest node and transferred into classes by clustering the segments whose corresponding nodes are in each other's topological neighborhood. From here, stops and transfer points were determined. Ashbrook and Starner (Ashbrook & Starner, 2003) used K-means algorithm to find important places in trajectories. However, the disadvantage of this algorithm is that the number of clusters must be given *a priori*.

In a branch of the clustering-based approach, Palma et al. (Palma et al., 2008) proposed a two-step algorithm: the Clustering-Based Stops and Moves of Trajectories (CB-SMoT) to extract stops and moves as an extension of a density-based clustering algorithm (the neighborhood is defined as an area around points in 2-dimensional space), called cluster based-stops and moves of trajectories. The key idea of density-based clustering is that for each object of a cluster, the neighborhood of a given radius has to contain at least a minimum number of objects. First, potential stops (the slower parts of a trajectory) are detected, then the algorithm identifies where these stops are located, considering the geography behind the trajectories.

Since 2008, Microsoft's Geo project has presented another methodology to infer a travel model from the raw GPS data (Zheng, Li, Chen, Xie, & Ma, 2008a). In the first step of the work, in addition to using density-based clustering algorithm, the authors used a change point algorithm: ordering points to identify the clustering structure (OPTICS) (Ankerst, Breunig, Kriegel, & Sander, 1999). Change point as a terminology stands for a place where people change transportation mode with an assumption that the start point and end point of a walk segment could be a change point with very high probability. The authors distinguished all possible walk points and non-walk points by grouping multiple users' change points into nodes, and built edges between different clusters based on user-generated GPS trajectories.

Thresholds for velocity and that of acceleration for each data point were used for segmentation. The study opens an opportunity for segmenting GPS trajectories research based on change points that appear naturally in the data themselves instead of depending on *a priori* information about likely destinations.

3.2.2 Multiple changepoints detection

Substantial effort has been devoted to efficiently identifying multiple changepoint locations, particularly as applied to time series data, a class of data under which GPS data falls. In the literature, two major issues have been addressed along this topical area: determining which parameters (either simply the mean or both the mean and variance) are affected, and selecting the number of changepoints. The first issue has led to two main research streams: Hidden-Markov-based methods and segmentation methods. To our knowledge, segmentation methods are the more popular of the two, usually based on such segmentation criteria as likelihood ratio, least-squares, and cumulative sum of squares. In this research, we consider a likelihood-based framework to split time series occupancy data into segments.

There are three commonly used likelihood-based segmentation methods for multiple changepoints detection: Binary Segmentation (BS) (Scott & Knott, 1974), Segmentation Neighborhood (SN) (Auger & Lawrence, 1989), and optimal partitioning (OP) (Jackson et al., 2005). The first (BS) is an approximate algorithm with an $O(n \log n)$ computational cost because it considers only a subset of the 2^{n-1} possible solutions, where n is the number of data points. First, BS finds a single changepoint in the entire data, splitting the data into two at the changepoint location, if it is found. The procedure is repeated on the two sides of each location

until no further split in the data is justified. BS is fast, but it comes at a cost of not guaranteeing that the global minimum will be found. At $O(Qn^2)$, where Q is the maximum number of given changepoints, the SN method has better computational cost. It begins with setting up Q , then defines all possible segments between 0 and Q . The OP method, which has a computational cost of $O(n^2)$, begins by first conditioning on the last point of change. The algorithm then relates the optimal value of the cost function to the cost for the optimal partition of the data prior to the last changepoint plus the cost for the segment from the last changepoint to the end of the data.

In 2012, Killick et al. (Killick, Fearnhead, & Eckley, 2012) improved OP by proposing a new approach to search for changepoints—the pruned exact linear time (PELT) method—that has a computational cost that is linear in the number of data points. This pruning reduces the computational cost of the method, but does not affect the exactness of the resulting segmentation. PELT also produces quicker and more consistent results than identification “by eye” or assuming that the variability is constant (Killick, Eckley, Jonathan, & Chester, 2011; M.-H. Yang, Luong, & Recker, 2014).

3.2.3 Travel mode detection using GPS data

At this stage, a mode detection algorithm is built based on the period of staying or moving detected in the previous stage. This procedure could be a rule-based algorithm (Houston et al., 2013; Wu et al., 2011), fuzzy logic (Schönfelder, 2009; Schüssler, Montini, & Dobler, 2011), or data mining algorithm such as decision tree (Zheng, Li, Chen, Xie, & Ma,

2008a), neural networks (P. A. Gonzalez et al., 2010), and Bayesian Belief Network (Feng & Timmermans, 2013), etc.

Bohte and Maat (Bohte & Maat, 2009) determined whether the person walked, cycled or drove a car using the average and maximum speeds. They used GIS data to determine whether the trip was completed by train instead of by car, due to the similarity between speed of these two modes. The rule for checking train trip is whether at least 50% of all points of a trip lie within 50 m of the center of a rail track and whether the maximum speed was at least 20 km/h; if so, it is determined that a trip is likely to have been made by rail and that the respondent therefore probably took the train. A “missed” train trip in the data was imputed based on the starting and ending points of a trip, which were compared against the locations of train stations via underlying GIS maps. After the data was automatically interpreted, they were rechecked by the respondents.

Schüssler and Axhausen (Schuessler & Axhausen, 2009) and Schüssler et al. (Schüssler et al., 2011) used fuzzy logic to estimate the most likely transportation modes: walk, bike, car, urban public transport, and rail from GPS data. Four fuzzy variables were chosen: the median of the speed distribution of a stage, the median acceleration variation, the 95th percentile of speed, and the squared distance of the end of a stage to the nearest public transport stop. Once the membership functions for each variable were set up, fuzzy rules were derived. Each mode is described by at least one rule. According to the study, the accelerometer data show significant differences for the respective modes. Overall, 83% of the modes were identified correctly. The biggest misclassification problems occur between car and urban public transport, which was expected given the very similar speed and acceleration patterns of these two modes.

One particular application in artificial intelligence, neural networks, has been used to automatically detect the mode of transportation using such approaches as a Multi-Layer Perceptron (P. A. Gonzalez et al., 2010). A neural network takes a series of inputs, processes them through hidden nodes/layers and then generates an output that classifies the input data. The ability of a neural network to generalize conclusions for data, which do not exactly match the training data, is one of the significant strengths of this type of artificial intelligence. In Gonzalez et al.'s study, using cellphone to collect GPS data, a 10-fold cross-validation was performed to train and test the neural network. The N-fold cross-validation randomly partitions a set of data with known correct outputs into N folds, and then utilizes one fold as a testing data set while the remaining N-1 folds are used as the training data set. This process was repeated N times with a different fold being used as the testing data for each iteration. The highest accuracy accomplished for mode detection using this validation process was 92.11 percent of the car trips, 81.58 percent of the bus trips, and 100 percent of the walking trips for this test series. However, the primary attribution to this high accuracy is data sampling method. The research used only critical point data as inputs for the neural network. The Critical Point algorithm provides a dynamic update method without a fixed location update interval, and is capable of providing GPS data points collected at strategic locations of the path instead of random locations governed by a static update interval.

In another cellphone application, the work of Reddy et al. (Reddy et al., 2010) compared: (a) such instance classifiers as C4.5 Decision Trees (DT), K-Means Clustering (KMC), Naïve Bayes (NB), Nearest Neighbor (NN), and Support Vector Machines (a method to find a separating hyperplane in data space, which maximizes the margin between the two separated data sets), (b) a continuous Hidden Markov Model (CHMM), and (c) a two-stage system

involving the most accurate instance-based classifier (which is the decision tree) combined with a discrete Hidden Markov Model (DHMM) using a mobile phone equipped with a GPS receiver and an accelerometer. A decision tree followed by a first-order discrete Hidden Markov Model outperformed other methods and achieved an accuracy level of 93.6% in distinguishing between being stationary, walking, running, biking, and in motorized travel.

The work of Zheng et al. (Zheng, Liu, Wang, & Xie, 2008b) is again in the direction of inferring travel mode from non-walk GPS segments using machine learning techniques. The features from each segment then were extracted and sent to the inference model for four different values including Bike, Bus, Car and Walk. The features included: the length, the mean velocity, the expectation of velocity, the variance of velocity, the top three velocities and the top three accelerations of each segment. Two alternative ways were tested. In one approach, a classifier was employed to perform inference with the consideration that the segments of GPS tracks were independent instances. In a second approach, Conditional Random Field (a framework for building probabilistic models to segment and label sequence data) was applied to GPS data that were considered as a kind of sequential data. Their results indicate that DT outperformed such other inference models as Support Vector Machines, Bayesian Net, and Conditional Random Field. The study also showed the disadvantages of the other inferences: for example, BN does not hold the probabilistic independencies among such selected features as average velocity and maximum velocity; in Linear Support Vector Machines, it is hard to handle multiclass classification problems and probabilities of each class in its inference results are not obtainable.

In another trend of detecting transportation mode, Feng and Timmermans (Feng & Timmermans, 2013) combined accelerometer-GPS data. The accelerometer data provide

information about the change in three-axis acceleration with respect to the device, moving or non-moving of the device, and state of the device (turned off, sleeping, etc.). A Bayesian Belief Network model is used to infer 8 transportation modes (walking, running, bicycle, motorcycle, bus, car, tram and metro) and activity episodes simultaneously. Average speed, maximum speed for the three minutes epoch, non-moving time duration, average value and standard deviation of the three-axis acceleration change, horizontal measurement error, number of satellites used for position calculation, and car/bike/motorcycle ownership were used as inputs for the inference model. The model led to increasing in accuracy by 10% for validation data and 15% for calibration data when adding accelerometer data. This knowledge provides a potential advantage of the combining accelerometer and GPS data to infer transportation mode and activity episode. However, regarding the characteristics of the accelerometer sensitivity, the study also suggested that the effects of where the device was need further investigation. Different profiles, then different classification methods, may exist in the accelerometer data between holding the device and putting the device into the pocket when running.

3.2.4 Summary

Although the research described has generally done well in detecting transportation modes from GPS data, there are a number of open questions left. The first open question is the automatic detection change points—the point to separate moving and not-moving segments. Although some authors have attempted to tackle this problem, a robust method for automatically detecting any change point is still lacking. The second question is detecting travel mode. Although some existing research has provided relatively high accuracy in such prediction,

it is nonetheless limited in terms of the number of modes and/or specific data set. Furthermore, most of the studies rely heavily on such transportation network information as distance to transit stop or transit line. Further research is needed to consider a more comprehensive list of modes as well as to improve the prediction method based solely on such raw GPS data as speed.

The motivation for this study stems from the limitations of the above research. In the following, the focus is on development of a model of extracting segments from raw GPS trajectory, using multiple changepoints detection method, and inferring travel mode for each segment.

3.3 Methodology

3.3.1 Study framework

The proposed framework for predicting travel mode from raw GPS data is shown in Figure 3-1. After the data are cleaned by Naïve filter and Kalman filter, the cleaned GPS tracks (trajectories) are segmented by a multiple changepoints algorithm (stage 1). Then characteristics of each segment are extracted. A sample of the data is manually processed and labeled with moving/non-moving status and travel mode. A part of this sample is then used as training data to calibrate the models. The rest of the sample is used as testing data for validation.

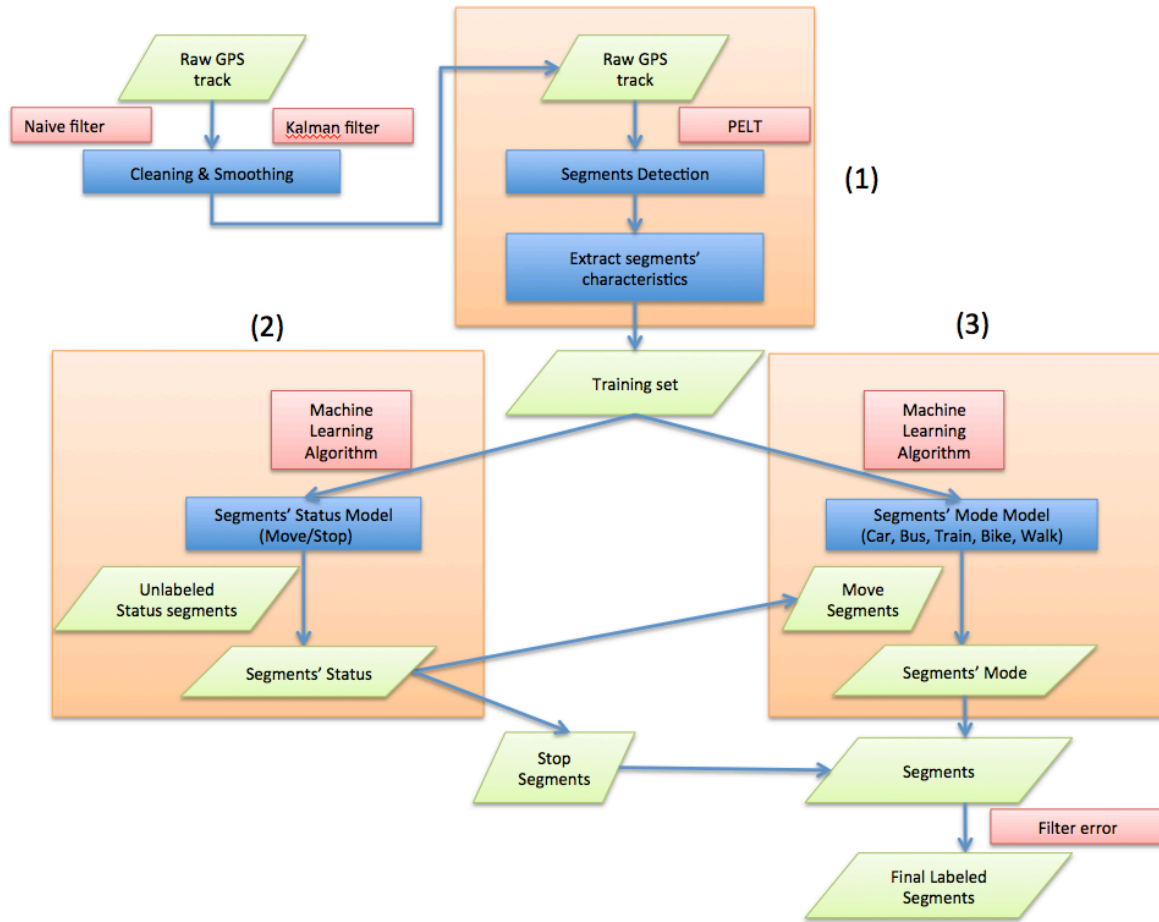


Figure 3-1 Travel Mode Detection Framework

In the next stage, a model of classification of a segment's status will be built based on the whole segments in the training set. This model then is applied to predict segments' status for unlabeled data.

At stage 3, only moving segments in the training set are selected for building the travel modes model. Then we apply this model to the predicted moving segments in the previous stage to get travel mode for each segment.

The combination of predicted stop segment and travel mode for each segment is a sequential segment. Because the prediction models could have error, a final step needs to be applied to filter the error.

The details of each stage are explained in the following sections. From the final results, we could connect the sequential segments with same travel mode into a trip and sequential stop segments become locations.

3.3.2 Data Cleaning

3.3.2.1 Common methods in data cleaning

Since GPS monitoring produces volumes of temporal and spatial data, substantial post-processing is needed to: 1) identify the locations and activity types associated with sequential GPS locations, 2) account for missing data associated with participant noncompliance or device errors, 3) address spatial positional inaccuracies associated with satellite signal obstructions, 4) identify unique trips during travel periods, and 5) tabulate daily trip counts by mode and walking duration (Bricka, Sen, Paleti, & Bhat, 2012; Cho, Rodríguez, & Evenson, 2011; Schuessler & Axhausen, 2009).

Data cleaning detects and corrects data that are either inaccurate, or incomplete or irrelevant. Normally, speed is used to detect unrealistic position jumps. Speed is directly derived from the calculated positions. Errors in maximum speed could be caused by poor satellite reception, causing misplaced registration and wrong estimation of speeds. There are three main techniques that are often used for data filtering and smoothing: using rules or sequential state estimation techniques, probabilistic algorithms like Kalman filter, and particle filtering schemes.

Roider (Roider et al., 2012) filtered unrealistically high speeds (URHS) movements by applying a rule. First, if two consecutive points are URHS, the authors deleted the first point and repeated until no URHS points remain. Bohte and Maat (Bohte & Maat, 2009) neglected the speed of the points with the highest speed, and the speed recorded for the following point was used.

The basic idea of a Kalman filter (KF) (Kalman, 1960) is to reduce noise in data. The derivation of the KF is based on a state-space model, with two assumptions: linearity of the model, and Gaussianity of both the dynamic noise in the process equation and the measurement noise in the measurement equation (Pantic, Pentland, Nijholt, & Huang, 2007).

The scheme of KF is shown in Figure 3-2 with two steps in each cycle: prediction and correction. The time update equations are responsible for projecting forward the current state and error covariance estimates. This step obtains the *a priori* estimates for the next time step. The states get updated based on the known system inputs and system dynamic where the state is the smallest vector that summarizes the past of the system in full. The state covariance matrix is updated by the state matrix and process noise covariance matrix. The second equations are responsible for the feedback in order to obtain an improved *a posteriori* estimate.

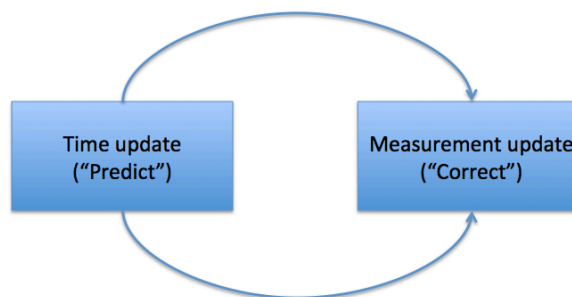


Figure 3-2: Kalman Filter scheme

Consider a linear discrete-time system as follows:

$$x_k = F_{k-1}x_{k-1} + u_{k-1} + w_{k-1}$$

$$y_k = H_k x_k + v_k$$

where x is state and y is output with some noise processes w and v , respectively. These noises are white, zero-mean, and uncorrelated with covariance matrixes Q_k and R_k .

$$w_k \sim (0, Q_k)$$

$$v_k \sim (0, R_k)$$

The goal is to estimate the state x_k at time k based on the known system dynamics and from the noisy measurements $\{y_k\}$. Given the initial conditions $x_{0|0}$ and $P_{0|0}$, the complete recursion in the Kalman filter is computed iteratively with the following steps (Andrade-Cetto & Sanfeliu, 2006):

- Predict the a priori state, error covariance, and observation estimates

$$x_{k|k-1} = F_{k-1}x_{k-1|k-1} + u_{k-1}$$

$$P_{k|k-1} = F_{k-1}P_{k-1|k-1}F_{k-1}^T + Q_{k-1}$$

$$y_{k|k-1} = Hx_{k|k-1}$$

- Compute the Kalman gain matrix K to minimize the a posteriori error covariance and correct the state and state error covariance estimates

$$K = P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1}$$

$$x_{k|k} = x_{k|k-1} + K(y - y_{k|k-1})$$

$$P_{k|k} = P_{k|k-1} - KHP_{k|k-1}$$

Nitsche et al. (Nitsche, Widhalm, Breuss, Brändle, & Maurer, 2013) pre-processed GPS tracks with a KF by combining the raw GPS and cell location data with predictions of a linear motion model assuming zero mean, Gaussian-distributed accelerations. The method's limitation is in its key assumptions—speed changes are generally nonlinear and quite often non-Gaussian. In particular, in cases of a sudden or rapid movement when a signal jumps, the prediction of positions from KF might be off significantly. However, the method is popular and easily applied.

Some researchers have proposed particle filter—a nonlinear estimation strategy—as an alternative method to the classical KF (Vernaza & Lee, 2006). The procedure finds an approximate solution using a complex model rather than an exact solution using a simplified model (as in KF). The main idea behind particle filtering is to maintain a set of solutions that are an efficient representation of the conditional probability $p(\alpha / Y)$, where α is the state of a temporal event to be tracked given a set of noisy observations $Y = \{y_1, \dots, y^-, y\}$ up to the current time instant (Pantic et al., 2007). This means that the distribution $p(\alpha / Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k , then it is as if s_k were drawn from $p(\alpha / Y)$. By maintaining a set of solutions instead of a single estimate (as is done by KF), particle filtering is able to track multimodal conditional probabilities $p(\alpha / Y)$, and it is therefore robust to missing and inaccurate data, and particularly attractive for estimation and

prediction in nonlinear, non-Gaussian systems. However, the method requires high computational effort, and is not applied in this research.

In the following subsections, we compare the rule-based and Kalman filter techniques to decide which is more suited for this research.

3.3.2.2 Data cleaning results

The following figure shows results of two methods: rule-based and Kalman filter based on 1 day example. For rule-based method, a “naïve” rule is applied: if a point has URHS (over 100 mph), its speed is replaced by the speed of the previous point until there is no URHS in the data. Based on our observation, smoothing data using Kalman filter reduces the data noise the most, however the method does not remove URHS entirely. Therefore, in this study, we combine both methods for data cleaning and smoothing task in this study: first, Kalman filter is applied then rule-based algorithm is used to remove URHS.

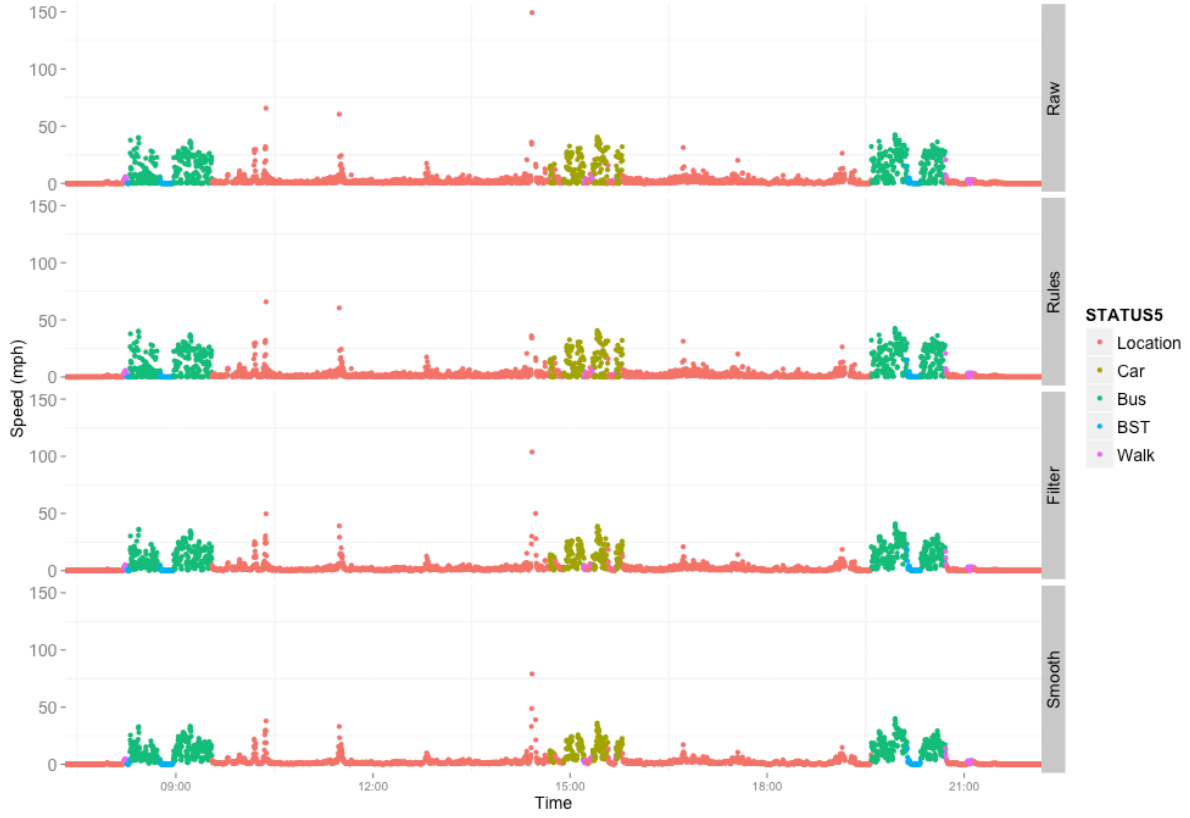


Figure 3-3: Data cleaning and smoothing result

3.3.3 Segmentation algorithm

In this study, PELT is adopted as the segmentation approach because the algorithm produces a computational efficiency of $O(n)$. Assume that we have n data points $y_{1:n} = (y_1, \dots, y_n)$ with m changepoints splitting the data into $m + 1$ segments, together with their positions, $\tau_{1:m} = (\tau_1, \dots, \tau_m)$. We assume that $\tau_0 = 0$ and $\tau_{m+1} = n$ and assume that the changepoints are ordered such that $\tau_i < \tau_j$ if, and only if, $i < j$. According to Killick *et al.* (Killick *et al.*, 2012), assuming C is a cost function for a segment and $\beta f(m)$ is a penalty to guard against overfitting, the commonly used approach to identify multiple changepoints is to minimize:

$$\sum_{i=1}^{m+1} [C(y_{\tau_{i-1}+1} : \tau_i)] + \beta f(m)$$

A key assumption is that of a penalty linear in the number of changepoints m . Pseudo-code for the PELT method is as follows (Killick et al., 2012):

Input: Time series data with n data points: $y_{1:n} = (y_1, \dots, y_n)$

A measure of fit $C(\cdot)$ dependent on the data

A penalty constant β which does not depend on the number or location of changepoints

A constant K such that for $t < s < T$

$$C(y_{(t+1):s}) + C(y_{(s+1):T}) + K \leq C(y_{(t+1):T})$$

Initialize: Set n = data size

$$F(0) = -\beta, \text{cp}(0) = \text{NULL}, R_1 = \{0\}$$

Iterate For $\tau^* = 1, \dots, n$

1. Calculate $F(\tau^*) = \min_{\tau \in R_{\tau^*}} [F(\tau) + C(y_{(\tau+1):\tau^*}) + \beta]$
2. Let $\tau^1 = \arg\{\min_{\tau \in R_{\tau^*}} [F(\tau) + C(y_{(\tau+1):\tau^*}) + \beta]\}$
3. Set $\text{cp}(\tau^*) = [\text{cp}(\tau^1), \tau^1]$
4. Set $R_{\tau^*+1} = \{\tau^* \cup \{\tau \in R_{\tau^*} : F(\tau) + C(y_{(\tau+1):\tau^*}) + K < F(\tau^*)\}\}$

Output: The changepoints recorded in $\text{cp}(n)$

The purpose of this pruning process is to remove those values of τ that are known not to lead to optimality, improving computational efficiency.

3.3.4 Classification and prediction algorithm

Artificial Intelligence (AI) is a field that initially was used to create artificial machines after World War II. The field was first identified by Alan Turing in 1950 in his paper: "Computing Machinery and Intelligence". Since then, the field has been highly developed and widely applied to many areas of science and engineering. One branch of AI is Machine Learning, which deals with creating algorithms and techniques that enable learning from input data to create patterns or prediction functions to be used to predict an unknown dataset. Here, we use machine learning principally because, for the problem considered, there are no rules that are both simple and reliable.

We apply a machine learning technique for classifying travel modes. A decision tree approach often has overfitting problems on the training set. Random forest (RF) is a group of decision trees (Breiman, 2001) for classification and regression with each tree depending on a collection of random variables. This is the most accurate learning algorithm currently available, and can mitigate the particular shortcomings of decision tree algorithms. The idea is to build a classification method by combining multiple models in order to obtain better predictive performance compared to a single model. Each tree is built depending on a collection of random variables. Because each tree is grown to its maximal depth instead of pruning, this ensemble method is less biased. Although each individual tree might overfit the data (has irrelevant predictors), the final model is outweighed by the multiple trees using different variables that (over) fits the data differently (Williams, 2011). A model that overfits the data has low bias but high variance, which means that it predicts the targets in the training set very

accurately, but any slight changes to the predictors result in vastly different predictions for the targets.

The RF algorithm is popular in the life sciences because it is robust to large amounts of noise, requires little parameter tuning and requires no predictor transformation (A. Cutler, Cutler, & Stevens, 2012). Further, RF provides information about which variables are the most significant in the prediction process; it can also be used to avoid any features that make no difference in the prediction model.

In addition, RF has the capability to handle non-linearities and multi-collinearity problems (has redundant predictors) in multinomial logit models. These characteristics make RF a promising classification algorithm for the study. Our procedure is to use a RF algorithm in a choice model to build and analyze the transportation mode based on a label sample. Once a RF has been trained, the trees function as an “ensemble of experts” to make predictions for all of the GPS tracks.

The following is a brief description of the types of classification trees used in the RF algorithm. In constructing each tree of a RF, one begins with a “training set” containing many instances of predictor variables along with an associated “truth” value. The random forest training procedure creates a “bagged” sample by drawing a random subset of n instances from the n -member training set, with replacement after each draw so that any instance may be selected more than once. This means that, on average, each tree is trained on roughly $2/3$ of the dataset, half of which are duplicated. Then, at each node of the tree, a subset of only m randomly-selected feature variables are chosen as candidates for splitting, contrasting with the usual practice of choosing the best split from all the feature variables. A typical choice for m is

the square root of the number of predictor fields. The trees are grown without pruning. The resulting trees are combined by unweighted voting if the response is categorical (classification) or unweighted averaging if the response is continuous (regression).

To provide background on the algorithm, we follow the work of Cutler et al. (A. Cutler et al., 2012). Let $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ denote the training data, with $x_i = (x_{i,1}, \dots, x_{i,p})^T$ denoting the p predictors and y_i denoting the response. A tree is denoted as $h_j(X, \theta_j)$. Here, θ_j is a collection of random variables and θ_j 's are independent for $j = 1, \dots, J$. The fitted tree is denoted as $\hat{h}_j(x, \theta_j, D)$ where θ_j is a particular realization of θ_j . Randomness is considered in two ways. First, as with bagging, each tree is fit to an independent bootstrap sample from the original data. The randomization involved in bootstrap sampling gives one part of θ_j . Second, when splitting a node, the best split is found over a randomly-selected subset of m predictor variables—instead of all p predictors—independently at each node. The randomization used to sample the predictors gives the remaining part of θ_j .

The Pseudo-code is shown as below (A. Cutler et al., 2012):

For $j = 1$ to J :

1. Take a bootstrap sample D_j of size N from D .
2. Using the bootstrap sample D_j as the training data, fit a tree using binary recursive partitioning
 - a. Start with all observations in a single node.
 - b. Repeat the following steps recursively for each unsplit node until the stopping criterion is met:
 - (i) Select m predictors at random from the p available predictors.

(ii) Find the best binary split among all binary splits on the m predictors from Step (i).

(iii) Split the node into two descendant nodes using the split from Step(ii).

To make a prediction at a new point x ,

$$\hat{f}(x) = \arg \max_y \sum_{j=1}^J (\hat{h}_j(x) = y)$$

where $\hat{h}_j(x)$ is the prediction of the response variable at x using the j th tree.

The error rate in each class is defined as the proportion of misclassified observations in just that class, whereas the overall misclassification rate is the proportion of misclassified observations for the entire data set. For each tree, the prediction accuracy for the out-of-bag (OOB) (training) portion of the data is recorded. The proportion of times the result is not accurate averaged over all samples is the OOB error estimate. If OOB error value is $\alpha\%$, this suggests that the model has $(1-\alpha)\%$ out-of-sample accuracy for the training set. Then the same is done after permuting each predictor variable. The differences between the two accuracies are then averaged over all trees, and normalized by their standard deviation (Breiman, 2001).

To evaluate a classifier, we use criteria calculated from the confusion matrix (Table 3-1). For each cell in the matrix we have fields as: True Positives, False Positives, False Negatives and True Negatives. These are defined as:

- False Positive (FP) means falsely predicting a label. It measures the number of negative cases that are incorrectly identified as positive cases.
- False Negative (FN) means missing label. It measures the number of positive cases that are misclassified as negative cases.

- True Positive (TP) means correctly predicting a label. It measures the number of positive cases that are correctly identified as positive, and
- True Negative (TN) means correctly predicting the other label. It measures the number of negative cases that are correctly identified as negative cases.

Table 3-1: Classification matrix for binary classification problem

	Actual +	Actual -
Predicted +	True Positives (TP)	False Positives (FP)
Predicted -	False Negatives (FN)	True Negatives (TN)

The criteria for measuring the model's accuracy are as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Class\ error = \frac{FP}{TP + FP}$$

$$Precision = 1 - Class\ error$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$OOB = 1 - Accuracy$$

For the multi-class classification problem (where the number of classes is more than two), the confusion matrix is a square of $m \times m$, where m is the number of classes. The above

measurements are extended into m dimensions, where the True for each class is the cell in the diagonal of the confusion matrix.

In addition, to measure a variable's importance to the measurements for each model, we use Mean Decrease Accuracy, which measures how much inclusion of this predictor in the model reduces classification error (Williams, 2011).

3.4 Framework of experiments

In this section, we describe the framework of experiments including features of each algorithm, results, and validation. The overall scheme is shown in Figure 3-3.

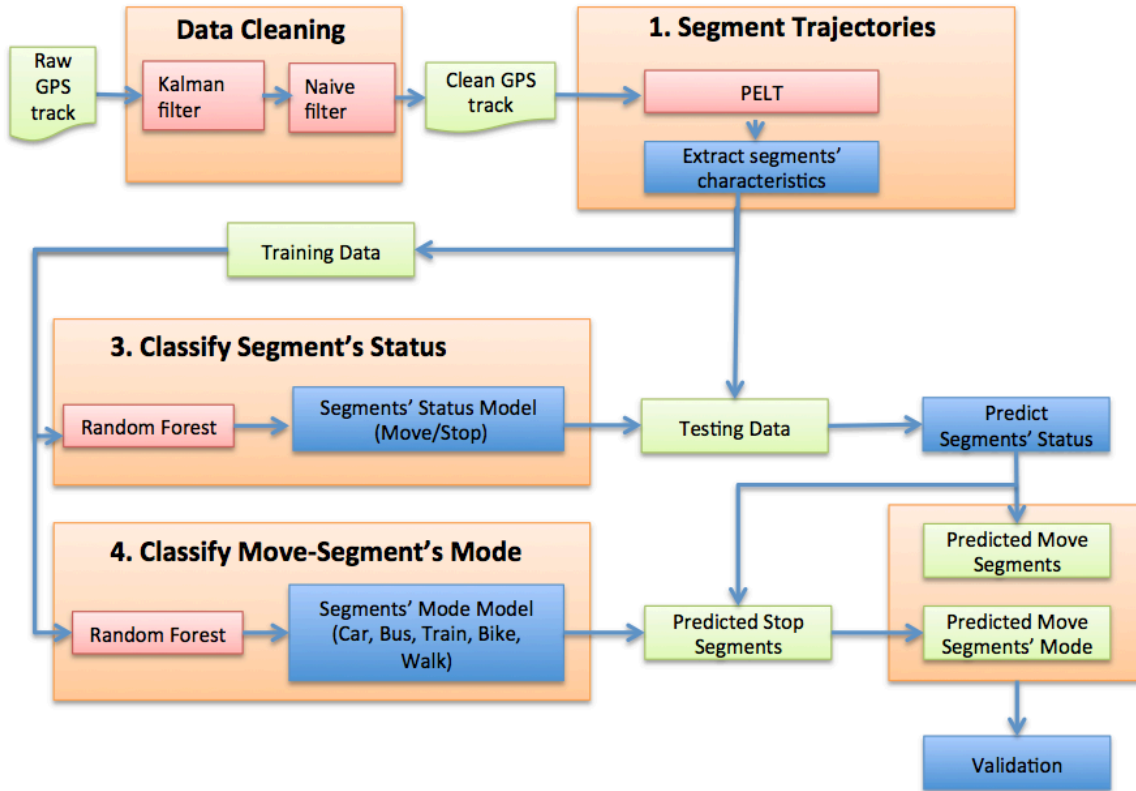


Figure 3-4: Experiment on Expo data

3.4.1 Segmentation results

In order to calibrate and validate the methodology, 60 travel days from 24 participants were chosen from wave II data. First, the clean data pertaining to all of the days then were segmented by PELT. The penalty value was set as $f(m) = \log(n)$. The cost function was set to be minus a penalized log-likelihood. We set the constant K equal to the penalization factor $\log(n)$. The data after extraction is saved in a shorter format (see Table 3-2). Figure 3-5 shows partial data of the sample period after segmentation. The blue ribbons illustrate the area between [mean – one standard deviation, mean + one standard deviation]. Estimated mean value is represented by the red line.

Table 3-2: Sample of data after segmentation

Segment	HID	Date	Start	End	Duration	Mean.	Sd.	Label
447	33BE	9/28/12	8:03:14	8:04:14	1	0.05105	0.01108	Location
448	33BE	9/28/12	8:04:29	8:05:59	1.5	0.25881	0.19129	Location
449	33BE	9/28/12	8:06:14	8:07:14	1	0.01571	0.00842	Location
450	33BE	9/28/12	8:07:29	8:08:14	0.75	0.02086	0.00034	Location
451	33BE	9/28/12	8:08:29	8:08:44	0.25	0.03169	0.00372	Location
452	33BE	9/28/12	8:08:59	8:10:14	1.25	0.04308	0.00359	Location
453	33BE	9/28/12	8:10:29	8:10:44	0.25	0.08879	0.03509	Location
454	33BE	9/28/12	8:10:59	8:11:14	0.25	0.49684	0.31116	Location
455	33BE	9/28/12	8:11:29	8:11:59	0.5	2.43622	0.55027	Location
456	33BE	9/28/12	8:12:14	8:12:29	0.25	3.32018	0.01737	Walk
457	33BE	9/28/12	8:12:44	8:13:29	0.75	4.26103	0.39041	Walk
458	33BE	9/28/12	8:13:44	8:13:59	0.25	3.6126	0.02213	Walk
459	33BE	9/28/12	8:14:14	8:14:59	0.75	2.06855	1.21475	Walk
460	33BE	9/28/12	8:15:14	8:16:29	1.25	0.155	0.09285	Location
461	33BE	9/28/12	8:16:44	8:17:14	0.5	2.20451	1.88666	Location
462	33BE	9/28/12	8:17:29	8:19:29	2	14.82846	5.73056	Bus
463	33BE	9/28/12	8:19:44	8:20:44	1	4.14838	0.67621	Bus
464	33BE	9/28/12	8:20:59	8:22:14	1.25	15.09673	3.4461	Bus
465	33BE	9/28/12	8:22:29	8:22:59	0.5	4.6098	1.14482	Bus
466	33BE	9/28/12	8:23:14	8:24:59	1.75	13.21358	6.06358	Bus

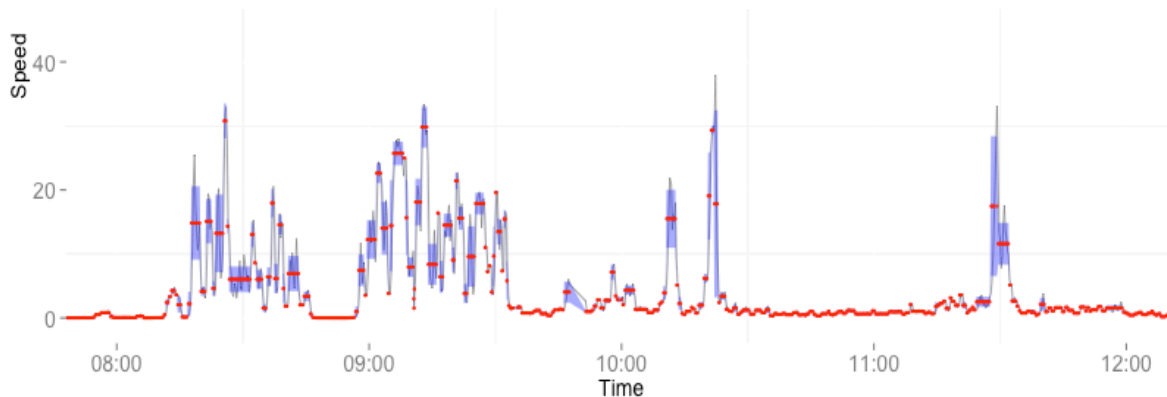


Figure 3-5: Changepoints in speed, penalty value = $\log(n)$

3.4.2 List of predictors

After using PELT to divide GPS tracks into segments, the first stage of classification using RF is based on the characteristics of these segments and the label of each segment. Each segment is classified as one of five modes: location (not moving, including transit station), car, bus, train, and walk. For each segment, a total of 44 features are used as attributes for the RF implementation. In general, speed is measured by miles per hour, and duration is measure by minutes.

- Mean, standard deviation, maximum, and 75 percentile of speed
- Mean, standard deviation, maximum, and 75 percentile of acceleration
- Duration
- Distance of travel includes: total distance, straight distance from the first point to last point, centroid distance: distance from centroid of a segment to the centroid of its previous segment.

- Characteristics of 4 segments before current segment: duration, mean speed, mean acceleration
- Characteristics of 4 segments after current segment: duration, mean speed, mean acceleration
- Distance to the closest transit stop: mean, variance, maximum, and minimum distance
- Distance to the closest transit line: mean, variance, maximum, and minimum distance

The name of the variable is created by using the calculation method plus the measurement's name. For example, "min_dline" is the minimum distance to transit line.

Acceleration is calculated by this formula: $a_i = |V_{i+1} - V_i| / \Delta T_i$ where i is the order of a segment, a is acceleration, V is speed, and ΔT is time difference.

The centroid point of each segment is calculated by one of two methods—average centroid or kernel estimation—depending on the number of data points in a segment.

Distance to transit stop/line is obtained from the SCAG 2012 data set using ArcGIS software (ESRI, Redlands, CA). We build two models using different combination of above variables. The first model-baseline model- uses all variables except distance to transit stop and transit line. The second model includes these two factors and is named improved model.

3.4.3 Classification Results

After segmentation of the data, two thirds of the days were used as the training set to define the patterns of biking, walking, driving, and taking train or bus. The training set includes

54,470 segments. The testing set has 19,026 segments. The pattern of each mode is shown as the following.

Table 3-3: Speed pattern of each mode

Mode	Mean	Sd	3 rd Qu.	Min	Max	# Points	# HID	# Day	# Hour
Bike	7.256	4.794	10.100	0.040	20.400	1129	3	6	5.729
Bus	12.202	8.708	16.460	0.125	66.320	9140	16	44	36.563
Car	18.230	14.365	24.260	14.840	74.810	6416	18	35	26.18
Train	16.275	11.287	23.260	0.000	50.490	2012	8	14	10.815
Walk	2.402	1.150	2.966	0.003	23.140	9060	20	48	40.088

Then, RF was used for building a model for classifying both move/stop segments and travel modes from moving segments. The number of predictors for each model is the same, and each model also uses 6 factors (speed, acceleration, duration, travel distance, characteristic of segments before and after each segment) at each split with 1000 and 2000 trees for the first and second models, respectively. There are many implementations of RF algorithm that differ in the accuracy assessment method used. In this study, the models are built from the R *randomForest* package (Liaw & Wiener, 2012) using the original code by L. Breiman and A. Cutler (Breiman, 2001).

For the moving points model, the OOB estimate of error rate is 2.18%. Table 3-4 shows the confusion matrix for the model using training data. There were very few false negatives in the moving class. The class error for moving segments is only 14.75%. The model identifies 4,133 segments (85.25%) as moving segments as they are labeled. There are 472 segments (10.25%) moving segments that are mistakenly identified as not moving ones (stops).

Table 3-4: Classify moving points model – baseline model – confusion matrix –**training set**

Predicted\Actual	Move	Stop	Class error
Move	4,133	715	0.1475
Stop	472	49,150	0.0095

The OOB estimate of error rate was 29.43% for the travel mode model on the training set. Table 3-5 shows how heavily the misclassification leans toward train trips (78.09%). We found that train tends to have substantially more confusion with bus and car than with bike and walk. Meanwhile, walking trips have the lowest error rate (4.99%). This is not surprising because walking's pattern is more unique with low speed and variance.

Table 3-5: Classify travel mode model – baseline model – confusion matrix –**training set**

Predicted\Actual	Bike	Bus	Car	Train	Bus	Class error
Bike	119	43	29	0	47	0.5000
Bus	14	1026	317	15	63	0.2850
Car	21	443	781	17	34	0.3974
Train	3	151	146	87	10	0.7809
Walk	12	47	11	4	1408	0.0499

For the model distinguishing moving and not moving status of a segment, distance from centroid of a segment to that of the previous segment and the 75th percentile of the segment's speed contribute the greatest importance. The results are consistent in both measurements. In the method of using mean decrease in accuracy, average speed of the 4th segment prior to a segment and standard deviation of acceleration are the next two variables in importance (Figure 3-6a).

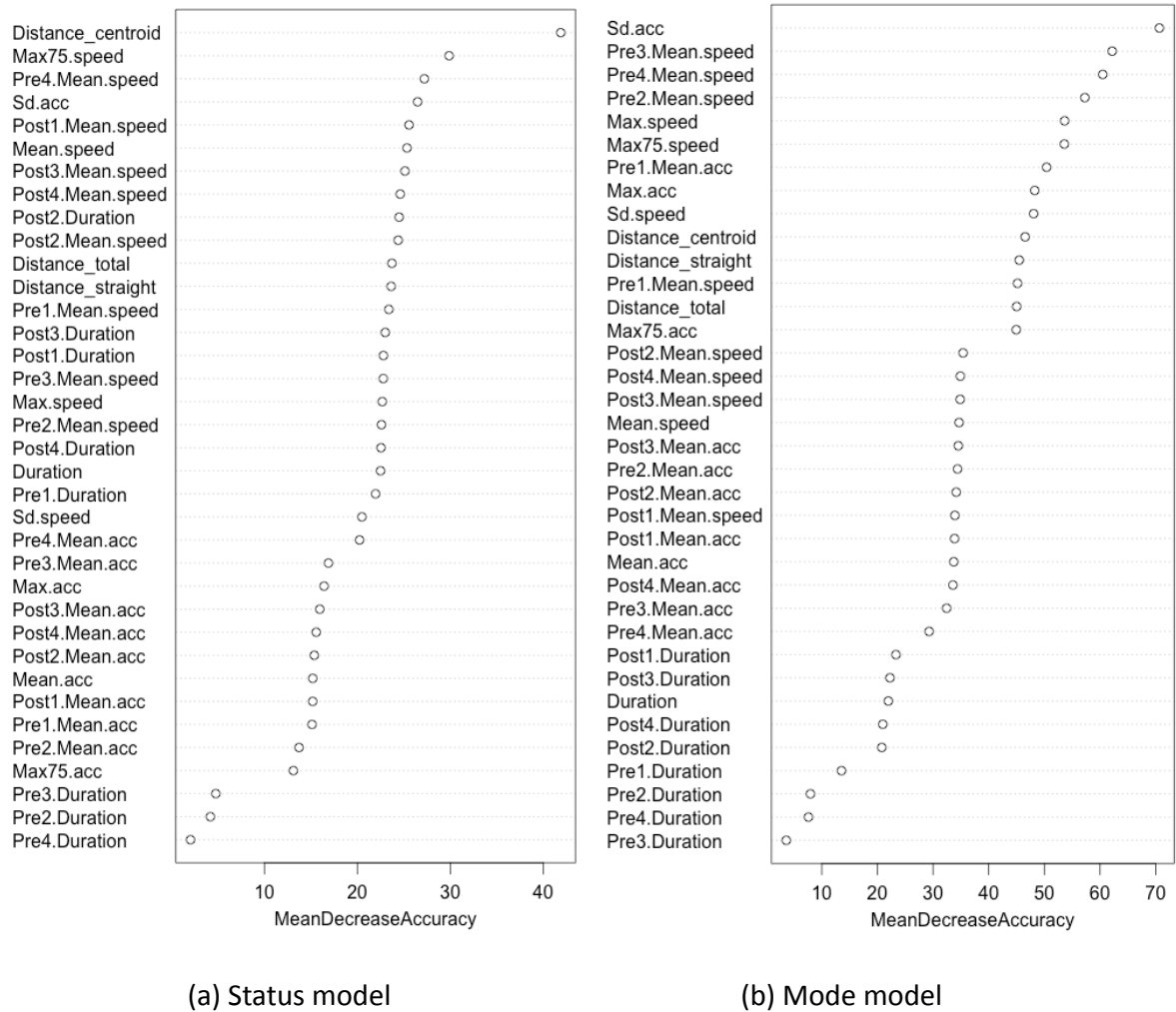


Figure 3-6: Important variables in status model

Figure 3-6b shows that the standard deviation of acceleration has the greatest importance in prediction of travel mode. The next five most important predictors are related to speed, including speeds of previous segments and maximum speed in the current segment. Post segment's characteristics have less impact on the prediction power. Duration factors affect the prediction accuracy the least (including duration of 4 previous segments). The break between Max75.acc and Post2.Mean.speed can help decide how to reduce dimensions of the final model.

3.4.4 Improved mode classification model

In order to improve the accuracy of the second model, we use two additional factors, which are measured for each GPS location: distance to the closest transit line and distance to the closest transit stop. For each segment, the following values are calculated: mean, variance, maximum, and minimum distance to the closest transit line and the closest transit stop. For the status model, the additional variables help decrease the OOB error rate to 0.47% for the training set. Table 3-6 shows the confusion matrix for the model using the training data. Nearly every segment that is moving is correctly identified as such with 95.40% recall; there were very few false negatives in the moving class. The class error for moving segments is only 4.6%. The model identifies 4,625 segments (99.25%) as moving segment as they are labeled. Only 35 segments (0.75%) moving segments are mistakenly identified as not moving ones (stops). This model has higher accuracy than the model with only 6 base factors.

Table 3-6: Classify moving points model with transit distance – confusion matrix – training set

Predicted\Actual	Move	Stop	Class error
Move	4625	223	0.0460
Stop	35	49587	0.0007

Table 3-7: Classify travel mode model with transit distance – confusion matrix – training set

Predicted\Actual	Bike	Bus	Car	Train	Walk	Class error
Bike	116	39	40	0	43	0.5126
Bus	5	1124	230	10	66	0.2167
Car	15	344	888	15	34	0.3148
Train	0	150	127	114	6	0.7128
Walk	6	31	20	1	1424	0.0391

The OOB error rate estimate is 24.38% for the travel mode model in the training set. Table 3-7 shows how heavily the misclassification leans toward train trips (71.28%). As with the previous model, we found that train tends to have more confusion with bus and car than with bike and walk—a result that may be explained by these three modes having similar patterns in speed (Figure 3-7). Meanwhile, walking trips have the lowest error rate (3.91%). Again, this is not surprising because walking’s pattern is more unique with low speed and variance. The most interesting result is the accuracy of classifying bus trips, which has an error rate of 21.67%, meaning that more than three quarters of segments that are bus are correctly identified as such, with 78.33% recall. With the exception of biking trips, the error rates of the other modes are lower than the error rate in the model without transit related measurements. However, the difference is small (less than 8%).

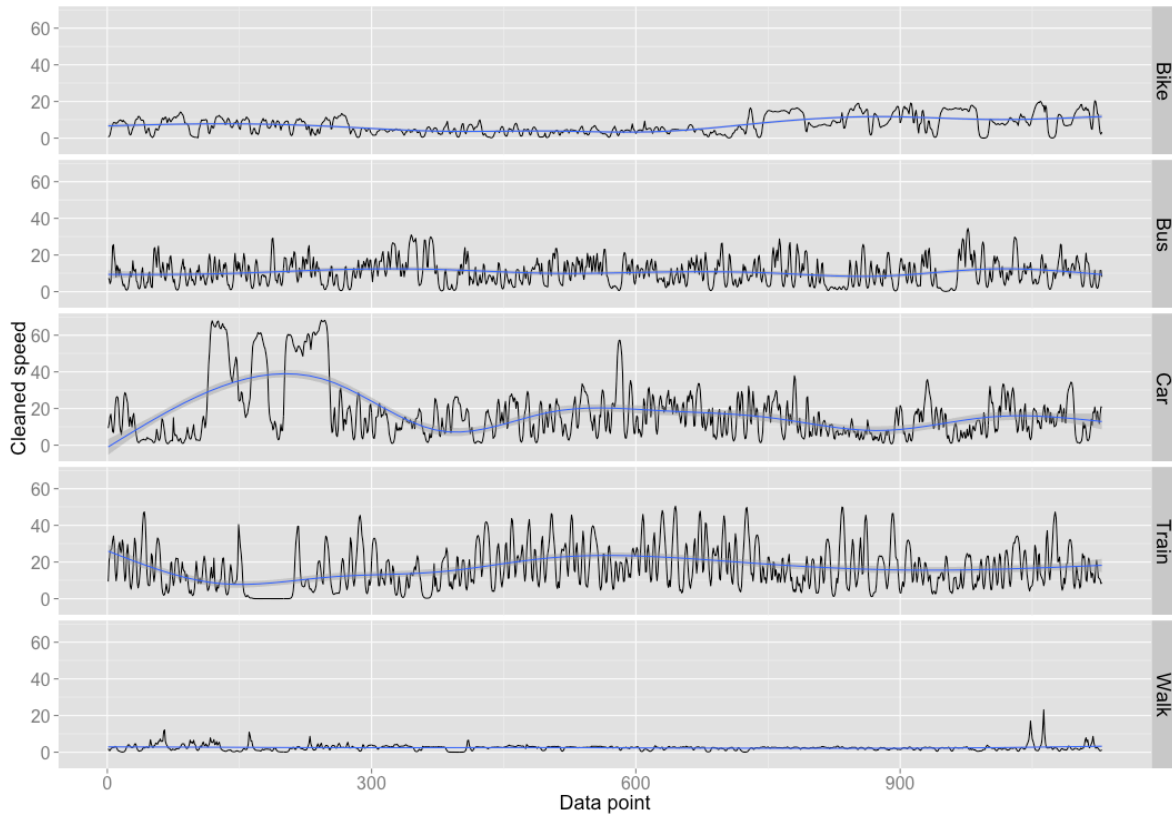
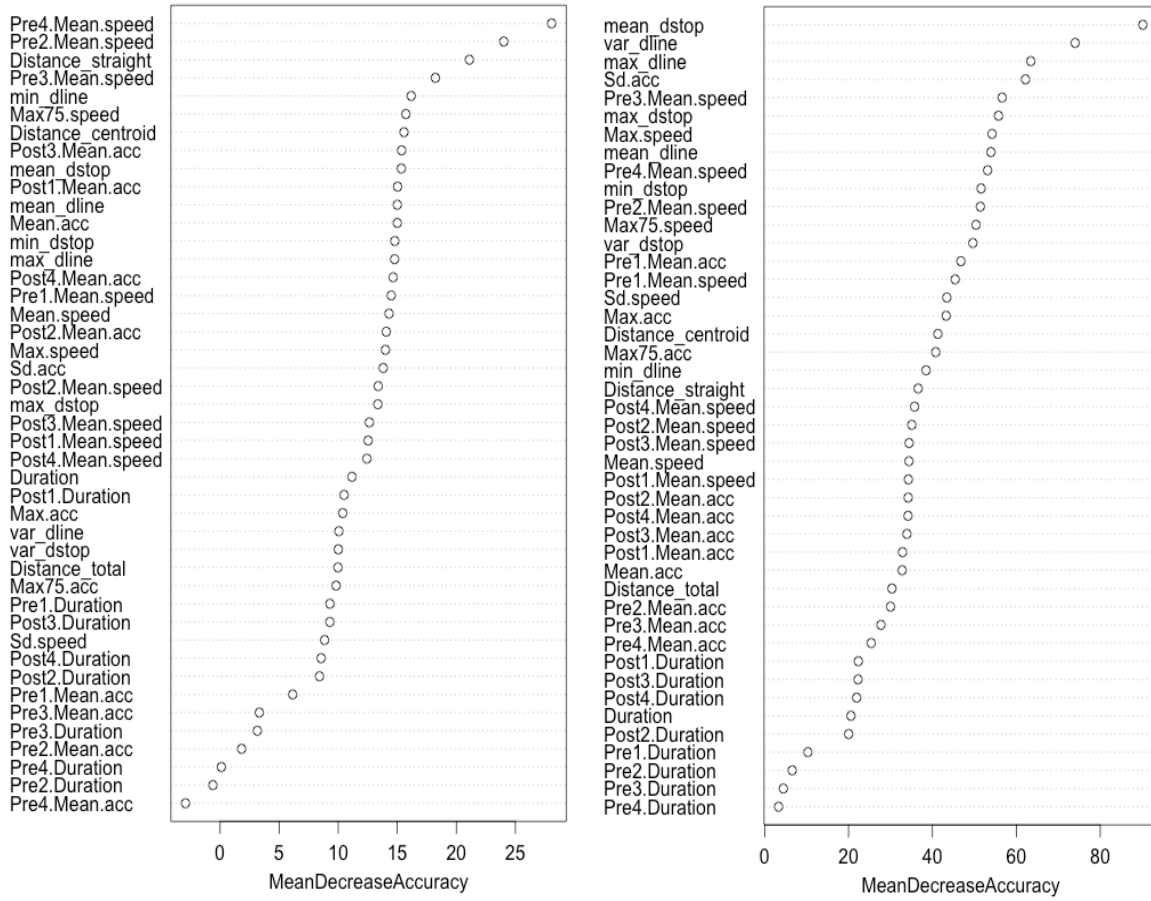


Figure 3-7: Speed pattern of each mode (first 1000 points)

The results of the models' variables importance measurements are shown in Figure 5. The values are sorted in decreasing order (from top to bottom) of attributes as assigned by the RF. For the model that distinguishes moving and not moving status of a segment (Figure 5a), the previous three segments' average speed, distance from the first point to last point of a segment, and minimum distance to transit line contribute the highest importance. Previous segments' duration and average acceleration are the least important factors. Figure 5b shows that mean distance to transit stop, variance distance to transit line, and maximum distance to transit line of each segment are the most important predictors of travel mode. All of the duration measurements rank last in importance.



(a) Status model

(b) Mode model

Figure 3-8: Important variables in mode model

3.4.5 Model validation

In this section, the accuracy of both baseline models and improved models will be compared using testing data. The baseline models' error rate is 2.63% and 39.61% for the first and second models, respectively, while these values are 0.49% and 23.90% when using improved models. Based on the improvement in accuracy obtained with the improved (full) model we use the full model as the model for detecting travel. Combining predicted none-moving segment model to the mode model, produces a result that reaches an accuracy of

97.63%—an increase of 2.33% compared to the baseline models' accuracy (95.30%). The final confusion matrix is shown in Table 3-8.

Table 3-8: Final result– confusion matrix – testing set

Predicted\Actual	Bike	Bus	Car	Location	Train	Walk	Class error
Bike	20	12	7	13	0	0	0.6154
Bus	12	385	139	29	4	13	0.3385
Car	6	125	94	31	3	2	0.6398
Location	11	55	22	17,284	1	71	0.0092
Train	1	24	28	1	16	1	0.7746
Walk	3	14	1	266	0	332	0.4610

(a) Baseline mode: Without transit distance

Predicted\Actual	Bike	Bus	Car	Location	Train	Walk	Class error
Bike	20	12	9	5	0	6	0.6154
Bus	6	425	98	21	2	30	0.2698
Car	5	103	121	9	4	19	0.5364
Location	0	3	3	17,435	0	3	0.0005
Train	0	20	23	0	27	1	0.6197
Walk	4	13	1	49	2	547	0.1120

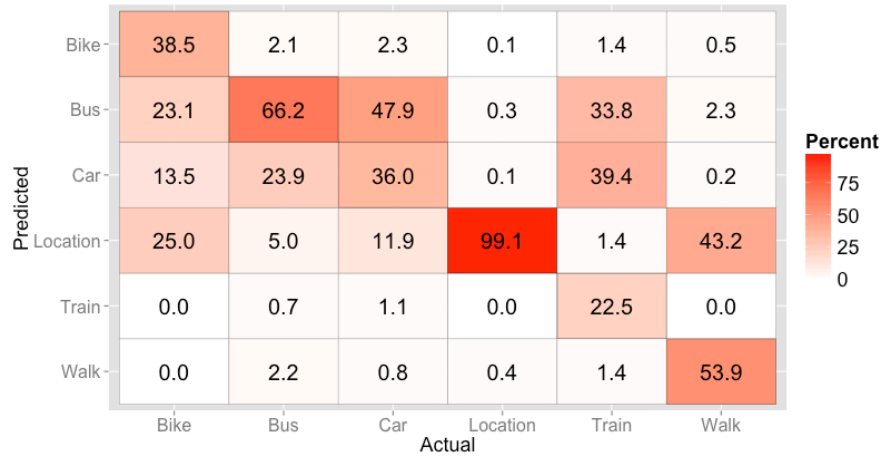
(b) Full model: With transit distance

The work of Zheng et al. (Zheng, Liu, Wang, & Xie, 2008b), suggests that the Decision Tree (DT) algorithm outperforms other machine learning algorithms in classifying travel mode. Consequently, we also build models using this algorithm in order to compare the results to those obtained with the RF algorithm. The same process was used for classifying move/non-move segments and for classifying modes. When applying the DT models to the testing set, the error rate of classifying modes is 31.29%, an increase of 6.94% compared to the error rate produced by the RF model (Table 3-9). Further, DT could not detect any bike or train trips in either the training set or the testing set.

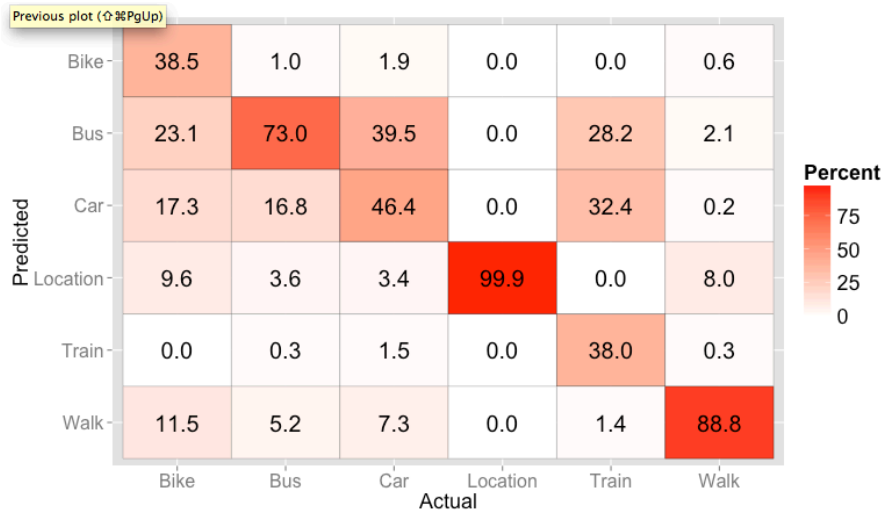
Table 3-9: Decision tree model - confusion matrix - testing set

Predicted\Actual	Bike	Bus	Car	Location	Train	Walk	Class error
Bike	0	17	23	5	0	7	1.0000
Bus	0	395	121	24	0	42	0.3213
Car	0	138	85	13	0	25	0.6743
Location	0	0	23	17420	0	1	0.0014
Train	0	37	33	0	0	1	1.0000
Walk	0	19	2	54	0	541	0.1218

Figure 3-9 presents the sensitivity of RF on the testing set. Sensitivity (recall or true-positive rate) is a measurement of the model's ability to correctly classify a segment as the correct mode or not. In other words, it refers to the model's ability to designate a segment with mode as positive. A highly sensitive model means that there are few false negative results, and thus fewer segments are misclassified. The diagonal cells show the ratio of correct positives predicted to actual number of positives. The figure shows that the full model has higher sensitivity than the baseline model for all classes. The sensitivity is high for both location and walking model (99.9% and 88.8%, respectively), fairly high for taking bus (73%), medium for car (46.4%), while low for biking and taking train (38.5% and 38%, respectively).



(a) Baseline mode: Without transit distance



(b) Full model: With transit distance

Figure 3-9: Sensitivity

Based on the above comparison, we conclude that including distance to transit line and transit stop increases the accuracy of the models. While the difference for the whole procedure is only 2.33%, the difference for detecting specific travel mode is higher. With distance to transit stop and transit line included, the models provides 35% lower error rate in detecting walking trips compared to the model without these factors. Therefore, we suggest that adding these two factors contributes a significant improvement in the model's accuracy.

A comparison of the accuracy between manual label data (base case) and that predicted by our full models also can be interpreted by the visual inspection of moving patterns for an example participant during one sample day (5762 data points), (Figure 3-11). We can observe that the list of travel mode for the two data sets is the same: bus, location, and walk. At this scale, the sequence of travel mode is very close. Comparisons made on other participants of course varied, but generally produced results similar to this particular case.

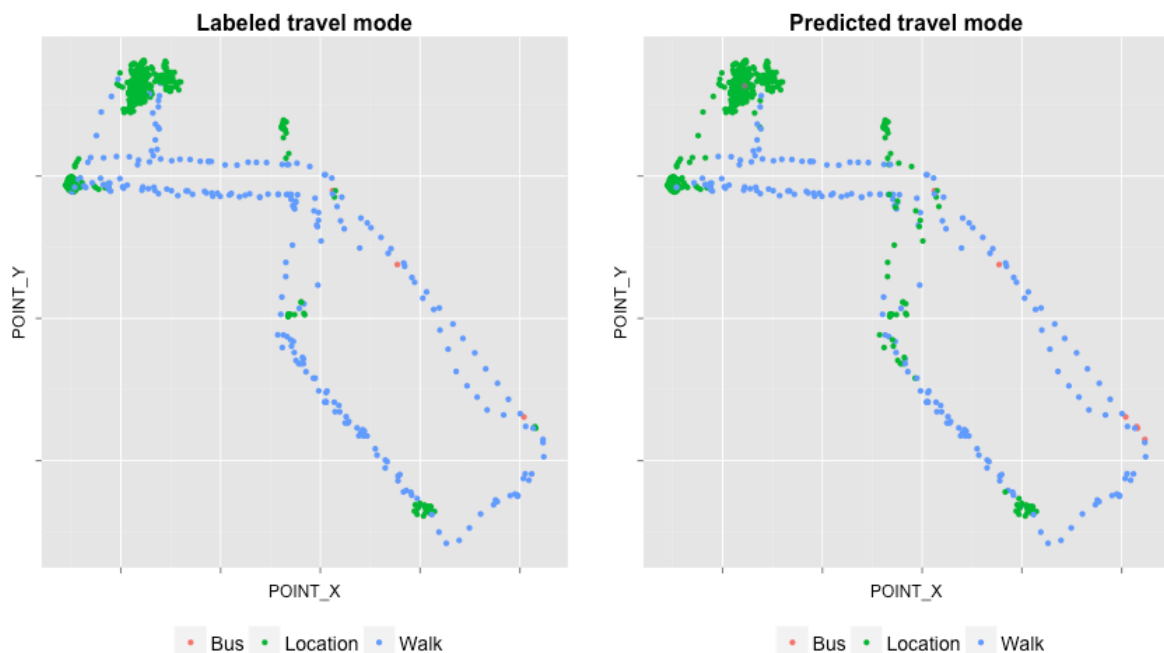


Figure 3-10: Sample movement patterns

3.5 Conclusion

Traditionally, travel data are collected by asking participants to describe their travel times, trip origins and destinations, travel modes and other different travel details during the surveyed period. Such traditional methods as face-to-face interview, telephone interview, mail,

Internet, or some combination of these, rely on the accuracy of self-reported information. It has been shown that data collected using these methods deviate systematically from actual behavior (Stopher, 2004), such as neglecting to report trips, especially short trips with increasing duration of the survey (Golob & Meurs, 1986) and non-home-based trips (Stopher, 1992) because of the burden of taking notes on all days. If shown to have efficacy, using GPS data to investigate people's travel behavior, such as travel mode and travel purpose, can be an important tool in transportation planning.

GPS technology has been used increasingly to supplement traditional travel surveys in public health and travel behavior studies. In practice, this technology provides many applications that benefit both individuals and society, especially with smart devices, such as personal assistance or providing navigation, urban planning, traffic forecasting, and event detection. However, GPS data do not directly expose key travel characteristics of a participant's movement behavior, such as travel mode. This study represents a complete and implementable framework using temporal and spatial patterns of human movements, while reducing manual work in processing data and recall from participants.

With the improvement of tracking devices and satellite signals, the volume of GPS data that needs to be processed for such purposes is getting bigger. In particular, there is need for automatic clustering/infering of travel mode—arguably, the most basic element of travel demand forecasting—from these data. Here, we propose a framework for automatic inference of the travel mode of people during their movement utilizing the PELT algorithm in conjunction with a RF algorithm. The method has not been previously applied to the analysis of GPS trajectories segmentation. The basic idea of our approach is to detect the stops and moves as a part of a trajectory which, solely based on speed data, identifying these two aspects of travel is

automatically decided by the data itself instead of by arbitrary thresholds of radius or duration of stops as in other research. Therefore, the framework requires no *a priori* assumptions on movement characteristics or geographic context and information as in previous research. PELT also produces a computational efficiency of $O(n)$ which is suitable for application to big data processing tasks (M.-H. Yang et al., 2014), tasks in which dealing with GPS data falls.

A RF algorithm—a famous machine learning algorithm with high prediction accuracy—is used to identify moving/non-moving segments, and travel mode in each moving segment. It is shown that the method is capable of using information on speed to predict travel mode from the GPS data. The classifying of travel mode model (car, bus, train, walk, and bike) has an accuracy of 75.65%, while the whole process reaches accuracy of 97.63% on the testing set. Any significant misclassification that is likely to occur is due to the similarity in movement patterns of travel mode such as car vs. bus. In this study, RF represented an increase of a 6.94% increase in accuracy over DT, a baseline algorithm.

The above results suggest that real-world travel mode recognition systems can rely on classification models that are trained on data sets. Although modes such as bus or walk may be accurately recognized, modes with data restricted by environment such as train (e.g. subway runs underground where signals could be inaccurate) still need improvement. In the future, we plan to incorporate features from the accelerometer to help disambiguate some of the train/car/bus confusions. Also, we will apply a data balancing technique and/or collect more data for these modes in order to increase the prediction accuracy. Additionally, constructing a vector of a sequence of movement using the same classifying techniques described, and then using that vector to learn the current mode for a segment, may provide a higher accuracy. Finally, with the improvement of GPS technology where inaccurate signals and time gaps are

reduced, we believe the models would produce greater accuracy. Although the procedure proposed in this study was developed to apply to one phase of the Expo study, it could be applied to any future GPS data sets from this study or from other sources. The framework could be an asset in both research and practice involving GPS data analysis.

CHAPTER 4: TRIP/ACTIVITY DETECTION AND CLASSIFICATION

4.1 Introduction

Recent studies have made advances towards building activity pattern detection algorithms based on GPS data. Comparisons of data from travel diaries and corresponding GPS monitoring in regional travel surveys reveal that the rate at which trips not reported in travel surveys are identified in GPS data varies substantially (Houston, Luong, & Boarnet, 2014). For example, the rate of under-reporting was 7% in a survey in Sydney (Stopher, FitzGerald, & Xu, 2007), 10% in a survey in Kansas City (Wolf, Bricka, & Ashby, 2004a), 18–40% in a survey of four California counties (California Department of Transportation, 2002), 30% in Ohio (Pierce, Casas, & Giaimo, 2003), and 81% in a survey in Laredo, Texas (Wolf, 2004).

Traditionally, researchers have simultaneously collected travel diaries and GPS tracking data in order to evaluate the accuracy of GPS derived trips/activities. A travel log format (such as the one used in the Expo Study) could help reduce participant reporting burden and provide sufficient seven-day travel data. In contrast to a travel diary format which requires trip-level details, the travel log is simpler, requires respondents to only count the total number, duration, and mode of trips for each day, and enables the collection of data for longer periods than with a typical one-day travel diary survey. Although our log format does not generate the type of trip-level information needed to calibrate the travel demand models used by transportation planning organizations, it does provide substantial day-level information which will be used in our future longitudinal evaluation of whether or not a major transit investment has succeeded in achieving its policy goals of encouraging reduced vehicle travel and increased transit usage and physical activity.

Person Name:

	Car Driver	Car Passenger	Motor-cycle/Scooter	Bus	Train	Bicycle		Walk		Other	Notes? Problems? Please describe below.
						# of Trips	Total Minutes	# of Trips	Total Minutes		
Monday											
Tuesday											
Wednesday											
Thursday											
Friday											
Saturday											
Sunday											

Instructions

- Count each trip you take during each day
- Include walk/bike trips over 5 minutes
- Count trips you take for recreation or exercise
- Log the total minutes you walk or bicycle each day
- Count each trip mode as a separate trip (car, walk, etc)

Suggestions

- Carry and complete the log as you travel
- Or you can complete the log at the end of each day
- Note any problems each day (forgot to fill out one day)
- See the back of this log for examples

Figure 4-1: Expo Study trip log format

To get an accurate estimate of trip counts using GPS data, the ending point for each trip needs to be identified within GPS data streams. A common approach to detect these points is using dwell time. Dwell time is the minimum amount of time for a valid stop between two trips. In previous studies, there are several variety of dwell times, depends on GPS data quality, for example, 45 seconds was found to be appropriate threshold for the first household travel survey conducted with a GPS in Austin, Texas in 1997. Other research used a 3-minute activity duration threshold and 60-second trip duration threshold for Borlange GPS data. Wolf (2001) found that anything less than 120 seconds tended to find many false trip ends that were attributed to heavy congestion or signal delays experienced within the Atlanta region. A value of 120 seconds was also used in studies in California (Schönfelder & Axhausen, 2010; Wolf et

al., 2001b), Australia , and Belgium (Beckx, Panis, Janssens, & Wets, 2010). The greatest value of time for dwell time is that found in Doherty et.al ’s research (Doherty, Noel, Gosselin, Sirois, & Ueno, 2001), 3 minutes. However, the required dwell time for such determination is expected to decrease over time because of the improvement in GPS technology that helps detect movement more accurate.

In this chapter, we assess which value of dwell time is suitable for the Expo study’s GPS data. We use data from wave 1 and wave 2 to compare dwell times from 30 seconds to 180 seconds. The process is shown in Figure 4-1.

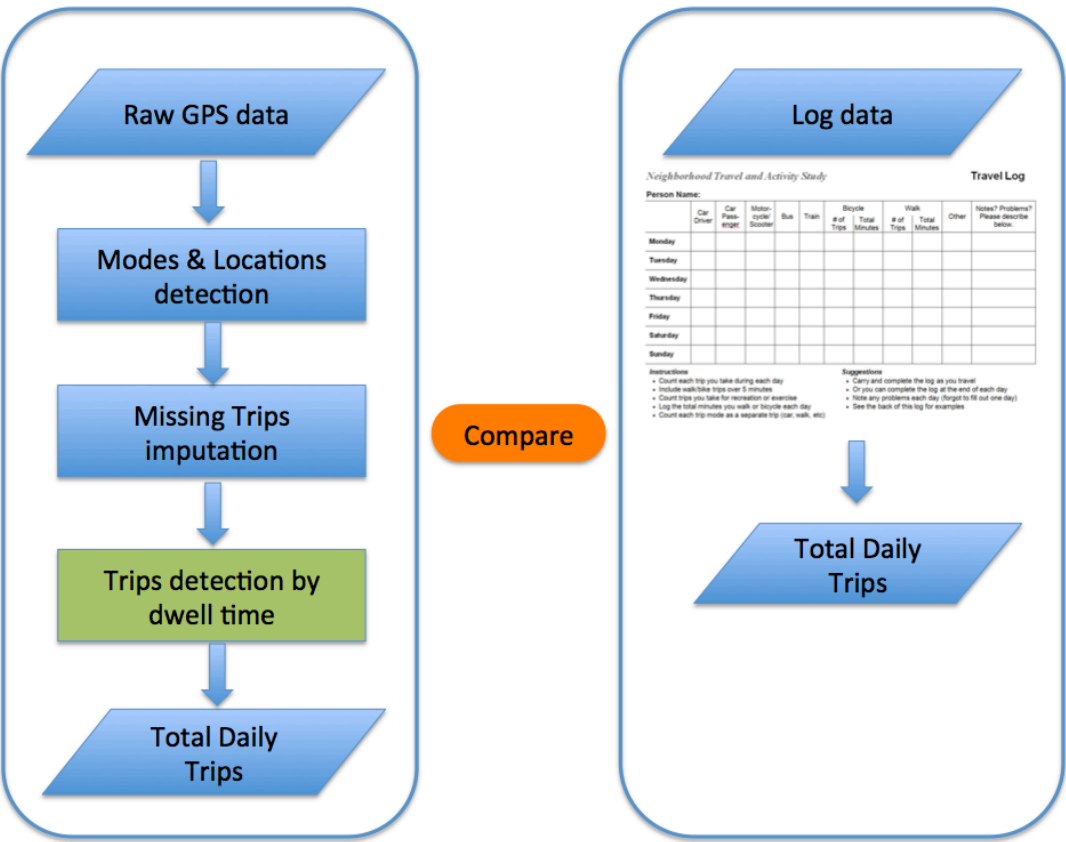


Figure 4-2: Trips detection process

4.2 Data Preparation

In this chapter, we use hand-coded GPS data from EXPO study. We developed a set of diagnostic thresholds to assess periods of missing data in GPS data and to identify days with sufficient information to be included in our analysis of discrepancies in day-level trip counts and walking travel duration. Periods of missing data could occur due to participant noncompliance, device error, or to satellite signal obstructions during indoor periods within a steel and concrete structure, underground periods on the subway, or outdoor periods with obstructions such as tall buildings, vegetation, or geography. We classified days with periods of missing data based on the length of the gaps in the GPS data and in terms of whether the location or travel status was the same or different before and after the gap.

4.2.1 Data selection

First, we included data for only the 119 participants who provided complete responses to the demographic information used in the analysis from the baseline survey (age, gender, race, educational attainment, household income, number of household cars, and employment status), which corresponded with 1,125 days of GPS data. Next, we eliminated 83 of these days from the analysis since they included a gap lasting more than 30 minutes which had a different location/travel status after the gap compared to the status before the gap (Table 4-1)—a pattern that suggests that GPS device may have missed at least one change in travel mode or location. However, we kept days with a gap lasting over 30 minutes which started and ended at the same location under the assumption that no trips or activities were unobserved during the period of missing data. We removed an additional 130 days with gaps lasting more than two

hours and in which the ending time of lost signal period was after 10:00 am, under our assumption that such a large daytime gap may have included unobserved activities. We also eliminated 32 days in which the participant did not leave the original location but the recorded duration was less than 12 hours or ended by noon (12:00 pm), and 3 days in which a participant flew on an airplane. The final dataset used for analysis included daily travel log and GPS data for 892 days, representing the patterns for 116 participants.

Table 4-1: Periods of signal loss (N = 1,125 days, 119 participants)

Criteria	Days	HH	Average daily lost period	Days in which changed status after lost signal ^a		Days in which same status after lost signal ^a	
				Number	%	Number	%
30-60 Minutes	82	53	1.56	23	28.05	63	76.83
61-120 Minutes	102	57	1.11	35	34.31	69	67.65
> 121 Minutes	202	70	1.14	73	36.14	142	70.30

^a A given day could have periods in more than one gap category.

4.2.2 Data imputation

In this part of the analysis, a set of rules is applied to impute the missing coding status for each GPS point and the missing link between locations.

Impute no-recorded status

- If the missing status duration is less than 10 minutes and the status after that is the walking, convert the period to the walking period.
- If the duration is more than 10 minutes and the status after that is the walking, convert the period to the status before missing period.
- If the status after missing period is the not walking, convert period to the status before missing status.

Missing mode between a location and a bus stop

If the trip link is Location-Bus stop and the time participant stopped at the location is less than 10 minutes, change Location to Bus stop.

Link criteria

- If a link contains Bus-Car-Bus or Car-Bus-Car with no stop between any two modes, remove the link
- Delete car trip if before this trip is a bus stop and after that is bus trip
- Remove all transfer transit trip to follow the counting logic in travel log.

For example, if a participant rode bus and then transferred to a second bus, these two single bus trips are counted as one bus trip.

4.3 Discussion of results

We analyzed the implications of using four different dwell time criteria to identify trips in the GPS data: 30, 60, 120, and 180 seconds for a sample of 892 days from wave 1 and 2 in the Expo study data. As expected, the aggregate number of trips identified in the GPS data was lower with longer dwell times, and this pattern was most pronounced for vehicle trips which, based on a 30-second dwell time, had 27% more aggregate GPS-derived vehicle trips compared to log-derived vehicle trips, and based on a 180-second dwell time had 2% fewer aggregate GPS-derived vehicle trips compared to log-derived vehicle trips using a 180-second dwell time.

Table 4-2: Aggregate GPS trip counts by dwell time and mode (N= 892 days, 116 participants)

Trip Mode	GPS Trip Counts by Dwell Time				
	Log Trip Counts	30 sec	60 sec	120 sec	180 sec
		Total	Total (% Diff.)	Total (% Diff.)	Total (% Diff.)
Total	4,522	5,068 (12.1%)	4,870 (7.7%)	4,532 (0.2%)	4,124 (-8.8%)
Vehicle	3,149	3,996 (26.9%)	3,797 (20.6%)	3,466 (10.1%)	3,074 (-2.4%)
Transit	429	290 (-32.4%)	290 (-32.4%)	286 (-33.3%)	285 (-33.6%)
Walking (>5 min.)	944	782 (-17.2%)	782 (-17.2%)	780 (-17.4%)	765 (-19.0%)

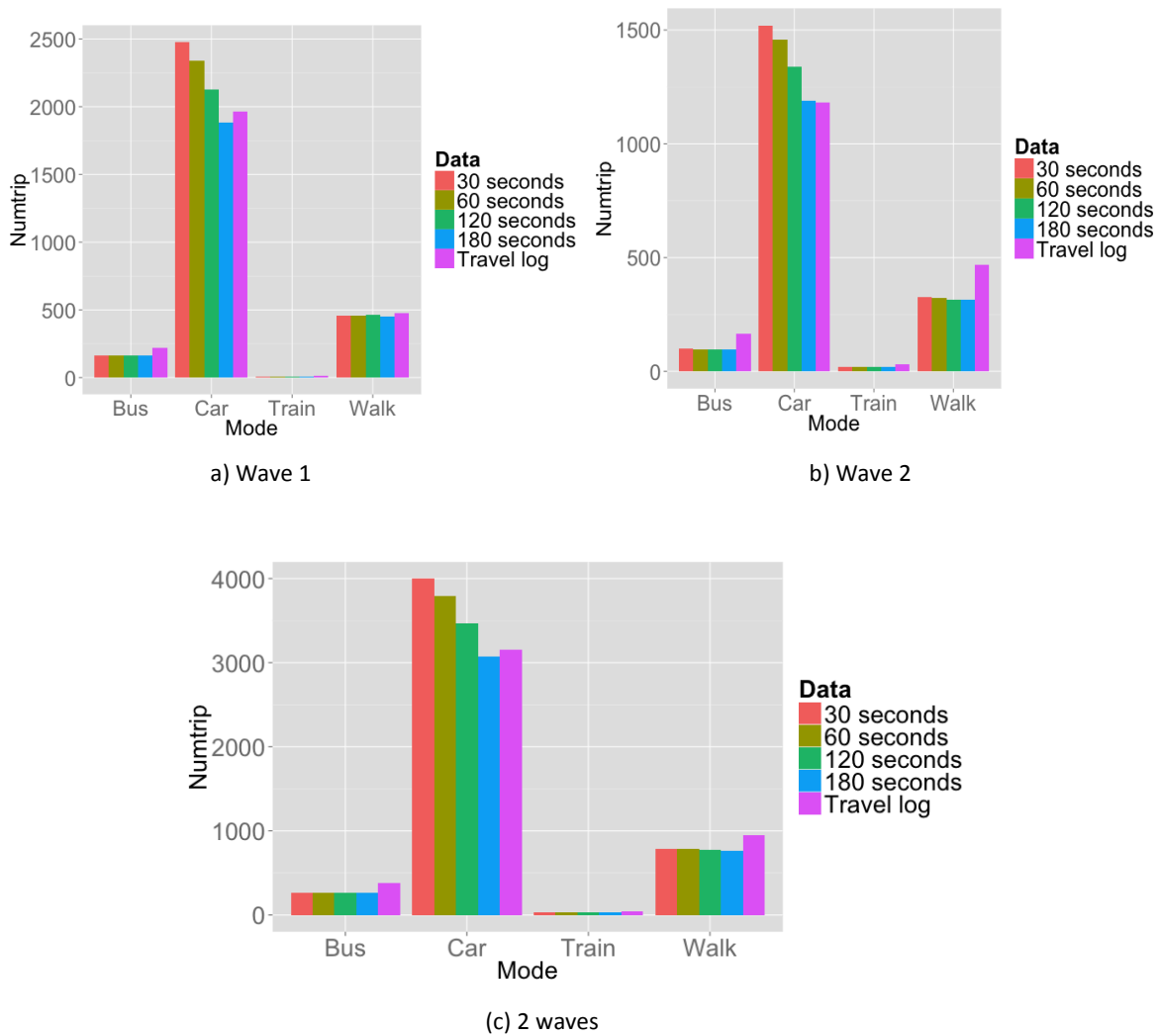


Figure 4-3: Compare dwell time results

We found substantial agreement between daily travel patterns derived from GPS data and those from our trip log instrument (Table 4-3). Although significant in most cases, the magnitude of the difference in mean total trips and mean trips by mode between the two measures was relatively small. For example, the difference for total daily trips ranged from 0.61 with a 30-second dwell-time threshold to -0.45 with a 180-second threshold. Likewise, the rate of good trip reporting (± 1 trip) for all trips ranged from 46% with 30-second threshold to 48% with a 180-second dwell time. The average discrepancy in daily vehicle trips ranged from 1.14 to -0.10, and the rate of good vehicle-trip reporting ranged from 45% to 50%.

On average, the logs had more daily transit and walking trips reported than were identified from GPS data, but these modes tended to have higher rates of good reporting (51% to 54%) compared to vehicle trips. Participant logs reported on average about 10–12 more minutes of walking than GPS-derived data, and only about a third of observation days agreed on daily walking duration (± 10 minutes). We found substantial variation in the difference in daily walking duration between GPS-based and participant-reported estimates (standard deviation = 43 minutes). Although we are not aware of comparable published estimates of daily walking duration, a meta-analysis of previous studies examining trip duration across various travel modes found that participants tended to self-report between 2.2–13.5 more minutes per trip than were identified through GPS tracking (Kelly, Krenn, Titze, Stopher, & Foster, 2013). Such discrepancies may occur due to participants rounding travel times to the nearest 5- or 15-minute clock time or because participant estimates of trip duration vary depending on the perceived difficulty of a given trip (Stopher et al., 2007).

Table 4-3: Mean daily discrepancies between GPS-derived measures and self-reported trip counts

			GPS-derived measures		Log-derived measures		Difference (GPS-log)			Good reporting ^c
Dwell time & mode	HID	Days	Mean	SD	Mean	SD	Mean	Sig. ^b	SD	Mean
a. Dwell time 30 seconds										
All trips	116	892	5.68	4.51	5.07	3.80	0.61	*	4.1	0.46
Vehicle trips ^a	113	740	5.40	3.94	4.26	3.24	1.14	*	3.7	0.45
Transit trips ^a	39	142	2.04	1.92	3.02	2.32	-0.98	*	2.6	0.54
Walking trips ^a	93	450	1.74	2.32	2.10	1.93	-0.36	*	2.7	0.52
Walk minutes ^a	93	450	21.47	33.26	33.38	36.42	-11.91	*	43.1	0.31
b. Dwell time 60 seconds										
All trips	116	892	5.46	4.20	5.07	3.80	0.39	*	3.75	0.48
Vehicle trips ^a	112	735	5.17	3.57	4.28	3.23	0.88	*	3.25	0.46
Transit trips ^a	39	142	2.04	1.91	3.02	2.32	-0.98	*	2.62	0.53
Walking trips ^a	93	453	1.73	2.30	2.08	1.93	-0.36	*	2.66	0.51
Walk minutes ^a	93	453	21.35	33.33	33.15	36.40	-11.80	*	43.09	0.32
c. Dwell time 120 seconds										
All trips	116	892	5.08	3.86	5.07	3.80	0.01		3.53	0.48
Vehicle trips ^a	111	732	4.73	3.24	4.30	3.23	0.43	*	3.02	0.48
Transit trips ^a	39	142	2.01	1.89	3.02	2.32	-1.01	*	2.60	0.53
Walking trips ^a	93	457	1.71	2.22	2.07	1.93	-0.36	*	2.58	0.53
Walk minutes ^a	93	457	22.15	33.98	32.86	36.37	-10.71	*	43.54	0.32
d. Dwell time 180 seconds										
All trips	116	892	4.62	3.51	5.07	3.80	-0.45	*	3.37	0.48
Vehicle trips ^a	111	731	4.21	2.88	4.31	3.22	-0.10		2.86	0.50
Transit trips ^a	39	142	2.01	1.89	3.02	2.32	-1.01	*	2.59	0.53
Walking trips ^a	92	457	1.67	2.13	2.07	1.93	-0.39	*	2.51	0.53
Walk minutes ^a	92	457	22.84	34.90	32.86	36.37	-10.02	*	44.49	0.31

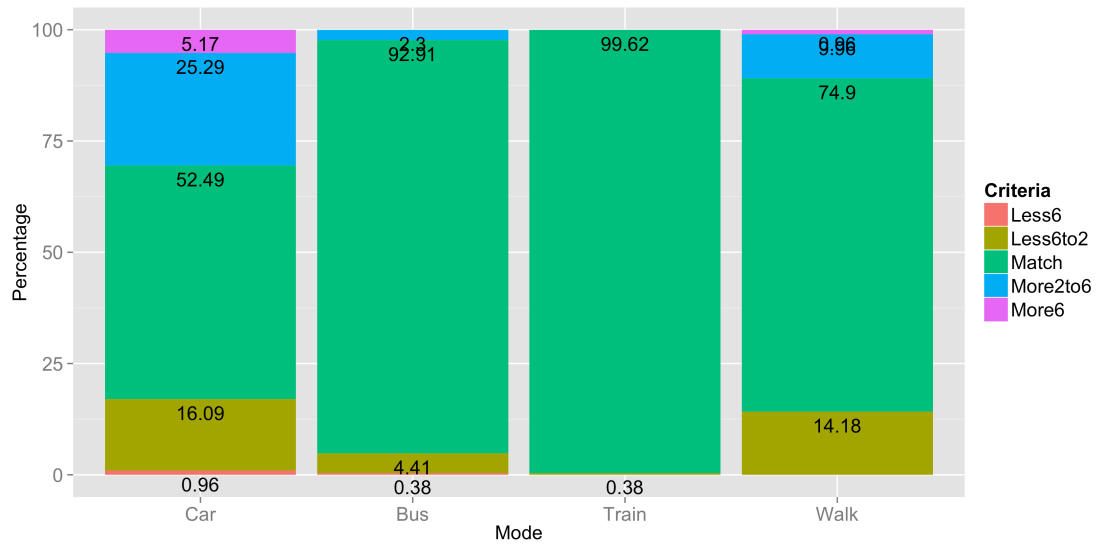
^a Results by mode are calculated including only days with at least one trip taken by that mode.

^b * Indicates the difference between the mean GPS-derived measure and the log-derived measure is significant (unpaired t-test, $P < 0.05$).

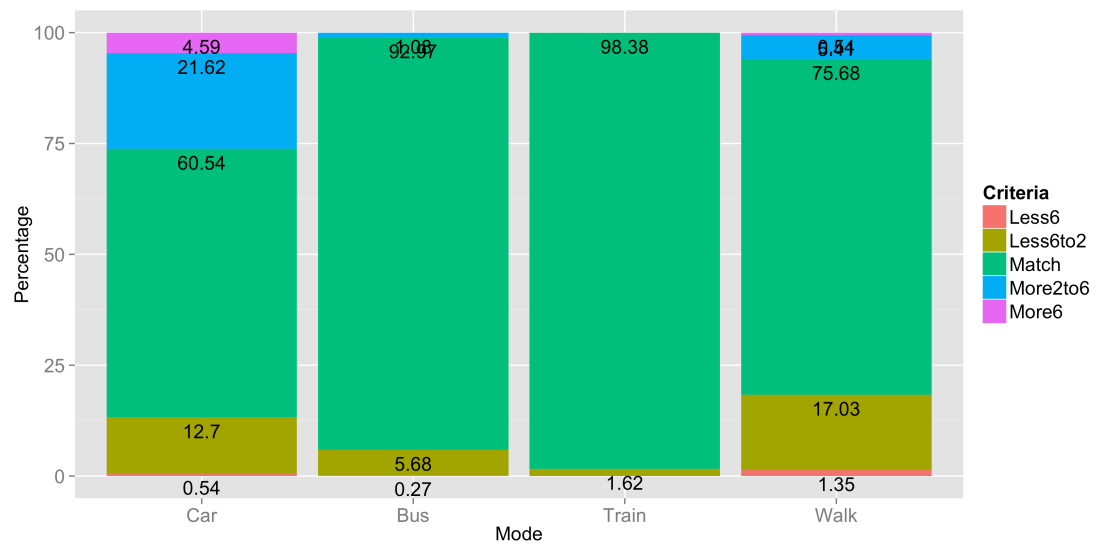
^c Indicates the daily difference between GPS- and log-derived measures was within ± 1 trip or ± 10 minutes.

Given that we found substantial agreement between estimates of reporting discrepancy across dwell-time thresholds used—7.7% difference in the total number of trip when we use 60-second threshold compared to 0.2% when we use 120-second threshold—the remainder of our analysis uses the travel patterns identified with the 60-second threshold. Figure 4-4 shows the match in the number of trip between GPS-derived count and log count. The difference is divided into ranges: fewer than 6 trips, from -6 to -2, match (from -1 to 1), from 2 to 6, and

more than 6 trips. Using these intervals, train trips have the best match between GPS and log count (more than 92% for both wave 1 and wave 2).



a) Wave 1



b) Wave 2

Figure 4-4: Compare number of GPS-derived trips with self-report trips
(dwell time = 60 seconds)

CHAPTER 5: INFERRING TRIP PURPOSE

5.1 Introduction

Travel diary data are a crucial resource for transportation and land use planners, particularly in calibrating travel simulation modeling processes. Travel demand models rely on the quality and reliability of the data used in their estimation. Depending on the specific model being estimated, these data requirements can apply to the purpose, time, distance, mode, and route choice of each trip as well as the sequencing of trips within the assigned travel day or travel period. Collecting these data via travel diaries, as is commonly the case, is a complex and expensive process.

Traditional travel diaries are collected by asking participants to describe their travel times, trip origins and destinations, travel modes, and other travel details during the surveyed periods. There are a number of methods for conducting travel diaries, such as: face-to-face interview, telephone interview, mail, Internet, or some combination of these. The validity and usefulness of data collected by these methods rely heavily on the accuracy of self-report information. However, it has been shown that data collected using these methods deviate systematically from actual behavior (Stopher, 2004). Participants may neglect to report trips—especially short trips—with increasing duration of the survey (Golob & Meurs, 1986), and non-home-based trips (Stopher, 1992) because of the burden of taking notes all days. Frequently, surveys are filled out at the end of each day or survey period, increasing the probability of forgetting partial or whole trips. Participants may also not want to report their detailed activities because of privacy concerns, or the trips that are considered unimportant (Wolf et al., 2001b). Still another reason may be participants' misunderstanding travel survey instructions.

To overcome this problem, over the past decade, the number of studies using Global Position System (GPS) for collecting individual humans' daily activities has increased. Recent developments in the technology have provided more accurate and better recording, especially in areas where high density of buildings and trees, under urban canyons, and in trains or buses where satellite reception is bad. Although GPS devices can collect the latitude, longitude, and the time of each location point, they do not directly provide trip purpose. Additional processing is needed to impute this information from GPS data. In most of the existing studies, these tasks have been done either manually or/and with some level of automation. Even so, researchers often still need to validate, correct, or add missing trips through an Internet recall survey, mailing, or telephone, which could be an additional burden for the participants, with the results relying on the participants' cooperation and memory. Few would argue that there is not a need for the improvement of algorithms to automatically derive this travel-survey-like information.

In this paper, we introduce the development and evaluation of models for deriving travel information from continuous GPS tracking data, focusing on trip purpose classification and imputation. We apply several machine learning algorithms to identify trip purpose through mining GPS tracking, land use and demographics characteristics of respondents. The study expands the existing literature on the use of GPS/GIS technology in an entirely passive and automatic way for its use in activity-travel behavior studies.

5.2 Literature review

According to Feng et al. (Feng, Moiseeva, & Timmermans, 2011), reported accuracy for trip purpose imputation from GPS data is rarely much higher than 70%—and more commonly

around 40%. The main reason is that dissimilar trip purposes may share similar characteristics, such as trip duration, land use type of the location, frequency of visit, etc. (Feng et al., 2011). Hence, the following issues should be considered when imputing trip purpose (McGowen & McNally, 2007):

- The model needs to deal well with correlated “independent” variables.
- The model needs to deal with non-linear relationships.
- The model needs to be a predictive model, not just descriptive.
- The model’s dependent variable is categorical.
- The model needs to be able to handle a large number of variables that include a mix of both categorical and continuous variables.

Several imputation techniques have been applied in the literature. they can be divided into two main approaches: 1) rule-based approach, and 2) statistical approach.

In the first approach, Wolf et al. (Wolf, Hoogendoorn, & Bachman, 2001a) used the trip purposes from the 1990 Atlanta regional household survey as criteria to impute trip purpose. Each land use description is associated with a primary, secondary and tertiary trip purpose, with some trip purposes further categorized by both time-of-day and duration of stay for the same land use code. In other research, Wolf et al. (Wolf, SchöUnfelder, Samaga, Oliveira, & Axhausen, 2004b) used GIS data to allocate trip purpose. The authors assigned the purpose “home” to trips ending within 200 m from the home location. Other trip purposes were assigned by matching the GPS data with available points of interest and land-use polygons as the primary means. The research also matched socio-demographic and trips/activities data with the National Household Travel Survey (NHTS) to define the most probable purpose for the other

trips. However, these studies were limited to GPS data loggers which were placed in the cars, missing both walk and public transit trips. In other research (Bohte & Maat, 2009), all of the potential trip purposes were classified into 13 categories. Locations of facilities were derived from GIS maps, listing points of interest by the x and y coordinates of the facilities. Once a respondent has assigned—using an Internet application—a trip purpose to an unknown trip purpose, this trip purpose is assigned to all following trips that end within 50 m of the same location. The work place and home address was known beforehand and trips ending within 100 m (instead of 50 m) were assigned to these two categories. Then, the study implemented seven IF-ELSE rules, which classify the purpose. For example, if the endpoint trip was within a shopping center polygon then trip purpose was “shop”.

In a series of research efforts focused on using GPS data, Stopher et al. (Stopher, FitzGerald, & Zhang, 2008b) used individual characteristics (household composition, possession of travel modes, home and work addresses) as input for rule-based algorithms that derive trip characteristics. However, the algorithms required pre-defined address of the participants' workplace or school, the two most frequently used grocery stores, and occupation. In addition to these addresses, Stopher et al. (Stopher, Clifford, Zhang, & FitzGerald, 2008a) divided trips into private-level destinations and public-level destinations. Activity duration, frequency of visiting at a location in the survey period, land use map, and bicycle ownership/availability are required for the process. Then, the trip purpose is deduced by heuristic rules, given the trip end's distance to the point of interest (200 m was used as buffering zone) and the respondents' occupation information.

In a more recent study, National Household Travel Survey (NHTS) data were used as source data to obtain basic information that can be applied to infer the trip purpose (Shen &

Stopher, 2013). The basic travel information includes the distribution of the activity duration, the distribution of the time when the activity occurs, and tour information. The study also used a series of special rules, including: a rule for activity duration of work/education trips, a rule for the time when work/education trips occur, and also a tour-based trip chain. However, the basis of activity duration, the time when activities occur and the distribution of tour types were derived from NHTS data, which is subject to the noted problems of self-reported surveys. The accuracy of trip purpose imputation for the processed GPS data was 58.7%.

Although the rule-based approach has been applied in travel data collection for much of the past decade, a method that can assign a probability to each purpose for each trip has been recommended to be better than a deterministic method (Shen & Stopher, 2013). Chen et al. (C. Chen, Gong, Lawson, & Bialostozky, 2010) applied a probabilistic model to evaluate how various factors determine the purpose of a trip probabilistically for those trips whose purposes cannot be deterministically determined. Three types of factors—time of day, history dependence, and land use characteristics—were considered as inputs for two multinomial logit (MNL) models. To calculate the probability that a trip is serving a particular purpose for home-based trips and non-home-based trips, four trip purposes were considered: work/school related, personal business, shopping, and social recreation. In both models, it was assumed that if a specific trip purpose were selected from the identified trip origins and destinations, then the participant derived a higher utility from conducting the trip for this purpose than for any of the other purposes. Time-of-day factors were found to be important to the prediction, which point to the regularity of human behavior patterns.

The imputation of trip purpose is considered more difficult and requiring more additional data than that of trip detection. According to Feng et al. (Feng et al., 2011), reported

accuracy for this task is rarely much higher than 70% and more commonly around 40%. The main reason is that the dissimilar trip purposes may share similar characteristics such as trip duration, land use type of the location, frequency of visit, etc. (Feng et al., 2011).

McGowen and McNally (McGowen & McNally, 2007) tested the feasibility of a model that predicts activity types based solely on GPS data, land use data, and demographic data using discriminant analysis and classification/regression trees that can handle non-linear relationships and correlated variables. Quick Unbiased Efficient Statistical Tree (QUEST) (Loh & Shih, 1997) was used for developing classification trees. The “unbiased” refers to the fact that other trees—including Classification and Regression Tree (CART) and Chi Squared Automatic iteration detection (CHAID)—are biased toward choosing variables with more categories first (McGowen & McNally, 2007). A set of variables were obtained from the reported activity times and geo-coded locations, such as: activity start time, activity duration, total duration of all out of home activities, distance from home (straight line), chaining = total out of home activities / total home based tours, etc. The models predicted in which of 26 different activity types the individual participated. Accuracy for the best model was 63% by using discriminant analysis for out-of-home activities (62% by using classification tree), 79% when including the “at home” activity, and 72% considering that GPS data may miss as much as 10% of trips.

Recently, Deng and Ji (Deng & Ji, 2013) formulated a machine learning procedure to incorporate four different types of information into a multiple decision tree-based classification. These information types included: GPS time stamps, land use, temporal behavior, and socioeconomic/demographic characteristics. With the use of the adaptive boosting strategy, the trip purpose identification task achieved a rather high correctness of 87.6%, a significant error reduction of 7.1% from classification with no boosting. The experimental results lead to findings

that even when part of information is missing—such as some sample records with no socioeconomic data—the learning and classification procedure still produces reasonably good results.

Lu et al (Lu, Zhu, & Zhang, 2012) explored the feasibility of using machine learning method to automatically derive trip purpose based on in-vehicle GPS data collected by University of Minnesota. Post-pruning and on-line pruning—a technique that reduces the size of the tree by cutting off some nodes from the tree—were applied to the trip purpose decision tree in the research. Once the decision tree is constructed and pruned, a 10-fold cross-validation is employed to estimate the error rate of the machine learning technique. The input variables used to derive trip purpose can be divided into three categories (GPS trip attributes, respondent's characteristics, and land use data) and five subcategories (trip attributes, next & previous trip attributes, respondent's social demographics, parcel-level land use type, and trip end location). Specifically, the trip purposes are coded into 10 categories (home, work, shopping, daycare, dining, driving others, services, school, social/recreation, and other) and decoded into five categories including Home-base Work, Home-base Shopping, Home-base Social/Recreation, Other home-base, and Not Home-base based trips. A classification accuracy of 62.77% was reported. Later, the authors (Lu & Zhang, 2014) applied the bagging method to National Household Travel Survey (NHTS) 2009 add-on data, with an overall accuracy of 77.7% with 10-fold cross-validation for 9 purposes. Random Forest (RF) also was used in a recent paper to predict trip purposes (Montini, Rieser-Schüssler, Horni, & Axhausen, 2014). The classifier performed well with overall mean accuracy above 80%, which is relatively high.

In summary, because the accuracy in classifying and predicting trip purpose is rarely higher than 70%, and such accuracy still strongly reliant on the prior knowledge from

participants, there is a need to improve these methods and to take better advantage of such available data sets as NHTS or other GPS data sets with fully recorded trip purposes.

5.3 Methodology

The “No Free Lunch” theorem (Wolpert & Macready, 1997) states that there is no one model that works best for every problem, so it is common in machine learning—particularly in supervised learning—to try multiple models and find the one that works best for a particular problem. Following this paradigm, in this study, we compare several machine learning algorithms for trip purpose classification modeling. Our goal is to run machine learning algorithms to build and analyze the purposes of trips identified in GPS data. The notion of 10-fold Cross Validation will be used to choose the best structure for each algorithm. The 10-fold cross-validation procedure randomly partitions a set of data with known correct outputs into 10 folds, and then utilizes one fold as a validation data set while the remaining N-1 folds are used as the training data set. This process is repeated 10 times with a different fold used as the testing data for each round. The best model’s structure is determined as the one with smallest average error rate after 10 rounds. Using this final structure and training data, we build the final model for each algorithm. Then, we use the testing set to assess these final models, compare their performance, and select the best algorithm for trip purpose prediction. The overall framework for the process is shown in Figure 5-1.

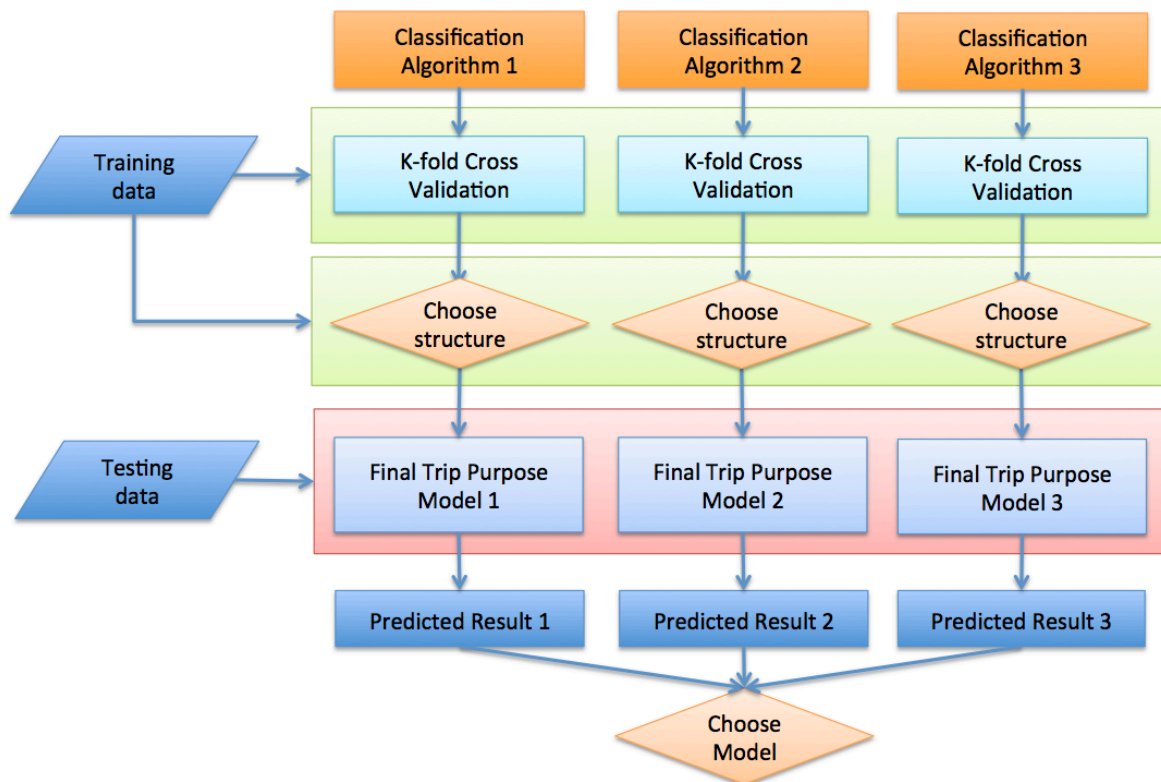


Figure 5-1: Process to predict trip purpose

Random forest (RF)—a machine learning technique that was used for predicting transportation mode in Chapter 3—is also used here for modeling trip purpose based on the BH data set. We also compare the results to the results obtained with the generative and discriminative methods. The first type of model learns each trip purpose’s characteristics, and then classifies the input using the knowledge gained. Typical models of this type are Naïve Bayes, Hidden Markov Models, Bayesian networks, etc. Alternatively, the second model determines the difference in the trip purpose models without learning their characteristics and then classifying the input. Logistic regression, Support Vector Machines, neural networks are

popular discriminative models. In this study, we use Naïve Bayes and Multinomial Logistic regression.

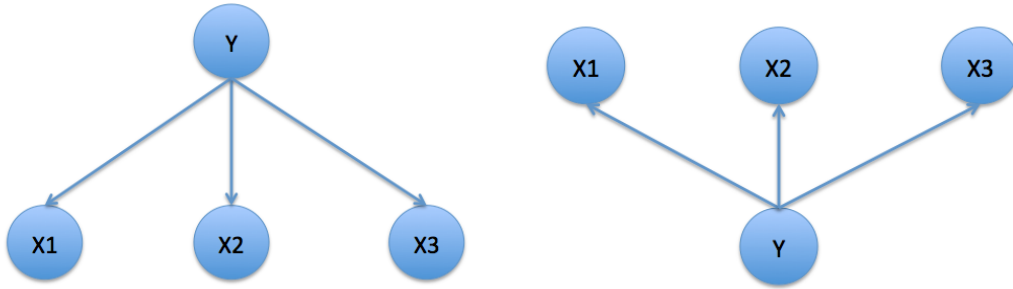


Figure 5-2: Generative method v.s discriminative method

5.3.1 Naïve Bayes model

The Bayesian Learning approach to the classification problem estimates the posterior probability that an input belongs to a class, given the observed attribute values for the input based on the Bayes Rule:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_y P(X|Y = y)}$$

where Y is response vector and X is predictors. The **a priori** probability of Y or $P(Y)$ is the probability of an event **before** the evidence is observed. The **a posteriori** probability of Y or $P(Y|X)$ is the probability of an event **after** the evidence is observed.

Naïve Bayes (NB) (Good, 1967) is the simplest (“naïve”) model of Bayesian Learning Algorithms; the class with the highest estimated posterior probability will be selected. In the NB model, the oversimplifying assumption is made that the inputs are independent given the label. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive. The input features are represented as a random vector $X = [X_1 X_2 \dots X_n]^T$. Each element of X is a discrete random variable. When discussing a particular element of X , we will address it with its subscript index.

$$P(X_1, X_2, \dots, X_n | Y) = P(X_1 | Y) P(X_2 | Y) \dots P(X_n | Y)$$

This also is the biggest disadvantage of NB algorithm. Under this assumption, the above equation becomes:

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^n P(X_i|Y)}{\sum_y \prod_{i=1}^n P(X_i|Y)}$$

The process of classification is:

$$y_{MAP} = \underset{y_j \in Y}{\operatorname{argmax}} P(y_j | X)$$

$$= \underset{y_j \in Y}{\operatorname{argmax}} \frac{P(X|y_j)P(y_j)}{P(X)}$$

$$= \operatorname{argmax}_{y_j \in Y} P(X|y_j)P(y_j)$$

where MAP is the maximum **a posteriori**, meaning most likely purpose. The model applies maximum likelihood estimates for a data size N . For categorical attribute, the frequencies in the data are used:

$$\hat{P}(y_j) = \frac{N(Y = y_j)}{N}$$

$$\hat{P}(x_i|y_j) = \frac{N(X = x_i, Y = y_j)}{N(Y = y_j)}$$

If an attribute is continuous, we do not have the counts of the number of times that that attribute occurred in the training dataset; we only have a range of values. $\hat{P}(x_i|y_j)$ is estimated through a Gaussian density function with a mean μ and standard deviation σ :

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\hat{P}(x_i|y_j) = g(x_i, \mu_{Y=y_j}, \sigma_{Y=y_j})$$

5.3.2 Multinomial Logistic Regression

Multinomial Logistic Regression (MNL) is a classical categorical regression model. The model generalizes logistic regressions for classification problems where the class label can take on more than two possible values. Using the same notation as for Naïve Bayes models, logistic regressions assume a parametric form for the distribution $P(Y|X)$, then directly estimate its parameters from the training data. Logistic regression models do not assume homoscedasticity between the dependent and independent variables.

The total response probability is equal to 1. When one takes N independent observations based on these probabilities, the probability distribution for the number of outcomes that occur in each of the m class is multinomial (Agresti, 2007).

$$\pi_j = P(Y = y_j) = \frac{e^{\beta_j X}}{\sum_{k=1}^m e^{\beta_k X}}, \quad j = 1, \dots, m$$

where β_j is a vector of coefficients, estimated by Maximum Likelihood.

The multinomial logit model also is written in terms of the log-odds. The logit models pair each category with a baseline category, normally the last class of Y (Agresti, 2007). The choice of the baseline category is arbitrary. The baseline-category logits are:

$$\log\left(\frac{\pi_j}{\pi_m}\right), \quad j = 1, \dots, m - 1$$

There are $m - 1$ models for m classes to show the effects of the explanatory variables on producing outcome j rather than the baseline outcome m . We fit each baseline-category logit model with predictors following a regression form:

$$\log\left(\frac{\pi_j}{\pi_m}\right) = \beta_j X, \quad j = 1, \dots, m - 1$$

For $j = m$

$$P(Y = y_m | X) = \frac{1}{1 + \sum_k^{m-1} e^{\beta_k X}}$$

For $j \neq m$

$$P(Y = y_j | X) = \frac{e^{\beta_j X}}{1 + \sum_k^{m-1} e^{\beta_k X}}$$

To summarize, although both NB and MNL can model multiple classes response, MNL directly estimates the parameters of $P(Y|X)$, while NB directly estimates parameters for $P(Y)$ and $P(X/Y)$. In practice, the NB classifier is very efficient since it is less computationally intensive (in both CPU and memory) and requires a small amount of training data. The model is characterized by the independence of irrelevant alternatives (IIA), a problem that results from the MNL assumption that the relative probability of choosing between two alternatives is unaffected by the presence of additional alternatives. Although the independence assumption is a limitation of the method, it often does not have an impact on reality. This method has two advantages over many other classifiers: easy to construct and efficient due to its assumptions (Cheng & Greiner, 1999).

5.3.3 List of predictors:

This section introduces a list of variables will be used for trip purpose prediction.

Land use type:

Land use type of the location where the activity happened. Land use type includes: service (SRV), school, (SCH), food (FOD), industrial (IND), residential (RES), office (OFF), retailer (RET), and recreational (REC).

In order to identify land use information for each location, we first collect all points that have been identified as a location in the previous chapter, and then apply the follow process to find the centroid point for each location:

- If the number of points in a location is large (more than 200), the centroid point was the point with highest frequency.
- If the number of points in a location is small (less than 20), the centroid point is the center point (average distance in each dimension).
- Otherwise, Kernel density estimation was used to find the centroid point.

The principal reason for using this particular method is because Kernel density estimation (KDE) is a popular tool for visualizing the distribution of data (Duong, 2007). First, a kernel is a special type of probability density function (PDF) with the added property that it must be even. KDE is a non-parametric method of estimating the PDF of a continuous random variable. It is non-parametric because it does not assume any underlying distribution for the variable. Essentially, at every datum, a kernel function is created with the datum at its center –

this ensures that the kernel is symmetric about the datum. In other words, KDE can find the high-density region in space or find the centroid point of a region.

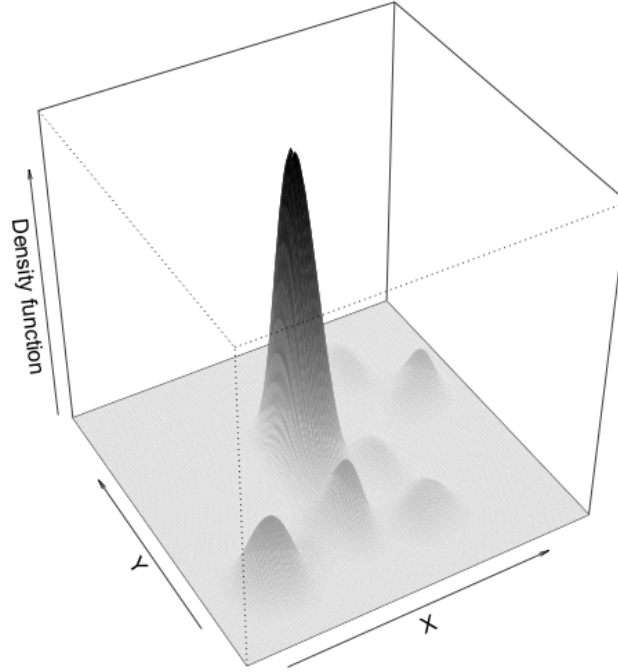


Figure 5-3: Example of Kernel density estimation

Two steps are used to estimate Kernel density: choosing the shape of the kernel, and identifying its bandwidth. The Kernel density estimate is defined as

$$\hat{f}(x, H) = n^{-1} \sum_{i=1}^n K_H(x - X_i)$$

where $x = (x_1, x_2, \dots, x_d)^T$ and $X_i = (X_{i1}, X_{i2}, \dots, X_{id})^T, i = 1, 2, \dots, n$.

$K(x)$ is the kernel, which is chosen as standard normal PDF

$$K(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}x^T x\right)$$

The bandwidth H is estimated by (Wand & Jones, 1993)

$$H = \text{diag}(h_1^2, h_2^2, \dots, h_d^2)$$

These centroids will be used in the next step to get the land use type from land use definitions and descriptions used in the 1990 Land Use Study of Southern California, which were developed by Aerial Information Systems, Inc.. This classification uses a hierarchical system, allowing for easy aggregation and disaggregation of classes. If a centroid point falls inside a parcel, the parcel's land use type is used. If it is not in a parcel, the closest parcel will be used.

Demographic variables:

- Gender (Female, Male)
- Age (20 to 29, 30 to 39, and over 40)
- Number of vehicles
- Employment status (not employed, part time, and full time)

Trip characteristics:

- Weekday (Weekday, Weekend)
- Duration the participant stayed at a location (in minutes)

- The hour point when the participant arrived the location (1 to 24)
- Daily visit frequency (Total number of visits in the period of survey/ total of day surveyed)
- All travel modes a participant used in the trip to travel from origin to destination
 - Walking: WLK,
 - Driving: VDR,
 - Driving plus walking: VRD_WLK,
 - Taking transit plus walking: TS_WLK,
 - A combination of taking transit, walking, and driving: TS_WLK_VDR

The training data was finalized with 548 activities—at 314 locations from 32 participants with 154 days—in the BH data with purpose that matched in base line file (exclude final destination in a day which often is home). In order to better classify trip purpose, the original purposes were transfer into more general purposes. The following table shows the frequency of each trip purpose and how they are classified into a new group.

Table 5-1: Trip purpose definition

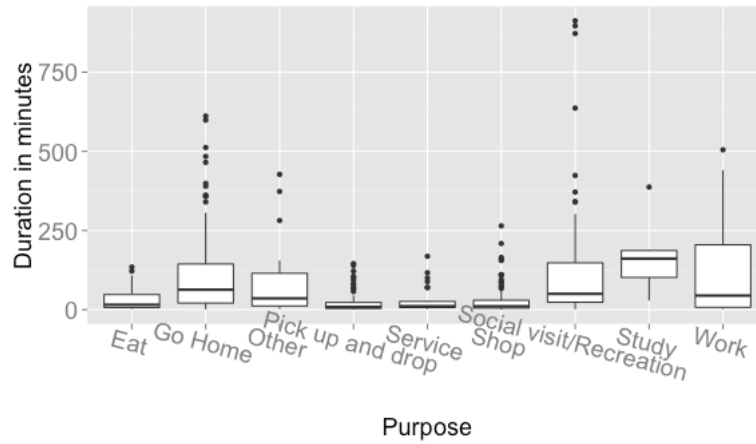
Old code		New Code	
Definition	Frequency	Definition	Frequency
Banking, post office, pay bills	20	Service	27
Medical	7		
Work	28	Work	29
Work-related (meeting, errand, etc)	1		
School/educational (attending class)	6	Study	6
Eating meal (restaurant, drive through, take out)	44	Eat	44
Shopping/Retail	131	Shop	159
Quick retail stop for ATM, buy gas, coffee, newspaper, etc.	28		
Church, temple, religious meeting	1	Social visit/ Recreation	101
Community meetings, political/civic event, volunteer work	8		
Entertainment (watching sports, move)	3		
Fitness/Sports/Recreation Activity	21		
Visit friends/relatives	68		
Pick up someone or get picked up	52	Pick up and drop	59
Drop off someone or get dropped off	7		
Home	102	Go Home	102
Other	16	Other	21
Change mode of transportation	5		
Total	548		548

After the model for trip purpose and other factors were defined, the trips for each participant were then processed to assign the appropriate trip purpose based on each activity's characteristics and related variables.

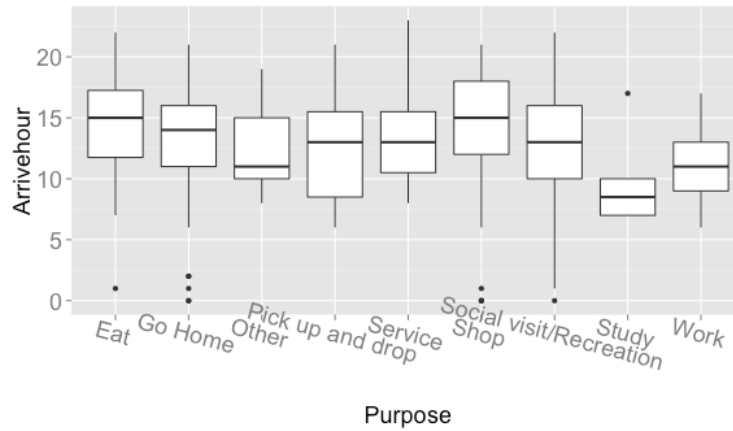
Table 5-2: Land use type vs purpose - BH data

	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Social /Rec	Study	Work	Total
FOD	39	0	3	0	0	1	0	0	3	46
IND	0	0	0	0	0	0	2	0	0	2
OFF	0	0	4	0	1	1	2	0	1	9
REC	3	0	0	1	0	0	22	0	3	29
RES	0	102	5	8	1	0	53	0	2	171
RET	2	0	2	12	1	143	15	0	9	184
SCH	0	0	5	33	0	4	3	5	2	52
SRV	0	0	2	5	24	10	4	1	9	55
Total	44	102	21	59	27	159	101	6	29	548

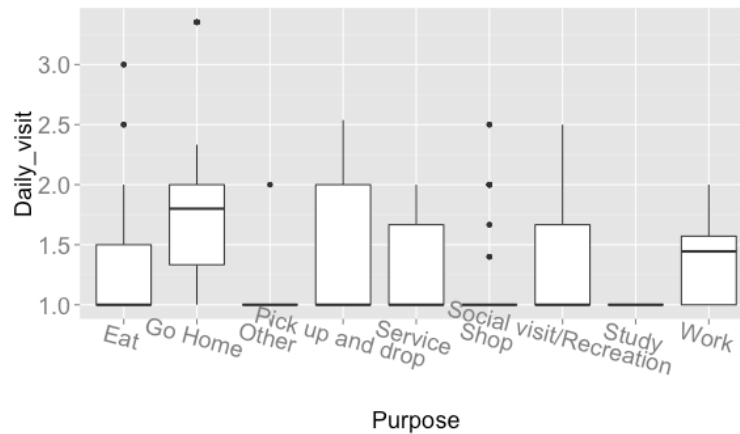
Figure 5-4 shows the box plot of activity duration, arrive hour and daily frequency for each purpose in the BH data. Outliers are plotted as individual points. The comparatively short box plot suggests that overall purpose have a high level of agreement with each other, for example box plot about activity duration (Figure 5-4a). While a tall one suggests the purpose holds quite different level of measurement, such as duration box plot for arrive hour (Figure 5-4b). The medians of each measurement (which generally will be close to the average) can be compared among purposes.



(a) Box plot in duration



(b) Box plot in arrive time



(c) Box plot in daily frequency

Figure 5-4: Box plot for trip purposes - BH data

Figure 5-5 and 5-6 display the patterns of activities in BH data in multi-dimensional plots.

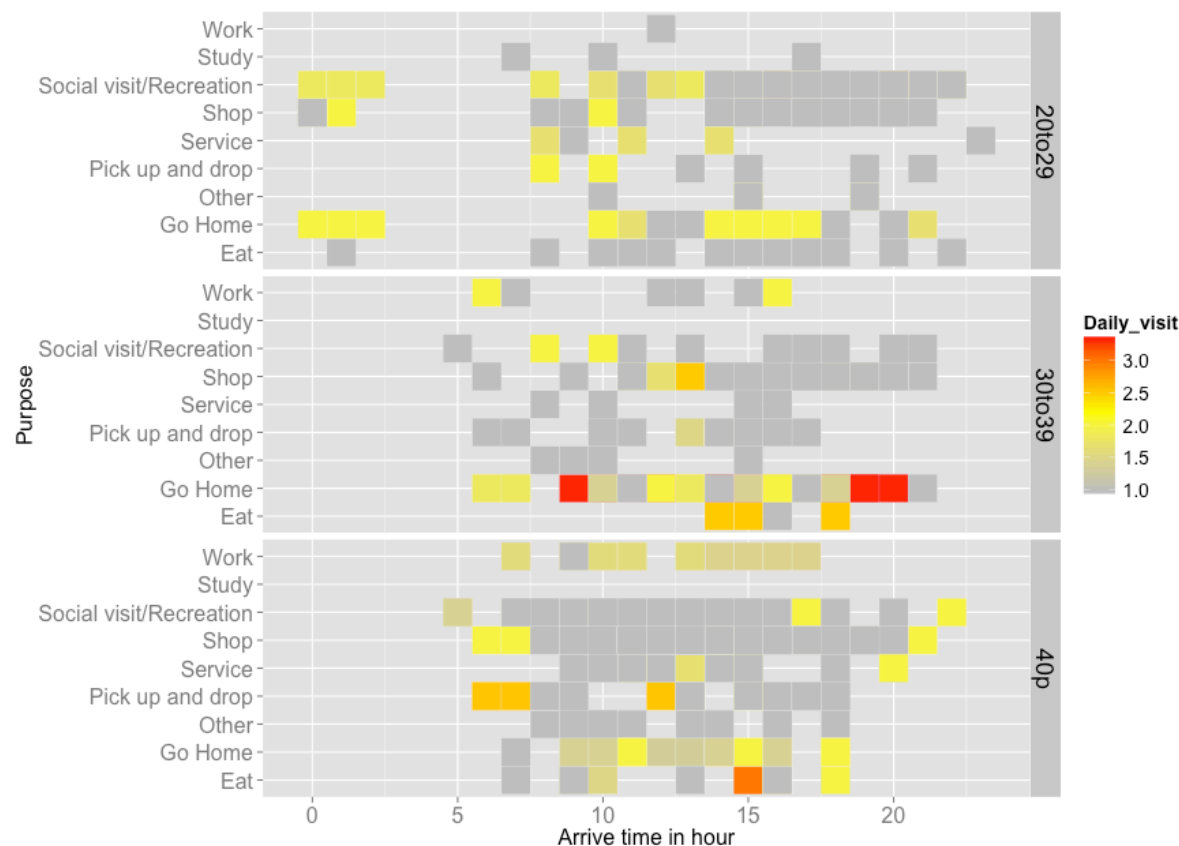


Figure 5-5: Association among arrive time, daily frequency, age, and purpose - BH data

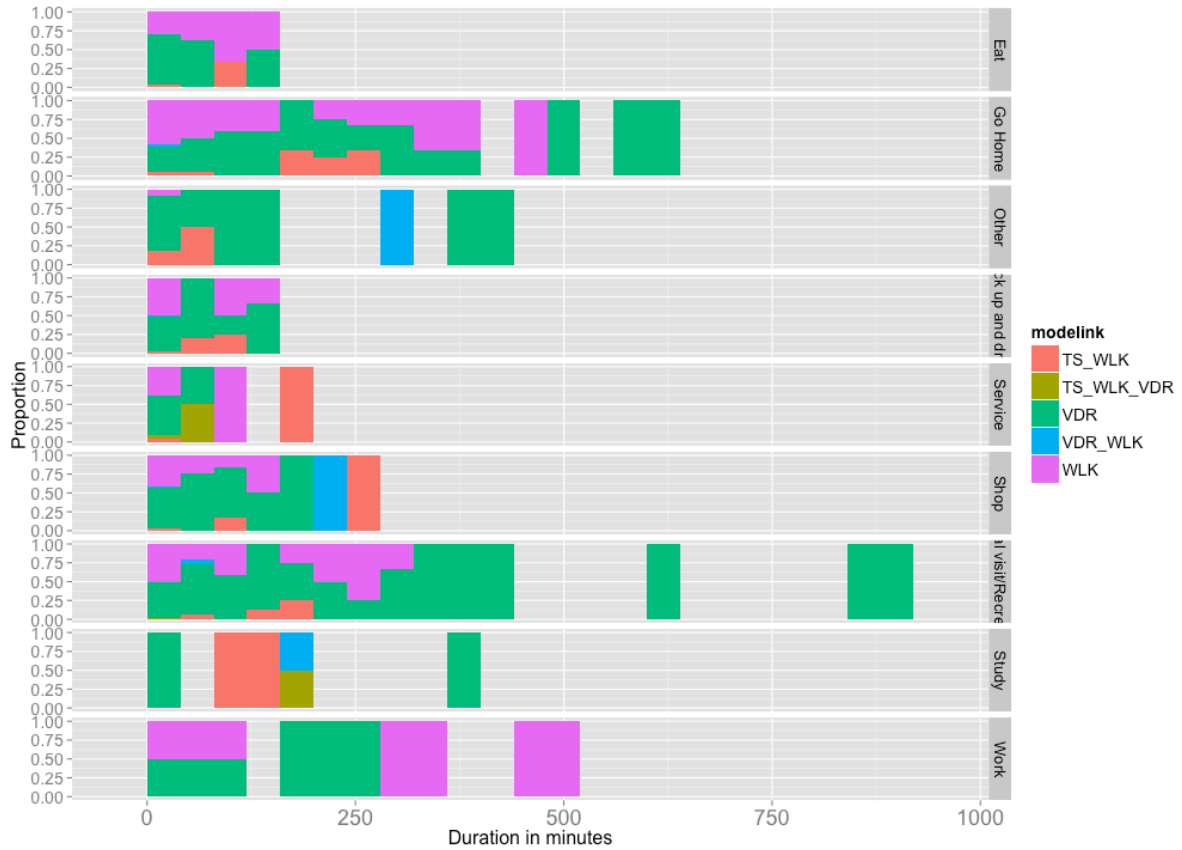


Figure 5-6: Distribution of activity duration by travel mode and purpose - BH data

We created a training set (2/3 BH data) of 432 trips and a testing set (1/3 BH data) of 145 trips for each classifier. The distribution of each purpose in each data set is as follows:

Table 5-3: Distribution of purposes - BH data

	Eat	Go Home	Other	Pickup /Drop	Service	Shop	Social /Rec	Study	Work	Total
Training set	35	76	17	42	23	121	72	4	21	411
Testing set	9	26	4	17	4	38	29	2	8	137
Total	44	102	21	59	27	159	101	6	29	548

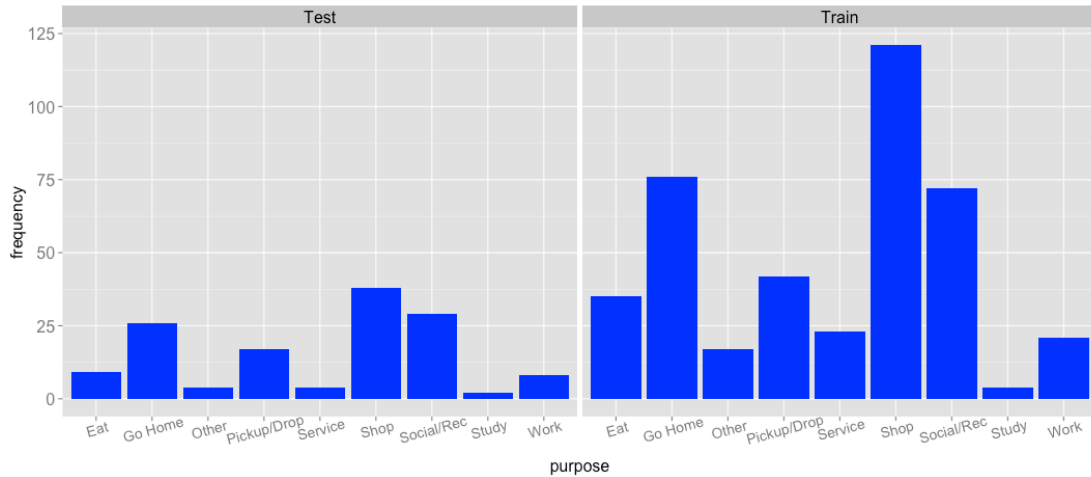


Figure 5-7: Distribution of trip purpose - BH data

5.4 Home location identification rules

In order to increase the accuracy of the classification model, we also create a set of rules to identify whether a location is a participant’s home or not. The rules will be validated with BH data, where the participants’ home locations are already known. We use these rules because the EXPO study did not tag the participants’ homes in the GPS data. Although we know their home address, the task of matching requires manual labor and it is difficult to apply to other studies, where this information may not be available.

Because the data were collected on multiple days, it is easier to identify which location is a participant’s home. Instead of using only one rule, such as land type of location, origin and destination of a tour, location staying at night, we use a combination of all these criteria. We could not use land type solely because the participant could visit a friend’s house, which also is located in a residential area. Additionally, the participant also could start/stop GPS device to be kept track/not track in the middle of a day and not at home.

First, the first origin and last destination of a day's trajectory were listed. We calculated the total duration the participant stayed at these locations over the period of survey. Then we ranked these locations by the total duration. For each participant, if the highest rank origin matches the highest rank destination and the land type of this location is residential, the location is set as home of the participant based on the activity duration criterion.

Second, the frequency of visiting a location, which appeared as the first origin and last destination, was calculated. Then, for each participant, if the highest ranked origin matches the highest rank destination and the land type of this location is residential, the location is set as home of the participant based on activity frequency criterion.

Finally, we compare the results of the two home identification criteria. If the two criteria generate the same conclusion, we have higher confidence in identifying whether or not a location is the participant's home. For a small number of samples, or with the participants' home address specified, this process of course could be unnecessary. However, in the case of large samples that do not provide home addresses, the process can help reduce time in manually checking.

5.5 Classification models

5.5.1 Results from Random Forest model

Since RF stores each of their training samples internally as they construct their decision trees, unlike many other machine learning methods, they can evaluate their own training error after the completion of training. On classification problems, this error is calculated as the percentage of misclassified training samples. A number of RFs are run to choose the best one.

Therefore, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error—it is estimated internally. The difference between the two accuracies is then averaged over all trees, and normalized by their standard deviation (Breiman, 2001).

The lowest error rate OOB of the 100 classification model for $n = 1,000$ trees and $mtry = 3$ is 18.98% for training RF on the training set. The variation in error rate for each class and OOB over trees is shown in Figure 5-6. Applying this model to the testing set also generates an OOB error rate of 18.98%. Although the OOB is the same, the error for each class varies. Trips for food (“Eat”) are detected with 22.22% in the testing set, an increase 4 times compared to the error rate in the training set. Conversely, trips for studying are detected 100% correctly in the testing set, compared to only 50% in the training set. The other classes’ error rates either remain the same (go home trips) or are just slightly different. The error rate can be visualized as a function of the number of trees grown (Figure 5-6). Figure 5-6 shows that the average training set OOB plateaus around 19% after a 250 trees.

Table 5-4: Trip purpose model - Confusion matrix – RF – training set

Predicted\ Actual	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Social/ Rec	Study	Work	Class error
Eat	33	0	0	0	0	0	2	0	0	0.0571
Go Home	0	76	0	0	0	0	0	0	0	0.0000
Other	2	0	2	1	2	2	8	0	0	0.8824
Pick up/ drop	0	0	0	29	1	6	5	0	1	0.3095
Service	0	0	0	4	13	5	1	0	0	0.4348
Shop	1	0	1	3	3	110	1	0	2	0.0909
Social /Rec	1	0	2	4	2	8	55	0	0	0.2361
Study	0	0	1	0	1	0	0	2	0	0.5000
Work	1	0	0	2	0	3	2	0	13	0.3810

Table 5-5: Trip purpose model - Confusion matrix - RF - testing set

Predicted\ Actual	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Socia /Rec	Study	Work	Class error
Eat	7	0	0	0	0	0	1	0	1	0.2222
Go Home	0	26	0	0	0	0	0	0	0	0.0000
Other	2	0	1	0	0	0	0	1	0	0.7500
Pick up/ drop	0	0	0	12	0	2	2	0	0	0.2500
Service	0	0	0	0	2	0	0	0	1	0.3333
Shop	0	0	1	2	0	34	2	0	0	0.1282
Social /Rec	0	0	2	3	0	1	23	0	1	0.2333
Study	0	0	0	0	0	0	0	1	0	0.0000
Work	0	0	0	0	2	1	1	0	5	0.4444

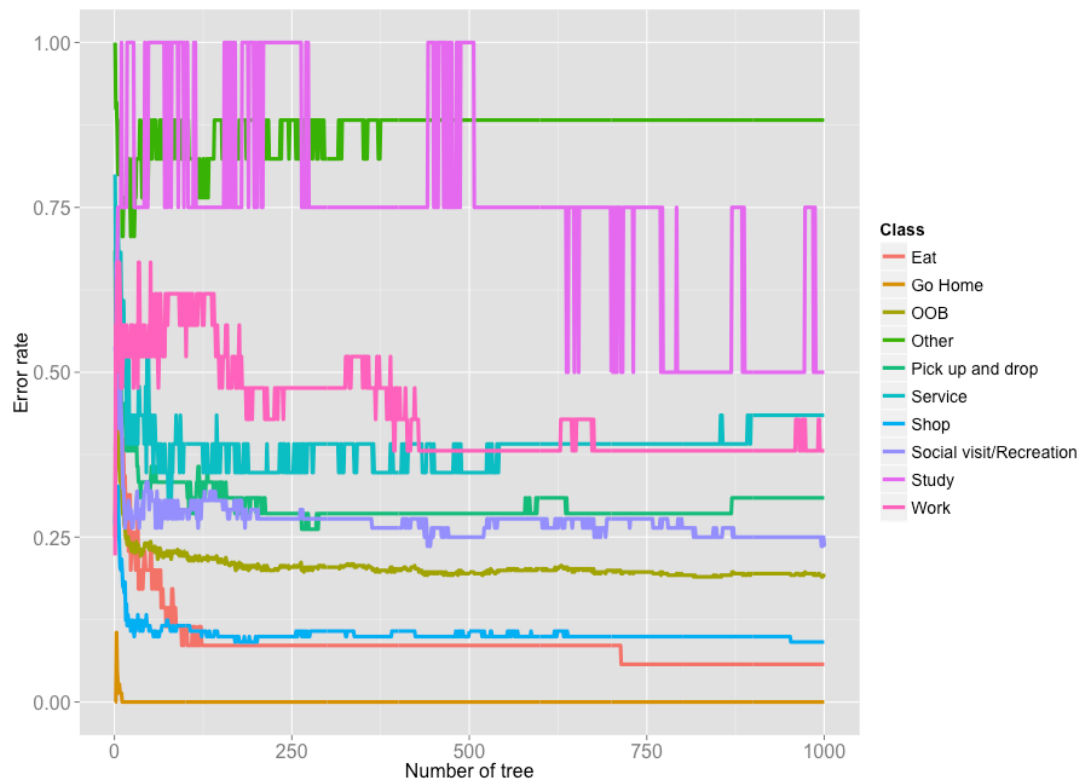


Figure 5-8: Error rate for trip purpose model

In addition to returning the classification results, RF provides a measure of the importance that each variable in the feature vector played in the result. These importance

scores are calculated by adding noise to each variable one-by-one and calculating the corresponding increases in the misclassification rate. The result is displayed in Figure 5-7, where the darker the bar's color is, the more importance the variables is. The results match our expectation that Location type is the main predictor. Combined with activity duration and the participant's home identification, these variables are the three most important factors for predicting trip purpose. Day-of-week variation factors affect the prediction accuracy the least.

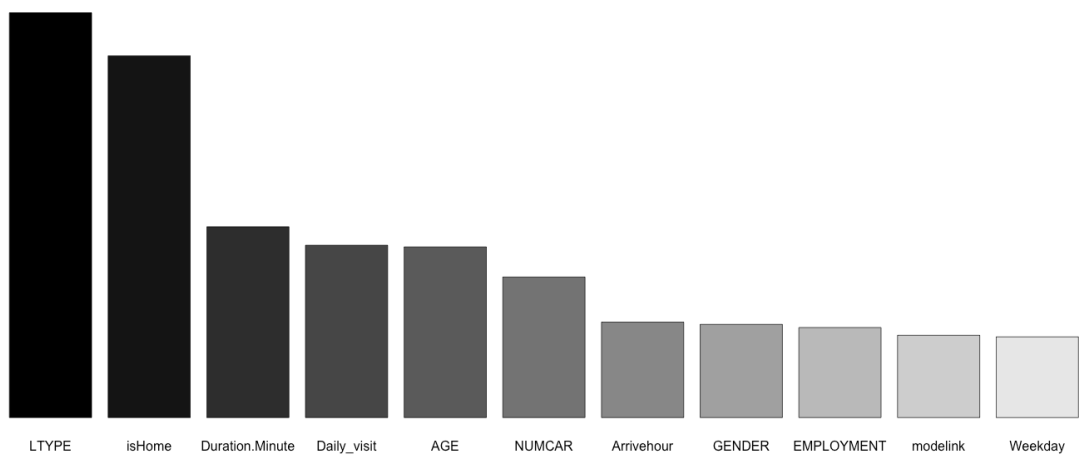


Figure 5-9: Trip purpose model - Variable importance ranking

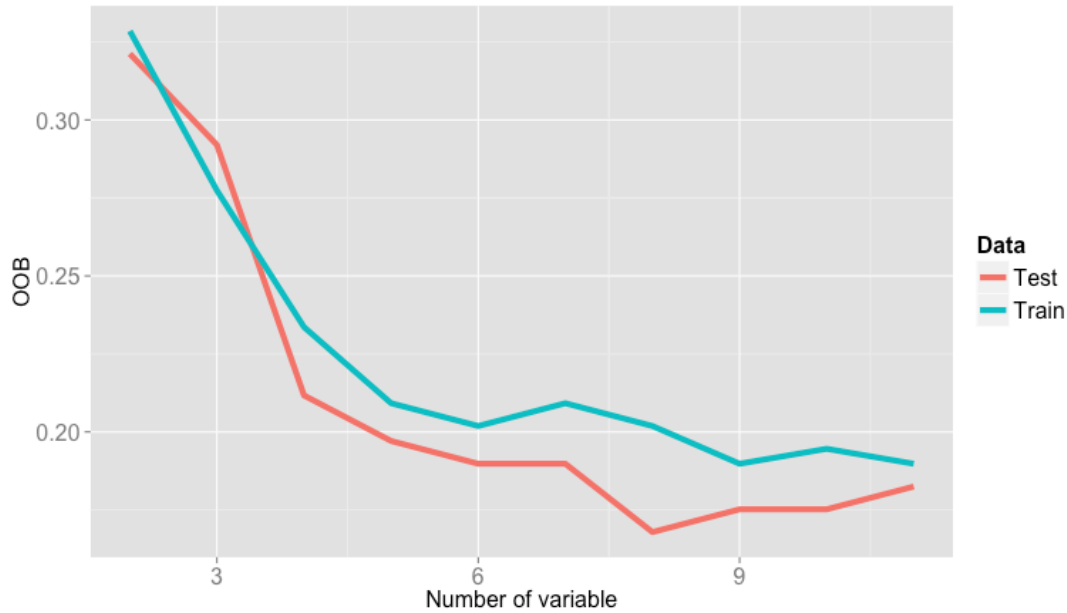


Figure 5-10: Trip purpose model - Error rate when adding variables

To improve the accuracy of the model, we apply a feature selection process to the model using backwards variable elimination to remove unimportant features for the model. The method throws out the least important variables until OOB prediction accuracy drops (Diaz-Uriarte & De Andres, 2006). This task also helps identify a small set of features that could be used for prediction purposes in practice, because all features may not be calculated in every case. This means obtaining the smallest possible set of segments' characteristics that can still achieve good predictive performance—even higher predictive power compared to the full model (Mode model in part 3.4.3). To select features we iteratively fit RFs. At each iteration, we build 100 new forests after discarding those features with the smallest variable importance based on mean decrease in accuracy (Figure 5-8) and choose the one with smallest OOB. The order of importance is kept through the iterations instead of being recalculated at a new iteration in order to avoid overfitting (Diaz-Uriarte & De Andres, 2006). Figure 5-8 shows that a

RF with 9 predictors provides the lowest OOB (18.98% using the training set and 17.52% on the testing set). The list of high-impact variables is:

- Location type
- Home identification
- Duration of activity in minutes
- Frequency per day
- Age
- Number of car
- Arrive time in hour
- Gender
- Employment status

Table 5-6: Trip purpose model with selected variables - Confusion matrix – RF –

Testing set

Predicted\Actual	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Social/ Rec	Study	Work	Class error
Eat	8	0	1	0	0	0	0	0	0	0.1111
Go Home	0	26	0	0	0	0	0	0	0	0.0000
Other	0	0	1	0	0	1	1	0	1	0.7500
Pick up/ drop	0	0	0	12	0	2	3	0	0	0.2941
Service	0	0	0	0	2	0	0	0	2	0.5000
Shop	0	0	1	1	0	35	0	0	1	0.0789
Social /Rec	1	0	0	1	1	2	23	0	1	0.2069
Study	0	0	1	0	0	0	0	1	0	0.5000
Work	1	0	0	1	1	0	0	0	5	0.3750

We also build another RF model for the training data without socio-demographic variables (age, gender, employment status, and number of vehicles in the household) but with

the identical setup for other parameters. The new model's OOB error rate is 23.11% for the training set, which is a 5.02% decrease in accuracy compared to the full model. Applying this model to the testing set, the OOB is 18.25%, which is only slightly different from testing RF full model (0.93%). The model could not correctly detect any study trips, with 100% error rate for this purpose. These results show that, with demographic variables, the model can predict the purpose of trips more accurately. However, without these predictors, the accuracy of the model is still high at 81.02% on training set.

Table 5-7: Trip purpose model without demographic variables - Confusion matrix –

RF – Testing set

Predicted\ Actual	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Social /Rec	Study	Work	Class error
Eat	8	0	0	0	0	0	0	0	1	0.1111
Go Home	0	26	0	0	0	0	0	0	0	0.0000
Other	1	0	1	0	0	0	0	1	0	0.6667
Pick up/ drop	0	0	0	11	0	1	0	0	0	0.0833
Service	0	0	0	0	3	0	0	0	1	0.2500
Shop	0	0	1	3	0	35	4	0	2	0.2222
Social /Rec	0	0	2	3	0	1	24	1	1	0.2500
Study	0	0	0	0	0	1	0	0	0	1.0000
Work	0	0	0	0	1	0	1	0	3	0.4000

5.5.2 Results from Naïve Bayes model

Because of the limited size of the training data, if certain count data for a categorical attribute is zero or if the mean and variance of a continuous attribute is such that the probability calculated is extremely small, the resulting total probability will be zero (or very close to zero), which may lead to biasness. It is customary to replace all probability values smaller than a given value by a small constant. All probabilities therefore have a lower limit,

which is represented by the *threshold* attribute. All values smaller than this threshold are replaced by the threshold value. We set the threshold value is 0.001.

To choose the best variables for the NB model, we compare these models by doing cross-validation with 10 folds for each of the model. The accuracy of each model is shown in the next table.

- Model 1: All variables (full model)
- Model 2: No demographic variables
- Model 3: No Weekday
- Model 4: No Home detection
- Model 5: No Age
- Model 6: No Employment status
- Model 7: No Gender
- Model 8: No Number of Car

Table 5-8: 10-fold cross validation for NB

Round	Model 1	Mode 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	0.5854	0.6341	0.6098	0.4878	0.5854	0.6341	0.6098	0.6098
2	0.6585	0.6829	0.6585	0.6341	0.6585	0.6585	0.6829	0.6341
3	0.7561	0.8049	0.8049	0.6585	0.8049	0.7805	0.8049	0.7805
4	0.7073	0.6829	0.6829	0.6585	0.7073	0.6585	0.6829	0.7073
5	0.7143	0.7857	0.6905	0.5952	0.7381	0.6905	0.7143	0.7143
6	0.7073	0.6829	0.7317	0.6098	0.7073	0.7073	0.7317	0.7073
7	0.7561	0.8049	0.7561	0.6829	0.7317	0.7805	0.7561	0.7805
8	0.7805	0.7805	0.7561	0.7317	0.7805	0.7805	0.7805	0.8049
9	0.5610	0.5610	0.5610	0.3902	0.5366	0.5610	0.5610	0.5610
10	0.6341	0.6098	0.6341	0.5854	0.6341	0.6341	0.6098	0.6098
Average	0.6861	0.7030	0.6886	0.6034	0.6884	0.6886	0.6934	0.6909

The best structure after this validation process is Model 2: without demographic variables include Age, Employment, Gender, and Number of Car. When applying this structure

to the training set, the overall accuracy is 74.45%. The next table shows conditional probabilities given the target class for each categorical variable. For each numeric variable, the table gives, for each target class, mean and standard deviation of the (sub-) variable. One can see that the land type has reasonable association with each trip purpose. The highest score by residential areas (100%) associated with the go home trip, followed by retailer areas (90.91%), service areas (86.96%), and selling food areas (85.71%), associated with shopping, service, and eating trips, respectively. Additionally, the average duration in the study trip is the highest duration (162.28).

Table 5-9: Prior probability

Purpose	Prior Probability
Eat	0.0852
Go Home	0.1849
Other	0.0414
Pick up and drop	0.1022
Service	0.0560
Shop	0.2944
Social visit/Recreation	0.1752
Study	0.0097
Work	0.0511

Table 5-10: Conditional probability

	Eat	Go Home	Other	Pick& drop	Service	Shop	Social /Rec	Study	Work
Arrive hour.m	14.57	13.01	12.76	13.00	13.70	14.65	12.67	10.25	11.00
Arrivehour.sd	4.04	4.60	3.67	3.79	4.40	4.21	4.53	4.72	3.21
Daily vist.m	1.33	1.87	1.12	1.35	1.23	1.16	1.29	1.00	1.36
Daily visit.sd	0.63	0.82	0.33	0.60	0.37	0.38	0.47	0.00	0.33
Duration.m	25.17	126.47	68.77	25.35	25.97	25.09	107.98	162.28	133.10
Duration.sd	25.21	149.34	103.76	38.49	31.80	34.39	147.17	157.51	151.51
FOD	0.86	0.00	0.18	0.00	0.00	0.01	0.00	0.00	0.05
IND	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
OFF	0.00	0.00	0.12	0.00	0.04	0.01	0.01	0.00	0.05
REC	0.09	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.10
RES	0.00	1.00	0.29	0.14	0.04	0.00	0.43	0.00	0.10
RET	0.06	0.00	0.06	0.21	0.04	0.91	0.17	0.00	0.24
SCH	0.00	0.00	0.24	0.52	0.00	0.02	0.04	0.75	0.10
SRV	0.00	0.00	0.12	0.12	0.87	0.06	0.04	0.25	0.38
TS_WLK	0.03	0.05	0.18	0.07	0.04	0.03	0.04	0.50	0.00
TS_WLK_VDR	0.00	0.00	0.00	0.00	0.09	0.00	0.01	0.00	0.00
VDR	0.66	0.49	0.71	0.52	0.48	0.56	0.58	0.50	0.57
VDR_WLK	0.00	0.01	0.06	0.02	0.00	0.02	0.01	0.00	0.00
WLK	0.31	0.45	0.06	0.38	0.39	0.39	0.35	0.00	0.43
Weekday	0.49	0.75	0.76	0.83	0.78	0.79	0.72	1.00	0.86
Weekend	0.51	0.25	0.24	0.17	0.22	0.21	0.28	0.00	0.14

Table 5-11 shows the confusion matrix when applying this NB model to the testing data, generating an overall accuracy of 75.91%. Although the model can detect 100% of going home trips, there are four other trips that are predicted with high error rate (over 60%). This problem could be due to the limitation of the sample size. Trips for social visiting or recreation, and shopping, have the second and third lowest error rates, respectively.

Table 5-11: Trip purpose model - Confusion matrix – NB – testing set

Predicted\ Actual	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Social/ Rec	Study	Work	Class error
Eat	6	0	0	1	0	0	0	0	3	0.4000
Go Home	0	26	0	0	0	0	0	0	0	0.0000
Other	3	0	2	0	0	0	1	0	0	0.6667
Pick up/ drop	0	0	0	11	0	1	3	0	0	0.2667
Service	0	0	0	0	2	2	0	0	1	0.6000
Shop	0	0	1	3	0	33	3	0	1	0.1951
Social /Rec	0	0	0	2	0	0	21	0	2	0.1600
Study	0	0	1	0	2	1	0	2	0	0.6667
Work	0	0	0	0	0	1	1	0	1	0.6667

5.5.3 Results from Multinomial Logistic model

We choose the first classes (“Eat”) in purpose classes as the baseline and calculate the odds that a member of activity i falls in category j as opposed to the baseline as π_{i1}/π_{ij} . A 10-fold cross validation will be processed using eight models with similar structure to the eight models validated with the NB algorithm. Again, the results suggest that, without demographic variable, the MNL algorithm has the highest accuracy (Model2, Table 5-11).

After 120 iterations, the Log-likelihood value is -263.01 and Residual deviance of 526.03 with 3,288 degrees of freedom. There are eight sets of coefficients for each independent variable that correspond to the eight different trip purposes that are not a work trip. The results are shown Table 5-13. The results indicate that an increase in daily frequency visiting a location increases the probability that the individual went there for food, while reduces the probability of other purposes when compared to work trip. If the participant used walking, transit, and driving (not separated by driver or passenger), it significantly increases the probability he or she went home, to service trips or social events instead of going out for food.

Further, if the participant only drove from an origin to a destination, he or she more likely went to work. A trip occurring during the weekend is least likely a studying trip.

Table 5-12: 10-fold cross validation for MNL

Round	Model 1	Mode 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	0.68	0.73	0.68	0.59	0.71	0.66	0.66	0.71
2	0.76	0.76	0.76	0.66	0.71	0.73	0.76	0.71
3	0.56	0.78	0.61	0.54	0.63	0.66	0.63	0.66
4	0.80	0.76	0.80	0.76	0.80	0.80	0.76	0.76
5	0.74	0.69	0.74	0.67	0.71	0.71	0.71	0.74
6	0.80	0.90	0.80	0.76	0.83	0.88	0.80	0.83
7	0.83	0.83	0.85	0.78	0.83	0.88	0.85	0.83
8	0.73	0.71	0.71	0.61	0.73	0.73	0.71	0.73
9	0.71	0.73	0.71	0.61	0.71	0.71	0.73	0.76
10	0.56	0.61	0.61	0.51	0.61	0.56	0.61	0.59
Average	0.72	0.75	0.73	0.65	0.73	0.73	0.72	0.73

Table 5-13: Coefficient for MNL - trip purpose

Vs Eat	Go home	Other	Pick& drop	Service	Shop	Social /Rec	Study	Work
(Intercept)	-23.58***	6.25***	-32.41***	-36.85***	1.29***	-34.65***	42.53***	-24.71***
Arrivehour	0.03***	-0.11***	-0.02***	0.04***	0.06***	0.03***	-0.36***	-0.10***
Daily_visit	-7.40***	-3.64*	-2.82***	-3.46***	-2.10***	-3.01	-73.88	-2.17***
Duration	-0.03***	0.00	-0.02	-0.04	-0.02	0.00	0.00	0.00
isHome	74.97***	-23.30***	-23.66***	-22.42***	26.23***	-33.83***	47.28***	-26.47***
LTYPE IND	0.28	-3.75	-5.92***	-1.60	-4.99***	115.27***	0.31***	-1.63
LTYPE OFF	1.83	43.33***	-20.30***	84.29***	44.09***	80.44***	-9.62***	42.70***
LTYPE REC	18.12	-22.60***	-23.88***	-12.79***	-29.57***	46.49***	6.08***	5.82***
LTYPE RES	52.79***	49.92***	86.78***	89.43***	-5.70**	89.72***	57.96***	48.64***
LTYPE RET	9.23***	3.03	42.13***	43.58***	9.07***	43.44***	14.85***	4.83***
LTYPE SCH	41.67***	60.48***	99.52***	4.37**	60.76***	97.53***	98.28***	58.92***
LTYPE SRV	49.81***	48.76***	86.92***	92.87***	51.36***	86.69***	86.80***	49.71***
TS_WLK_VDR	4.11***	-5.26***	-9.86***	42.84***	-17.76***	7.89***	-0.62	-5.49***
VDR	-11.38***	-3.24***	-2.50***	-3.10***	-2.94***	-3.07***	-4.98***	25.67***
VDR_WLK	16.38***	13.37***	13.98***	-5.12**	15.62***	12.13***	-31.25***	-15.16***
WLK	-18.38***	-4.92***	-3.66***	-1.70***	-3.37***	-3.05***	-30.40***	25.50***
Weekend	3.35**	-1.15	-2.16**	-2.01***	-2.11***	-1.94***	-24.03***	-1.98**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

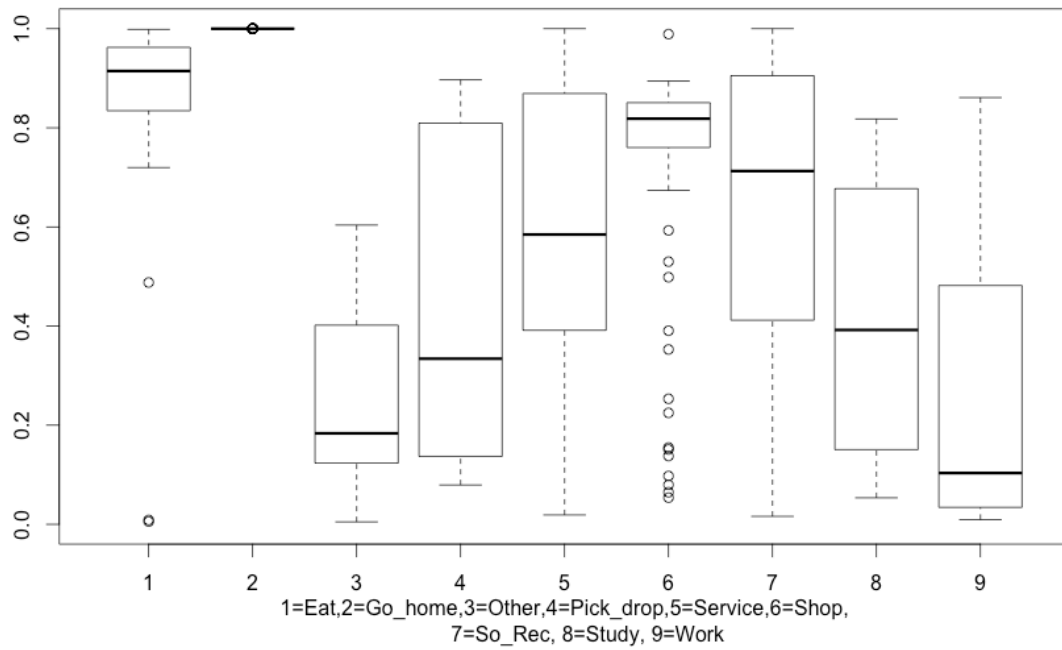


Figure 5-11: Box plot of MNL's fitted value

The model results in an accuracy of 78.10% when tested with the testing set and generates the confusion matrix displayed in Table 5-14. The model classifies going home trips with 100% accuracy and can classify picking up or dropping trips successfully at a rate of 91.67%. However, the model could not detect any study or work trips.

Table 5-14: Trip purpose model - Confusion matrix – MNL – testing set

Predicted\ Actual	Eat	Go Home	Other	Pick up/ drop	Service	Shop	Social /Rec	Study	Work	Class error
Eat	8	0	0	0	0	0	0	0	2	0.2000
Go Home	0	26	0	0	0	0	0	0	0	0.0000
Other	1	0	2	0	0	0	1	1	0	0.6000
Pick up/ drop	0	0	0	11	0	1	0	0	0	0.0833
Service	0	0	0	0	3	2	0	0	1	0.5000
Shop	0	0	1	3	0	33	3	0	1	0.1951
Social /Rec	0	0	1	3	0	0	24	1	4	0.2727
Study	0	0	0	0	1	1	0	0	0	1.0000
Work	0	0	0	0	0	1	1	0	0	1.0000

5.5.4 Predictive performance comparison

In this section, we compare the performance of the three methods used to classify trip purposes based on the BH data.

5.5.4.1 Accuracy

Table 5-15 presents the results for the comparison among methods used for trip purpose imputation. The measure of accuracy is obtained by dividing the sum of the diagonal elements of the confusion matrix by the total number of test points. It illustrates that the accuracy of all models for the testing data is lower than that of the RF models. The RF model has an accuracy of 82.48% while the lowest rate belongs to the NB model with 75.91%. This result again proves that this ensemble method outperforms other the methods with multinomial response data.

Although we decide to use the selected model design for trip purpose imputation in the next step as the result of cross validation process, we still calculate and show the class error as well as overall accuracy of each algorithm base for full model design (including demographics).

The results show a small difference in overall accuracy between each pair-model (selected vs. full), but the difference between classes has a bigger gap. In particular, neither RF nor MNL with reduced features format could predict any study trips.

Table 5-15: Error rate comparison

Purpose	RF-select	RF-full	NB-select	NB-full	MNL-select	MNL-full
Eat	0.1111	0.2222	0.4000	0.3750	0.2000	0.3636
Go Home	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Other	0.7500	0.7500	0.6667	0.7143	0.6000	0.7500
Pick up/ drop	0.2941	0.2500	0.2667	0.3125	0.0833	0.1538
Service	0.5000	0.3333	0.6000	0.5000	0.5000	0.5000
Shop	0.0789	0.1282	0.1951	0.1951	0.1951	0.2105
Social /Rec	0.2069	0.2333	0.1600	0.2083	0.2727	0.2500
Study	0.5000	0.0000	0.6667	0.7500	1.0000	0.0000
Work	0.3750	0.4444	0.6667	0.5714	1.0000	0.5000
Overall Error rate	0.1752	0.1898	0.2409	0.2555	0.2190	0.2263
Overall Accuracy	0.8248	0.8112	0.7591	0.7445	0.7810	0.7737

Figure 5-12 presents the sensitivity of each classifier on the testing set by purposes. Sensitivity (or recall) is a measurement of the model's ability to correctly classify an activity as having the correct purpose or not. In other words, it refers to a model's ability to designate an activity with purpose as positive. A highly sensitive model means that there are few false negative results, and thus fewer trips are misclassified. The diagonal cells show the ratio of correct positives predicted to the actual number of positives. In general, RF classifies all purposes at a relatively higher true positive rate compared to other methods. RF and MNL detect 5 purposes (going home, other, pickup or drop, social visit, and study) at the same sensitivity level. Leaving aside computational issues, NB has better sensitivity than MNL. This is consistent with the knowledge that discriminative classifiers are not always to be preferred to generative ones (Ng & Jordan, 2002).

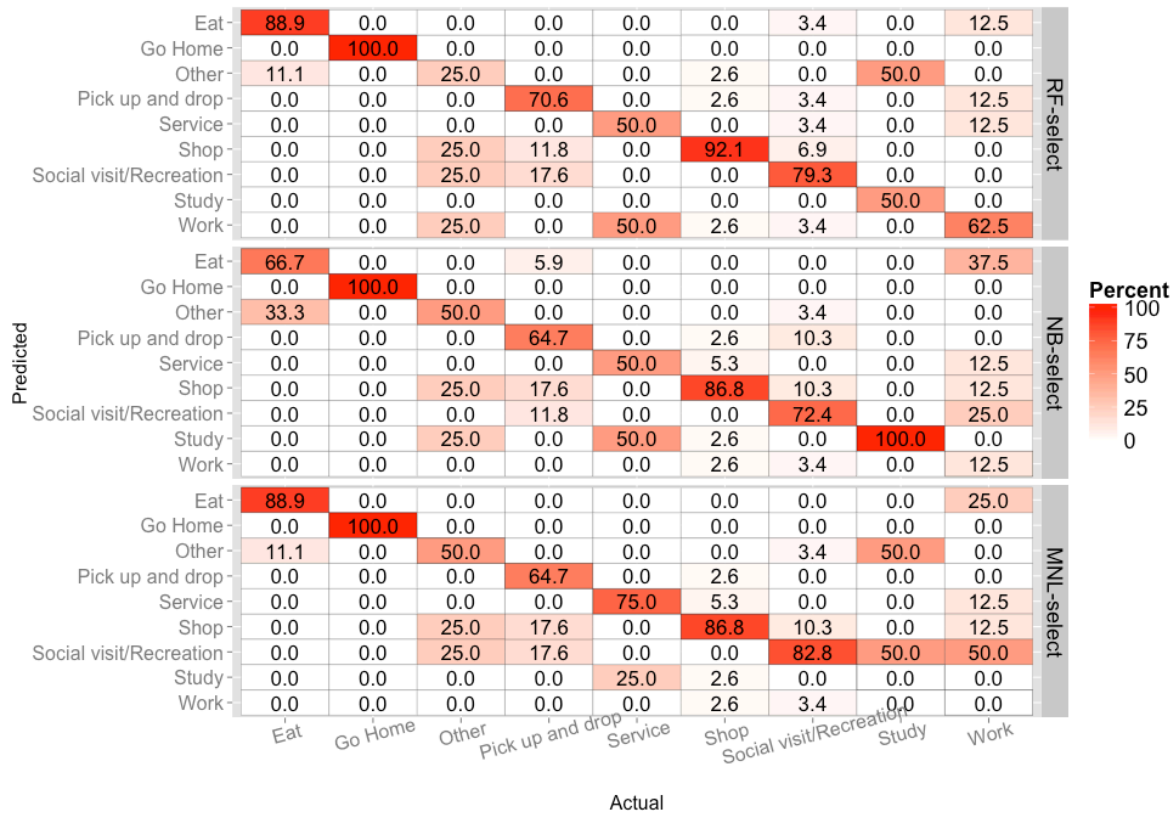


Figure 5-12: Sensitivity of classifying trip purposes

Unlike parametric models, RF models do not suffer from a specification bias and are able to model highly complex relationships without transforming variables. While MNL models provide easily interpretable coefficients and significance statistics, RF are very difficult to interpret and perceived as being “black boxes” which can neither explain how they reach an outcome nor provide an explicit representation of the relationship that they estimate. We consider the trade-off between interpretability and predictive power when predicting purpose of trips.

5.5.4.2 The area under the ROC curve

One of the more popular criteria for comparing the performance of classification rules is the area under the Receiver Operating Characteristic (ROC) curve (AUC). With multi-classes problems, it is calculated by averaging pairwise comparisons. For each pair, AUC compares the overall distributions of $\hat{p}(x)$ for a class points and the other class points. The classification rule will generally be better the more these two distributions differ. In other words, the bigger an area is, the more accurate the classifier is. An area of 1 represents a perfect classifier.

We transfer the name of class to a number as $0, 1, 2, \dots, m - 1$ with the order of the labels not reflecting any intrinsic order to the classes (here $m = 9$). Each model will provide the probability $\hat{p}(i|x)$ that each activity belongs to each purpose with $i = 0, \dots, m - 1$. For any pair of classes i and j , we can compute the measure $\hat{A}$ using either $\hat{p}(i|x)$ or $\hat{p}(j|x)$. We therefore define $\hat{A}(j|i)$ as the probability that a randomly drawn member of class j will have a lower estimated probability of belonging to class i than a randomly drawn member of class i . Note that $\hat{A}(j|i)$ is different from $\hat{A}(i|j)$. Hand and Till (Hand & Till, 2001) adopted the measure of separability between classes as:

$$\hat{A}(i, j) = \frac{\hat{A}(i|j) + \hat{A}(j|i)}{2}$$

Therefore, the overall performance of the classification rule in separating the m classes is then the average of this measure over all pairs of classes:

$$M = \frac{2}{m(m-1)} \sum_{i < j} \hat{A}(i, j)$$

The M values for RF, NB, MNL model are calculated as 0.9331, 0.8903, and 0.8925, respectively. This procedure generates the same accuracy order for these three classifiers as when we use overall error rate measurement in the previous part.

5.5.4.3 Summary

According to the simplest method of selecting models, i.e., to pick the one with the best performance according to a given testing data, RF is chosen based both on accuracy measurement and AUC measurement.

5.6 Conclusion

Identifying trip purpose can offer advantages in transportation planning and transportation engineering, within a wide range of applications. This task is particularly important with GPS travel data collection, where information on trip purpose is not available, unless we conduct a recall survey, or a parallel travel survey diary in the period of collecting data, which can be burdensome to the participants. In this study, we demonstrate and evaluate several methods of automating predicting trip purpose for the EXPO study data. Random forest, Naïve Bayes, and Multinomial logit algorithm are employed, validated, and tested to identify the most suitable model, i.e., one which has the highest predictive power at both class level and overall level. We take advantage of other data sets with available participant-derived trip purposes to train the model.

Estimation results show that, overall, RF is the best model. Because we want detailed, rather than general, trip purposes, nine specific purposes were classified: work, study, eat, go

home, shop, social/recreation, service, pick up & drop, and other. As the number of classes increases, class differentiation becomes more difficult. The overall accuracy for prediction is 81.12% on the testing set. The model also can predict accurately shopping trips with the sensitivity of 92.1%, go home and studying trips reach 100%.

Although the full model includes four demographic variables: age, gender, employment status, and number of household's vehicle, we also build a model without these variables in order to test how trip purposes will be imputed in the case that these factors are not observed. The result is interesting in that the accuracy of the model even increases to 81.75%. It implies that without socio-demographic variables, the trips' characteristics alone can predict the purpose of trips accurately. This conclusion is important in practice, when applying to large-scale trip purpose imputation, where it may be difficult or impossible to get all participants' personal information. The model's results also provide the clue that land use type of a location is the strongest indicator for imputing purpose of trips.

CHAPTER 6: COMPARISON OF ACTIVITY PATTERNS OVER TIME

6.1 Introduction

Activity-based analysis has received great interest for its ability to provide a more comprehensive understanding about the implications of various transportation policies on travel demand and, therefore, more effectively evaluate the efficiency of such policies. Although numerous researchers have had some success in determining the causal factors on travel behavior based on individual or household activities on the same days (interpersonal – variability), the variability in travel behavior for a given individual/ household over a span of time (intrapersonal – variability) remains only sparsely investigated. The objectives of this portion of the study are to complement and evaluate knowledge gained from data collection using GPS associated relative to the analysis of changing personal travel behavior over time. In this study, the variations of different kinds of activity patterns across a seven-day period will be explored. This time scale has the potential to reveal the underlying features of local patterns as being routine or habit. Further, the study extends the previous work to identify changes both in travel patterns as well as in activity patterns over different waves to understand how they are formed and how they are broken under changes in the living environment and transportation system.

6.2 Background

6.2.1 Activity-based modeling

The activity-based approach to travel demand analysis views travel as a demand derived from the need to pursue activities distributed in space (P. Jones, Koppelman, & Orfeuil, 1990). It

aims at predicting which activities are conducted where, when, for how long, with whom, the transport mode involved and ideally also the implied route decisions. It represents a shift in focus from trips to activities that began in the 1970s with the explicit assumption that most travel is not an end in itself but a means to bridge activities that are separated in time and space (Arentze & Timmermans, 2004). Activity-based models recognize interactions among various members of the household that arise when household members allocate resources (such as household vehicles) among themselves, assign and share tasks, and jointly engage in activities. They can also express restrictions to the physical and social resources that control activity participation in the models, for example, by including store open times or public transportation schedules. As such, it has been argued that activity-based approaches provide a stronger theoretical and conceptual framework that could reflect more accurately how people plan and organize their days.

Table 6-1: Comparison of Trip-based and Activity-based models

Activity based models	Trip based models
- Travel demand is derived from activities - 24 hour schedule with activities - Household interaction - Time and space constraints	- One-way trips - A day is a combination of broadly defined peak and off-peak time periods - Individual trips - Calibration is needed to fit the data to the real situation

There are common features across all models representing core concepts of the activity-based model paradigm:

- A tour-based structure, in which the tour—a closed chain of trips starting and ending at the base location (home or workplace)—is used as the main unit of modeling travel.

- An activity-based platform that implies that modeled travel is derived within the general framework of the daily activities undertaken by households and persons. This allows for the consistency of the typological, spatial, and temporal dimensions of individual activity patterns, the substitution between in-home and out-of-home activities, the duration of activities in a coherent framework with trip departure and arrival times, intra-household interactions, and other aspects pertinent to activity analyses.
- A modeling technique that is applied at the disaggregate level of persons and households, which converts activity- and travel-related choices from fractional-probability model outcomes into a series of ordinal or nominal decisions among the discrete choices; this method of model implementation results in more realistic model outcomes, with output files that look much like actual travel/activity survey data.

The modeling methods are divided into several groups: constraints-based models, utility maximization models, and computational process models. Constraints-based models have their roots in time geography and optimization models in microeconomic theory, while computational process models have been inspired by psychological decision process theories. Utility maximization models and computational process models are the most popular both in theory and practice.

6.2.2 Time use research

Time use research has been playing an increasingly important role in activity and travel behavior research because of the recognition that many travel choices are governed by time,

which is a limited resource that is consumed according to one's needs and preferences (Pendyala, 2002). According to Pendyala (2002), the most popular aspects of time use research are summarized as following:

- Daily time allocation: focusing on total time allocated to various activity categories or purposes.
- Activity episode duration: studying the factors that influence individual activity episode durations and the probability that a certain activity will be terminated given that a certain duration has elapsed.
- Activity timing and scheduling: examining when a certain activity or trip will happen.
- Activity sequencing: capturing trip chaining by several methods, such as sequence alignment techniques, discrete choice models, and heuristic rule-based algorithms.
- Activity frequency: examining the number or times an activity is pursued, regardless of the durations of the episodes.

Household daily activities are separated as two parts: habitual and nonhabitual activities. Routine or habitual activities are fixed in space and time while nonroutine/nonhabitual activities are more flexible because their performance depends on choices of location, start time, duration, and travel mode (Gärling, Kalén, Romanus, Selart, & Vilhelmson, 1998). Schonfelder and Axhausen (Schönfelder & Axhausen, 2010) defined habitual or routine travel behavior is "the reuse of behavioral segments, sequences" in identical or similar decision situations. In daily life, routine activities take precedence over nonroutine ones. Only when a nonroutine activity becomes very important, may it override a routine one. This idea also was shown in the research of Hannes et al. (Hannes, Janssens, & Wets, 2008): daily activity is

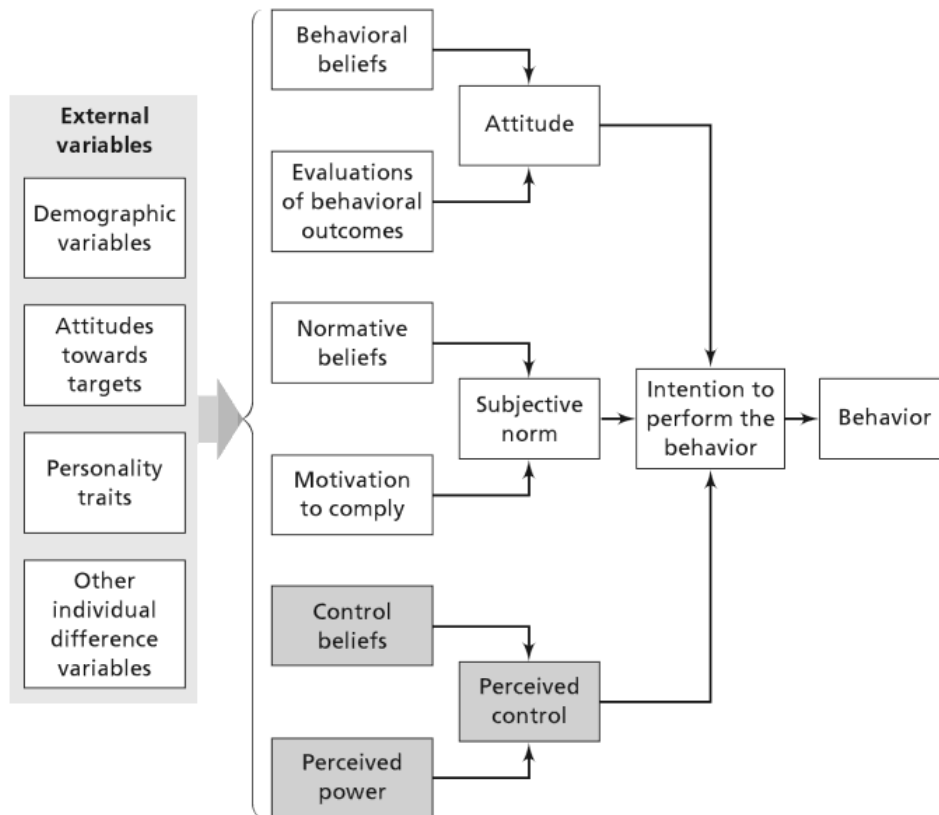
mainly fixed in mental scripts that are cued by certain situations. The actual and perceived choice sets of activities are very limited and justified by long-term decisions (residential location, vehicle ownership), decisions of others, logical reasoning, opinions, and habits. It is “highly automated routine behavior, often performed mindlessly” (Hannes et al., 2008).

Activity duration was found to play the most significant role in the models of Doherty and Mohammadian (Doherty & Mohammadian, 2011). It, rather than the nature of the event as indicated by activity type, is a primary driver of within-tour planning order. Activities with longer durations and greater travel times tend to be planned earlier. Therefore, study of the duration of activities continues to be one of the more interesting issues in activity-based modeling.

6.2.3 Psychological impacts

6.2.3.1 Theory

There is a large body of literature in psychology that studies the relationship between habit and behavior. Among several theories that have aimed at explaining habit and their behavioral relationship, Triandis’s Theory of Interpersonal Behavior (TIB) (Triandis, 1977) is one the most well-known(Matthies, 2003). It is in the same school of cognitive models with the Theory of Reasoned Action (TRA) and the Theory of Planned Behavior (TPB) (Fishbein & Ajzen, 1975). However, Triandis extended the framework by adding habits and facilitating conditions and affect. Bamberg et al. (Bamberg, Ajzen, & Schmidt, 2003) found that including such factors as habit increases the model’s predictive power over other models such as the TPB. A schematic of these theories is shown in the Figure 6-1.



Note: Upper light area shows the TRA, entire figure shows the TPB

Source: *Theory of reasoned action, theory of planned behavior, and the integrated behavioral model.* (Montano & Kasprzyk, 2008)

Figure 6-1: Theory of Reasoned Action and Theory of Planned Behavior

In Figure 6-1, the following definitions apply:

- Control beliefs: an individual's beliefs about the presence of factors that may facilitate or impede performance of the behavior (Ajzen, 2001).
- Behavioral intention: an indication of an individual's readiness to perform a given behavior. It is assumed to be an immediate antecedent of behavior (Ajzen, 2002).

- Behavior: an individual's observable response in a given situation with respect to a given target.

Verplanken & Aarts (Verplanken & Aarts, 1999) defined habits as '... learned sequences of acts that have become automatic responses to specific cues, and are functional in obtaining certain goals or end-states'. Habits have a history of repetition. The more frequently a behavior is performed, the more likely it is to become habitual. Thus, past behavioral frequency as a measure of habit strength is mostly used by social psychologists (Eagly & Chaiken, 1993; Verplanken & Aarts, 1999), and TIB is useful in explaining and understanding complex human behaviors (Figure 6-2).

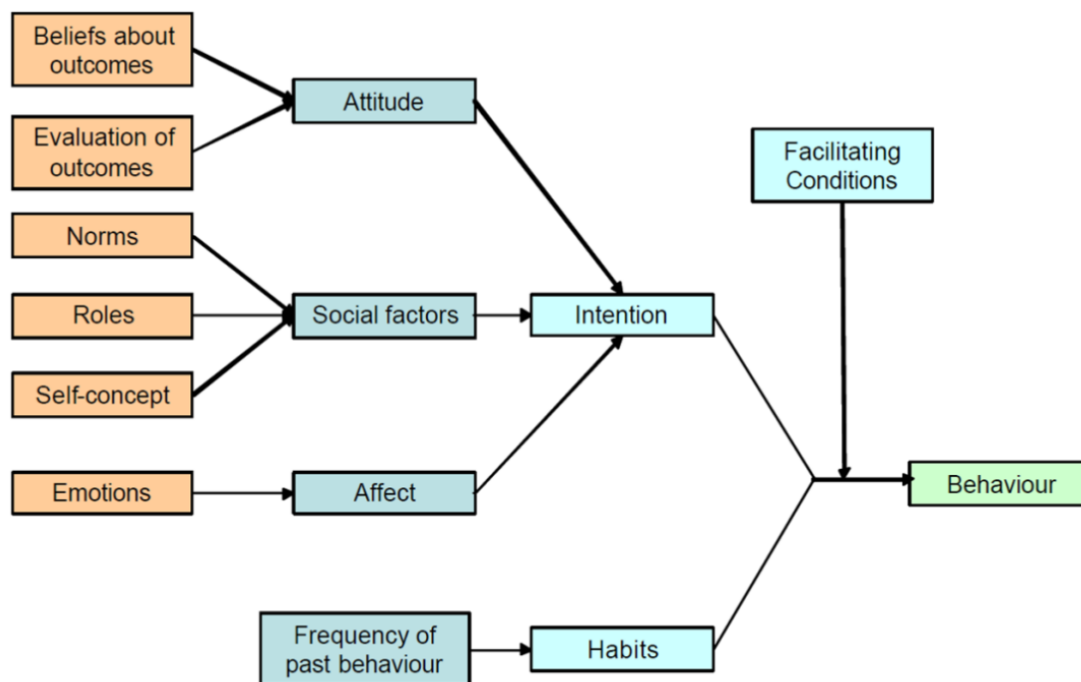


Figure 6-2: Framework of Theory of the Interpersonal Behavior

6.2.3.2 Applications in travel behavior research

Human psychological factors are often the basis for deep-rooted habits and preferences. Behavior is often shaped by habits. Put simply, habit is defined as the repeated performance of behavior sequences (Gärling & Axhausen, 2003). Gärling et al. (Gärling, Fujii, & Boe, 2001) showed that travel habits represent behaviors that require less conscious decision-making prior to their performance, such as in the case of the work trip. Individuals may become so familiar with a situation and the accompanying decisions that they may not consider any alternative mode and route choices (Burbidge & Goulias, 2013). Their costs are generally too high and the expected gains associated with new alternatives too uncertain. The travelers reuse past solutions to make their behavior easier and less risky (Gärling & Axhausen, Introduction: Habitual travel choice, 2003). In predicting students' car use for university routes (Bamberg et al., 2003) or mode choice decision (Domarchi, Tudela, & Gonzalez, 2008), estimation showed that attitudinal variables and habitual character presented a significant contribution to modal utility, and helped to improve both fitness and statistical significance.

Another study by Axhausen, et al. (Axhausen, Zimmermann, Schönfelder, Rindsfuser, & Haupt, 2002) also showed that the routine effect for some decision-making characteristics is similar across age. Routine was the second most frequent type of decision relative to activity duration, followed by plans made together to carry out the activity. The number of activity locations that both elderly and non-elderly age groups actually considered when making their travel choices was limited, and routine choices still usually guided their final decisions. For the non-elderly group, routine decisions were the most common, accounting for 35.8% of all decisions to engage in an activity. Additionally, empirical evidence derived from the CHASE

survey (Doherty & Miller, 2000) shows that households begin the week with a firmly established set of routine weekly activities. About 45% of the activities that take place on weekdays, and about 20% of those that take place on the weekend are pre-planned before the week starts on or before the first Sunday.

Both theory and research point to the appearance of psychological impacts, particularly habit, in travel behavior. There is a need for further understanding this factor in order to better understand travel behavior. Thus, this research uses TPB and TIB as foundations for such investigation.

6.3 Methodology

The objectives of this portion of the overall study are to complement and evaluate knowledge gained from data collection using GPS associated with the analysis of changing personal travel behavior over time. Therefore, the variations of different kinds of activity patterns across a seven-day period and a three-wave panel are explored. This time scale can reveal the underlying local patterns, such as routine or habit. Further, the study extends to identify changes in both travel patterns and activity patterns over different waves to understand how they are formed and how they are broken under changes of living environment and transportation system.

First, the travel behavior change can be described by a series of multilevel models which allow multiple relationships to be modeled simultaneously (Goulias, 2003). A multilevel model explains microlevel outcomes in two ways. First, parameters of microlevel covariates in multilevel models are expressed as a function of context as macrolevel variables. Second, this

micro-macro relationship can be interpreted by the characteristics of social groups (economic sectors, household, peer groups) and temporal contexts (different time points) (J. Ma & Goulias, 1996). A multilevel model structure is used in this study in the temporal context. Such a time effect is expressed in three different scales, which are analogous to the time frames of time of day, day-to-day and wave-to-wave, capturing short-term and long-term time effects, respectively. In this way, the time of day, day, and the wave effects can be examined in more detail, in contrast to a general regression model that may lack balance by estimating these effects separately. We apply this method to activity duration. Functionally, a multilevel model is equivalent to a standard OLS regression. For each level, a model (a regression equation with coefficients) is estimated describing the relationships among variables. These coefficients then become the dependent measures at the next level of analysis. The model can be written as:

$$\begin{aligned}
 Duration_{idw} = & \beta_{0idw} + \\
 & \sum_{a=1}^A \beta_{1a} Activity.Attribute_{aidw} + \\
 & \sum_{p=1}^P \beta_{1p} Person.Attribute_{pidw} + \\
 & \varepsilon_{idw}
 \end{aligned}$$

where $Duration_{idw}$ is the amount of time at a location in time of day i ($i = ("Early Morning", "Morning", "Afternoon", "Evening")$) in day d ($d = ("Weekday", "Weekend")$) of wave w ($w = (1, 2, 3)$). $Activity.Attribute_{aidw}$ is an indicator variable a that relates to that activity such as time of day, travel modes to that location, frequency, and activity type. $Person.Attribute_{pidw}$ is an indicator variable p that relates to demographics information of participant j , such as age, gender, employment status, and number of cars. ε_{idw} is the error for

time of day i in day d of wave w , where each group's error is assumed to be multivariate normal distributed. Also, to investigate the difference between people's behavior across day of week and waves, the K-group Multivariate Analysis of Variance (MANOVA) can be used.

6.4 Data

We applied the trip purpose models built in the previous chapter to EXPO data. The result is particularly important since we aim at semantically enriching raw GPS trajectories with more meaningful information that can be used in transportation planning, policy, sociology, and marketing. By applying the full model, selected model, and no-demographic variable model, we compare the results of predicted trip purpose for each land type.

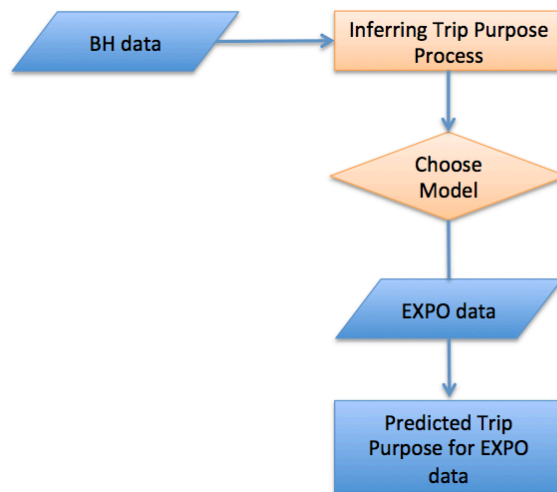


Figure 6-3: Imputing trip purpose process for EXPO data

We chose 110 participants that participated in three waves of data collection (6 months before, 6 months after, and 18 months after the EXPO light rail line was operated in Los

Angeles). Based on their GPS data collected from 922 days in this sample, there were 4,446 activities detected with 1,989 unique locations. The description of the data is shown in Table 6-2.

Table 6-2: Data description

Wave	Gender	Age	Employment	Count	Proportion
1	Female	20to29	not employed	1	1%
			part time	1	1%
			full time	1	1%
		30to39	not employed	2	2%
			part time	3	3%
			full time	6	7%
		40p	not employed	16	19%
			part time	8	9%
			full time	12	14%
	Male	20to29	part time	6	7%
			full time	4	5%
		30to39	full time	3	3%
			not employed	14	16%
		40p	part time	1	1%
			full time	8	9%
2	Female	20to29	part time	3	4%
			full time	3	4%
		30to39	part time	1	1%
			full time	3	4%
		40p	not employed	18	27%
			part time	5	7%
	Male	20to29	not employed	3	4%
			full time	4	6%
		30to39	not employed	1	1%
			full time	2	3%
		40p	not employed	6	9%
			part time	1	1%
3	Female	20to29	part time	1	2%
			full time	1	2%
		30to39	not employed	1	2%
			part time	1	2%
		40p	full time	3	5%
			not employed	15	27%
	Male	20to29	part time	6	11%
			full time	9	16%
		30to39	not employed	2	4%
			part time	1	2%
		40p	full time	2	4%
			not employed	1	2%
4	Female	20to29	not employed	8	14%
			part time	1	2%
		30to39	full time	5	9%
			not employed	1	2%
	Male	20to29	part time	1	2%
			full time	2	4%
		30to39	full time	1	2%
			not employed	8	14%
		40p	part time	1	2%
			full time	5	9%

After applying the trip purpose model chosen in the previous chapter (RF model with selected variables), the distribution of each trip purpose over land use types for each wave is shown in Table 6-3.

Table 6-3: Land use type vs purpose – Expo data

a) Wave 1

	Eat	Go home	Other	Pick & drop	Service	Shop	Social/ Rec	Study	Work	Total
FOD	82	0	1	0	0	4	0	0	0	87
IND	0	0	0	0	9	1	3	0	0	13
OFF	0	0	5	7	30	3	22	0	17	84
REC	17	0	0	0	0	1	43	0	0	61
RES	8	327	7	24	1	0	212	0	2	581
RET	1	0	0	5	0	778	6	0	17	807
SCH	0	0	1	57	1	4	14	2	2	81
SRV	0	0	2	5	79	20	13	8	34	161
Total	108	327	16	98	120	811	313	10	72	1875

b) Wave 2

	Eat	Go home	Other	Pick & drop	Service	Shop	Social/ Rec	Study	Work	Total
FOD	56	0	0	0	0	3	3	0	0	62
IND	0	0	0	0	5	0	4	0	3	12
OFF	0	0	1	12	5	1	14	0	13	46
REC	9	1	0	2	0	0	38	0	0	50
RES	2	193	2	15	0	0	163	0	3	378
RET	2	2	0	8	0	550	12	0	15	589
SCH	0	0	0	22	2	3	18	3	4	52
SRV	6	0	3	8	72	14	9	4	25	141
Total	75	196	6	67	84	571	261	7	63	1330

c) Wave 3

	Eat	Go home	Other	Pick & drop	Service	Shop	Social/ Rec	Study	Work	Total
FOD	93	0	0	0	0	4	0	0	0	97
IND	0	0	0	1	8	1	5	0	7	22
OFF	0	0	1	2	8	8	3	0	17	39
REC	0	0	0	1	0	3	35	0	0	39
RES	0	202	2	42	1	0	179	0	3	429
RET	1	0	0	10	0	373	8	0	7	399
SCH	0	0	4	22	3	2	14	2	4	51
SRV	0	0	6	5	106	15	12	1	20	165
Total	94	202	13	83	126	406	256	3	58	1241

6.5 Results

6.5.1 MANOVA results

In this section, we explore how/if the duration of each activity changed within and by: waves, day of week, type of activity, and other factors. First, the hypothesis that the two-dimensional mean-vector of activity duration and frequency of visiting the location where the activity happened is the same for three waves and seven days a week was tested by Hotelling-Lawley test in a multivariate analysis of variance (MANOVA) framework. The tests' results are shown in Table 6.4. The p-value associated with the Hotelling-Lawley statistic was very small for both tests and there was strong evidence that the mean vectors of the two variables were not the same among waves and days of week.

Table 6-4: MANOVA results

Factor	Df	Hotelling-Lawley	approx F	num Df	den Df	Pr(>F)	
WAVE	2	0.0023224	2.5784	4	8882	0.03553	*
Residuals	4443						
Day	6	0.014198	5.2497	12	8874	<0.00001	***
Residuals	4439						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1							

Analyzing the sample mean, we saw large differences in the three waves in activity duration, where a low number of visits was associated with high duration in the wave 2 and high number of visits with low duration in the wave 3 (Table 6-5).

Table 6-5: Sample mean for duration and number of visits

	Wave 1	Wave 2	Wave 3
Mean duration	67.74863	74.33511	61.88511
Mean # visits	1.551467	1.531579	1.623691

The difference in the association of activity type and activity duration among the three waves of data is described in Figure 6-6. The wave 3 data included 70% of trips for food taken by transit. The proportion of shopping trips using transit also was higher in the wave 2 and the wave 3, especially when the participant stayed for longer time at a location. The proportion of walking trips increased over the three waves for all activity types, except eating and picking up or dropping off. Social visit and recreational activities had the longest duration, and the duration increased from wave 1 to wave 3. The amount of time spent on shopping and other purposes increased in wave 3. All long-duration activities for other purposes were conducted by transit in the wave 3. The participants in the wave 1 spent shorter amounts of time for shopping if they used transit.

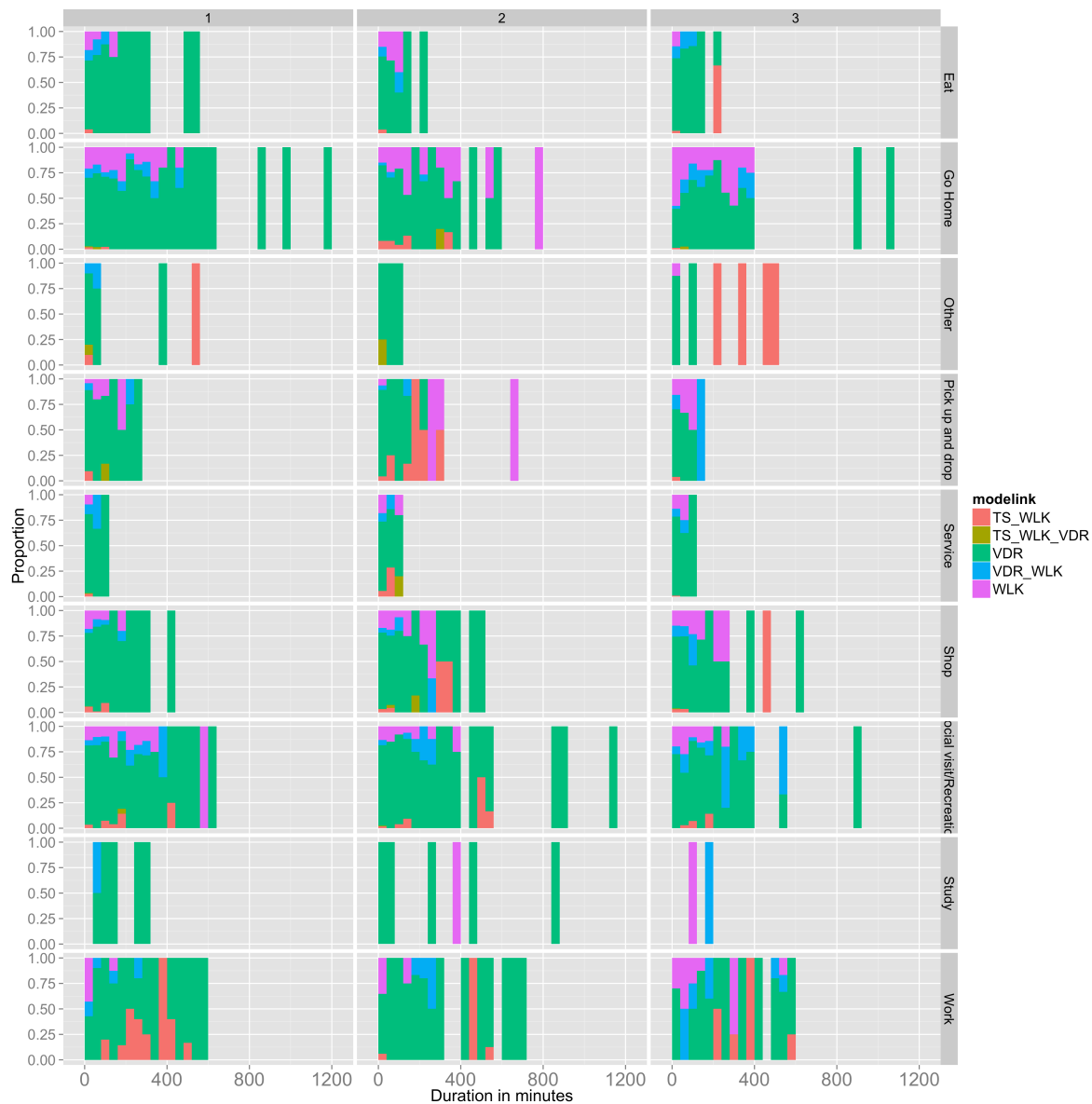


Figure 6-4: Activity duration: travel mode vs activity type

6.5.2 Multilevel model results

Table 6-6 shows the results of multilevel model for activity duration. Estimation of all of the fixed and random parameters was accomplished by the log-likelihood criterion. The fixed effects (the first part of Table 6-6)) give the means averaged over all random terms (waves, day,

and time of day). In general, if a participant arrived a location 1 hour later, the duration of activity that occurred at that location decreased by 8.1 minutes. If the participant drove car only or combined walking and driving, the duration of activity also decreased about 23 minutes. Furthermore, duration varied significantly with age—the activity duration of groups of participants older than 40 years old 21 fewer minutes than this number for the group of younger adults (20 to 29 year old).

Table 6-6: Multilevel model's components

Fixed effects	Estimate	se	t
(Intercept)	202.54	34.98	5.79
Modelink: TS_WLK_VDR	-20.08	22.40	-0.90
Modelink: VDR	-29.26	7.35	-3.98
Modelink: VDR_WLK	-29.88	8.89	-3.36
Modelink: WLK	-35.73	7.87	-4.54
Arrivehour	-7.44	0.90	-8.24
Purpose: Go Home	75.86	6.55	11.57
Purpose: Other	41.79	16.69	2.50
Purpose: Pick up and drop	-7.39	8.13	-0.91
Purpose: Service	-24.82	7.55	-3.29
Purpose: Shop	-5.87	5.99	-0.98
Purpose: Social visit/Recreation	49.96	6.42	7.78
Purpose: Study	157.83	21.65	7.29
Purpose: Work	193.35	8.83	21.89
Male	11.88	3.18	3.74
Age 30to39	0.05	5.89	0.01
Age: 40p	-9.26	4.79	-1.93
Number of car	-1.49	1.81	-0.82
Employment status: part time	-3.05	4.00	-0.76
Employment status: full time	24.63	3.42	7.20
Random effects	Variance	Std.Dev.	Proportion
Time of day (Intercept)	4742.43	68.87	0.3568
Wave (Intercept)	37.66	6.14	0.0028
Weekday (Intercept)	0.00	0.00	0.0000
Residual	8510.07	92.25	0.6403

The second portion of Table 6-6 shows the proportion of variance in temporal measurements for activity duration. The residual variance estimate for random effects can be thought of as the within groups variance, and each random effect variance estimate can be

thought of as a between groups estimate. We can see that the effect of time of day alone is significant (35.68%) and therefore not negligible. The key finding here is only 0.28% of the total duration variance of the random effects is attributed to the wave effect. The day of week's random effect is almost zero because this factor did not vary consistently across waves. These two effects can be removed. The within groups variance accounts for 64.03% of the total variance.

6.6 Conclusion

The objectives of this study are to complement and evaluate knowledge gained from data collection using GPS associated with the analysis of changing personal travel behavior over time. The temporal dimensions vary from time of day, to day of week, and wave-to-wave. This time scale can reveal the underlying local patterns such as routine or habit that typically are viewed as latent factors in single-day activity research. Activity duration was taken into the study as a measurement of travel behavior. The study found that though activity duration averages were significantly different over days of week and waves, the effect of these two factors on activity duration was minor, while time of day contributed over one third of the total variance in the duration. Previous multilevel studies have not explored these time variances at such a detailed level.

Habits and routines are deeply rooted into individual's behavior as the rhythms of daily life. There is a large body of literature in psychology use habit to explain behavior; however, the number of study in transportation has been limited. By measuring the changes of activity duration over three temporal dimensions, we expand the literature on the formation of

routines, the dynamics of major behavioral change due to upgrading the transportation system with an extended public transit network. The results provide knowledge about not only activity behavior but also can be synthesized in a micro-simulation model of travel behavior, particularly for activity scheduling model.

Although the study found interesting results related to time use behavior, these results may not represent the population since the sample was not balanced in gender (68% of participants were female), race (African-Americans comprised 56%) or income level (45% of households were low-income). Further, the study only includes activity duration as a measurement for activity/travel behavior. In the future, we will provide more detail analysis about other components, such as trip frequency and trip distance, to extend the knowledge of time use research.

CHAPTER 7: PUBLIC OPINIONS AND RAIL TRANSIT SERVICE IN LOS ANGELES; AN ANALYSIS USING TWITTER DATA

7.1 Introduction

With the explosive growth of social web services and mobile devices, individuals and organizations are increasingly using information from social networking media for decision-making. Twitter is a form of blogging that allows users to share brief text updates or micromedia such as photographs or video clips with other Twitter users. Through the service, users provide instantaneous information on their geo-location (based on GPS coordinates from smartphones), timing of where and when they provided information, and their information, observations, or opinions. Several social sectors (e.g. politics, entertainment and business) have been gathering and analyzing information from Twitter (Mittal & Goel, 2012; Tumasjan, Sprenger, Sandner, & Welp, 2010). Using information from this data source could reduce the need to conduct surveys, opinion polls, and focus groups of transit riders or interested parties because there is an abundance of information already publicly available. However, identifying opinions on Twitter and distilling the information into meaningful patterns is a time consuming task. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions from large numbers of daily Twitter posts. Therefore, there is a need for an analysis system based on text mining techniques to address this issue.

Many commuters use social media such as Twitter to communicate or report accidents, road congestion, and special events. In addition, many government organizations also use Twitter to communicate with the public and commuters. The contents of popular social media contain abundant information on transportation related issues, which could potentially be a

good source of data for transportation planners and researchers to better understand commuter and public opinion.

Although using Twitter data for opinion mining is popular in many fields, its use in the transportation planning and management sectors remain limited. In 2013, Iteris (Mai & Hranac, 2013) compared incident records from the California Highway Patrol with Twitter messages related to roadway events over the same time period. In another study, Twitter data were obtained from the riders of the rapid transit system of the Chicago Transit Authority which allows members of Twitter to monitor sentiments of transit users (Collins, Hasan, & Ukkusuri, 2013). However, these studies are still far from sufficient for public opinion mining.

Significant rail transit developments have occurred in Los Angeles during the last decade. The city will open six new rail transit lines before 2020 as the city moves further into its rail development with optimistic claims that the extended urban rail systems would increase transit ridership, draw people out of their cars, ease congestion, and improve the environment (Los Angeles County Metropolitan Transportation Authority, 2009). An evaluation program is crucial for policy makers, planners, and public to understand how efficient and effective these rail transit services are. However, different evaluation methods often lead to different conclusions concerning the value of a transit project. Taking advantage of available social media data and improvement in data mining techniques, we collected and analyzed twitter data relating to each line in 7-line Los Angeles Metro rail transit system. This chapter extends previous efforts by introducing ways to understand Twitter data in term of sentiment analysis, topic modeling, and the interaction between posters and retweeters. We demonstrate how this information can contribute to enhanced situational understanding and crisis communications, and can help reach the transformational objectives for the regional expanding transport system.

7.2 Literature review

Research about public opinion from Twitter data is not new either in industry or in academia. In the field political science, Tumasjan et al. (Tumasjan et al., 2010) collected roughly 70,000 tweets mentioning one of the 6 major parties represented in the Germany and 35,000 tweets referring to their politicians between August 13th and September 19th, 2009, prior to the German national election. To extract the sentiment of these tweets automatically, they used text analysis software developed to assess the emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. The mere number of tweets reflects voter preferences and provided a close approximation to traditional election polls, while the sentiment of Twitter messages closely corresponded to political programs, candidate profiles, and evidence from the media coverage of the campaign trail. The study demonstrated that Twitter could be considered a valid indicator of political opinion.

In another paper (Mittal & Goel, 2012), the authors applied sentiment analysis and machine learning principles to Twitter data to find the correlation between "public sentiment" and "market sentiment" in order to predict stock. The idea is that by having the Dow Jones Industrial Index (DJIA) closing values predicted one day in advance, the predicted values can be used to make intelligent trading decisions. They used only tweets that containing words more likely to express a feeling, (i.e. "feel", "makes me", "I'm" or "I am") to predict public mood and use the predicted mood and previous days' DJIA values to predict the stock market movements. The moods were classified as 6 standard Profile of Mood States (POMS) moods: Tension, Depression, Anger, Vigor, Fatigue and Confusion. Then, a simple word counting algorithm was applied to find the score for every POMS word for a given day. With a specific test period, the

research gives a 75.56% Directional Accuracy (compares the forecasted directions (up/downward movement) to the realized directions (up/downward movement)).

In a transportation application, Iteris compared incident records from the California Highway Patrol (CHP) with Twitter messages related to roadway events over the same time period (Mai & Hranac, 2013). Collected tweets appeared to correlate with CHP traffic incident records by keywords such as “accident”, “crash”, “traffic”, “road”, “freeway”, “highway”, “lane”, “wreck”, “car”, and “cars”. Findings indicate that tweets can be matched to traffic incidents by examining the content of the tweets for key words and comparing the timestamps and locations of the tweets and incidents. When compared broadly, it appears that tweets tend to be posted within 5-hours of the incident that they refer to, and are most often sent from locations within between 10 and 25 miles of the incident’s location.

Relating to transit service, Collins et al. (Collins et al., 2013) collected Twitter data from the riders of the rapid transit system of the Chicago Transit Authority (CTA) using Streaming API method, by tracking tweets containing the names of each line. They used SentiStrength (<http://sentistrength.wlv.ac.uk>), a machine learning program that detects the sentiment value of a short text, for analyzing the sentiments of rider tweets about the transit system. Output from the software can document the average negative and positive sentiment or the minimum negative or maximum positive sentiment. SentiStrength contains an original collection of 298 positive terms and 465 negative terms classified with relative sentiment strength. Ratings of 1 to 5 indicate a positive sentiment and -1 to -5 indicate a negative sentiment, with -5 and 5 being the maximum negative and positive sentiments that can be attained. A comprehensive list of emoticons annotated with sentiment values was incorporated in the analysis. The analysis showed that transit riders are more inclined to assert negative sentiments to a situation than

positive sentiments. In addition, findings indicate that the rider dissatisfaction related to incidents can be quantified through the sentiment analysis of the social media data.

In addition, Schweitzer (Schweitzer, 2014) analyzed 64,000 tweets regarding various modes of transportation and transit services in USA to come up with a look at how well public transportation ranks with the social media generation. Tweets' sentiment was estimated with an established lexicon of positive and negative English and Spanish words developed by computer scientists (Hu & Liu, 2004). In this study, the author discovered that these comments reflected more negative sentiments about public transit than did the comments about most other public services. However, in term of the content of the complaints, there were a much higher proportion of negative statements about fellow transit passengers than complaints about the quality of service. The results also showed that social media provided strategic opportunities for public agencies to improve customer service and image. When transit agencies responded directly to users, as opposed to merely "blasting" announcements, they had more positive statements about all aspects of services and fewer slurs directed at passengers. When the agencies chatted with users, they also had significantly more positive sentiments expressed about them on Twitter.

Though the above studies suggest the promising potential of using Twitter data in transportation planning and management, many methodological challenges remain to maximizing this valuable and rich source of information. The current research seeks to extend and clarify knowledge of mining tweets relating to public transit in order to develop a documental scheme that could be applied to any public transit system or transportation-related issues. In addition to using established methods for analyzing public opinion such as sentiment analysis, we demonstrate new methods which mine aspects of opinion that have not previously

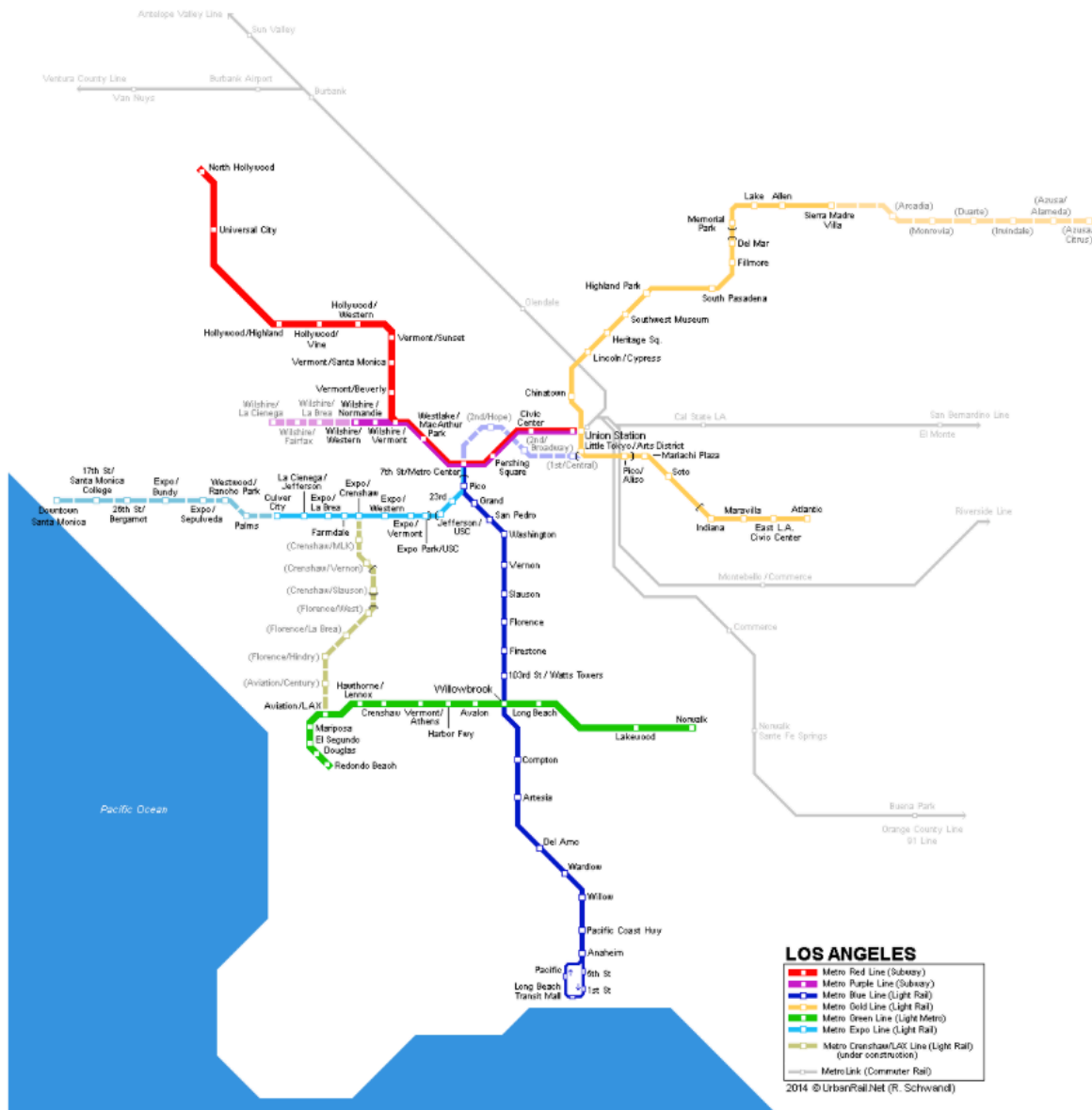
been evaluated in transportation field. Understanding how tweets might be amplified across a social network through retweeting has not been adequately studied. Therefore, we demonstrate networks analysis techniques which can detect how a tweets' content may have produced a change in information flows, between those who posted and those who retweeted for the whole Metro system and for specific rail transit lines in Los Angeles. Understanding patterns in periods of high system stress, such as during crisis such as train incidents or emergencies that cause delay, could be particularly helpful for transit agencies. Finally, we applied text-mining techniques to identify the opinion of rail transit users in order to support Metro in monitoring real-time feedback and sentiment along different lines in the area's rail transit system.

7.3 Methodology

7.3.1 Data collection

We collected tweets from May 1st 2014 to October 27th 2014 using the Twitter REST API, with radius 50 miles around Los Angeles using name/account's name of seven rail transit lines (Figure 7-1). REST APIs are based on the REST architecture, which use the pull strategy for data retrieval. After cleaning data, there were 13,392 tweets in the final analysis dataset, including 5,693 tweets that were retweeted. A Retweet is a re-posting of someone else's Tweet. Twitter provides only limited account information (e.g., location). In the data collected, about 10% of tweets had specified locations that could be parsed correctly into valid latitude and longitudinal coordinates for the study region. By mapping each tweet's longitude and latitude, we can see that the geographical distribution of the data is aligned with Los Angeles's rail

transit network (Figure 7-2). The darker areas in the map represent areas from which more tweets originated.



Source: <http://www.urbanrail.net/am/lsan/los-angeles.htm>

Figure 7-1: LA light rail network

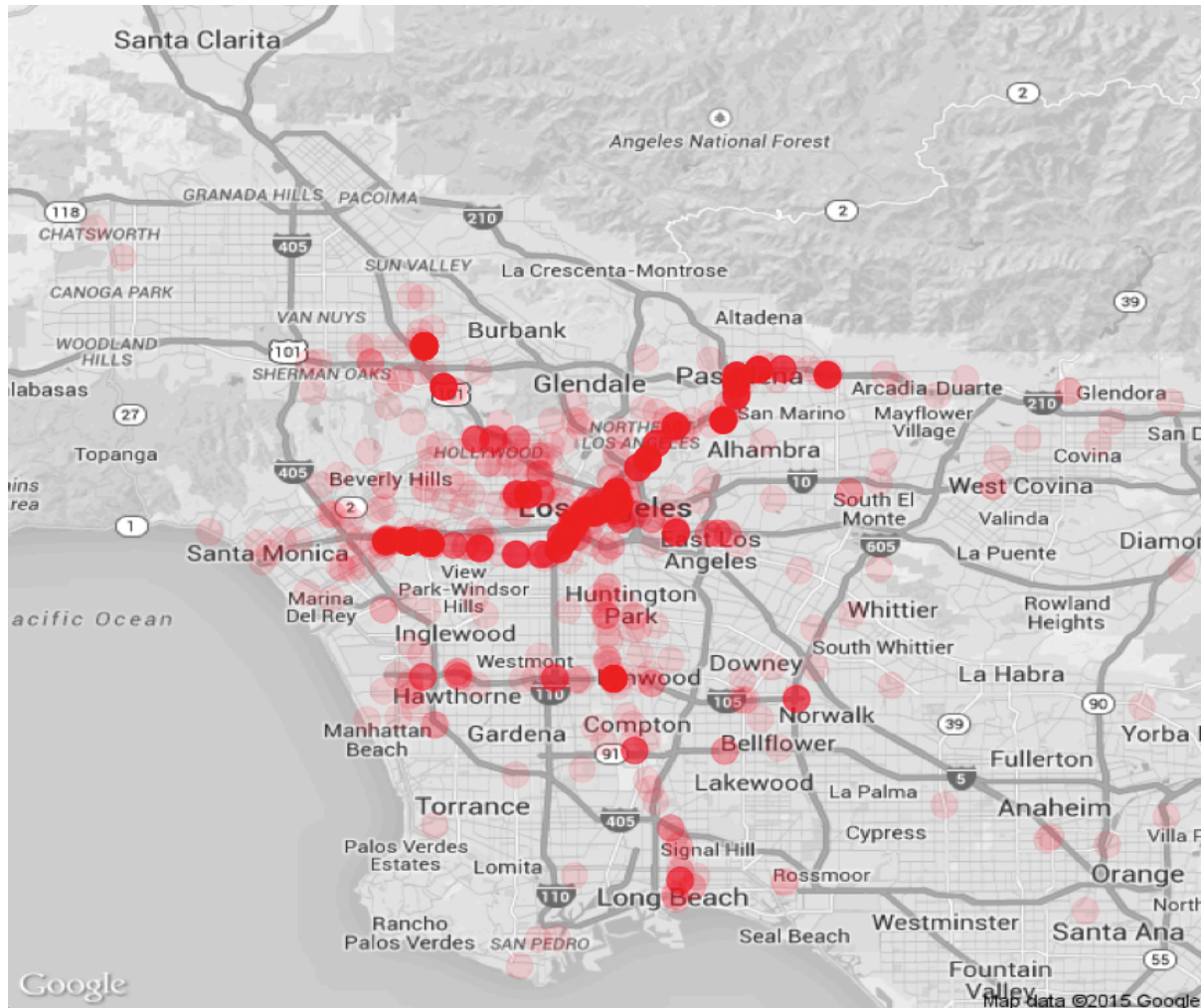


Figure 7-2: Geographical distribution of tweets

The architecture of data collection and analysis is outlined in Figure 7-3. First, the Twitter search is filtered by using topically relevant keywords, language filters, and geographic information. Then we use a list of criteria to identify the theme and estimate the sentiment score of each tweet. We store the tweets' text and metadata in a database for later analysis and visualization.

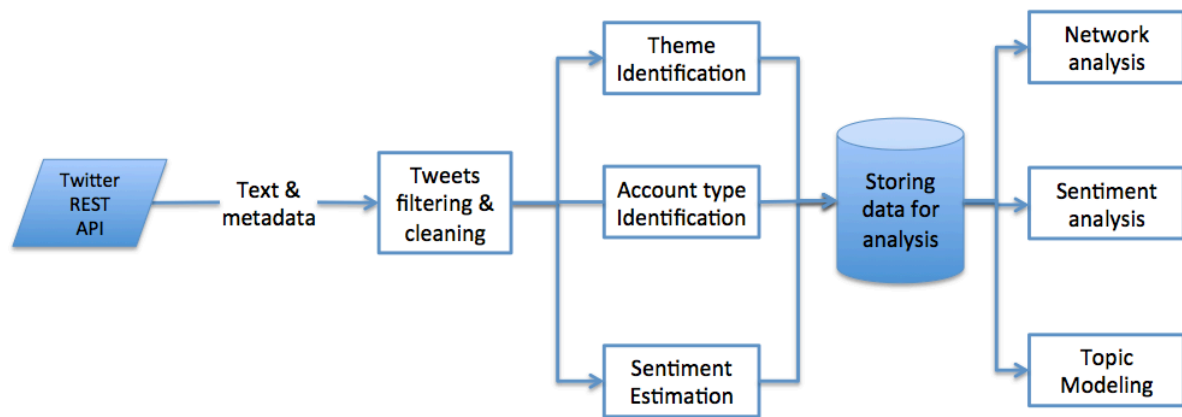


Figure 7-3: Scheme of Twitter data analysis

7.3.2 Data processing

Twitter data exists in a noisy structure. To clean the data, we use three processes: removed stop words, stemming, and emotion/emoji indicator extraction.

- ***Removing stop words***

The first step in preprocessing tweets is removing stop words. The stop words are words that do not add meaningful content to the data set (i.e., pronouns, prepositions, conjunctions, etc.). Removing them reduces the space of the items significantly in the text, prior to do topic modeling. Examples of such words include ‘the’, ‘a’, ‘of’, ‘and’, and ‘to’. We also add the name of each rail transit line as stop words.

- ***Stemming***

Stemming—the process of removing prefix and suffix leaving the stem or the root of the word—is a fundamental technique for processing large amounts of data in information retrieval and text mining. We chose the approach that stemming a word to the shortest version that appears in the whole data set. To apply this process, the Porter stemming algorithm (Porter, 1980) was applied. The Porter stemming algorithm (or ‘Porter stemmer’) is a process for removing the commoner morphological and inflexional endings from words in English. For example, the words “close”, “closed”, and “closing” would be reduced to the root word “clos”.

Since the stemming process cannot be directly reverted, it could make the short root words difficult to be interpreted after stemming. We implemented another step to complete the words after the stemming process by restoring them from a dictionary, which is setup as the whole data set’s texts before stemming. To find completions for individual stems along compatible semantic groups, we used the algorithm previously proposed by Feinerer (Feinerer, 2010). We used “shortest” method as the completion heuristic by choosing the completion word with the minimal number of characters of all suitable completions for a given root word term. For example, from the list of completions for “clos”: “close”, “closed”, and “closing”, “close” is chosen. Both the stemming and stemming completion process were done by using the source codes created by Feinerer (Feinerer, 2013).

- ***Emotion/Emoji indicator extraction***

Because tweets are short (at most 140 characters), Twitter users often use emoticons and emoji to express their emotion in a shorter way to save space and have more fun.

Emoticons are textual portrayals of a writer's moods or facial expressions in the form of icons. Emoji literally means "picture" plus "character", originally the ideograms or smileys used in Japanese electronic messages and webpages but it is used extensively outside of Japan nowadays. These pictorial representations of feelings play a significant role in communication. However, identifying these characters is different because it doesn't have a unique form or format as text. Therefore, we built a dictionary of emotions and emoji, and translated them into two layers of sentiment value. The first type is the emotion (happy, anger, etc.), the second one is sentiment (positive, negative). For example, an emoticon "☺" or "🙂" means happy and it is positive. While "☹" or ":((" means sad and it is negative.

7.3.3 Theme identification

We are interested in understanding how tweets' content might affect online behavior in terms of information transmission. Therefore, basic themes and user account classification were coded for each tweet and tweeter. We used exploratory data analysis to identify salient themes in the data. Our coding scheme includes 4 types of themes for rail transit service: Delay, Resume, Information, and Checkin. Delay tweets contained words relate to the disruption of the service, such as "delay", "cancel", or "suspend". Resume tweets included words such as "resume" or "reopen" referring to general updates about the status of the service. Information tweets were tweets relate to "construction", "maintenance", or "reminder", etc. that could inform or remind users about some events ahead of time or some real-time events that do not relate to Delay or Resume. Checkin tweets were tweets relate to location of users, were identified by key words like "I'm at" or "checked in". The rest of the data was classified as

“Other” theme because they were not identified by delay, resume, information, and checkin key words (Table 7-1). To identify specific themes of these Other tweets, we used techniques summarized in topic modeling section. The description of these themes is as following:

Table 7-1: Theme of tweets

Theme	Delay	Resume	Information	Checkin	Other
Number of tweets	1533	971	2281	208	8383

Because tweets are very short texts, the maximum number of characters for each tweet is only 140. Within the 13,392 tweets we collected, the average length of each tweet is 13.94 words, after filtering out usernames, hashtags, punctuation, links, emoji, and emoticon. The distribution of word count by theme category is shown in the Figure 7-4.

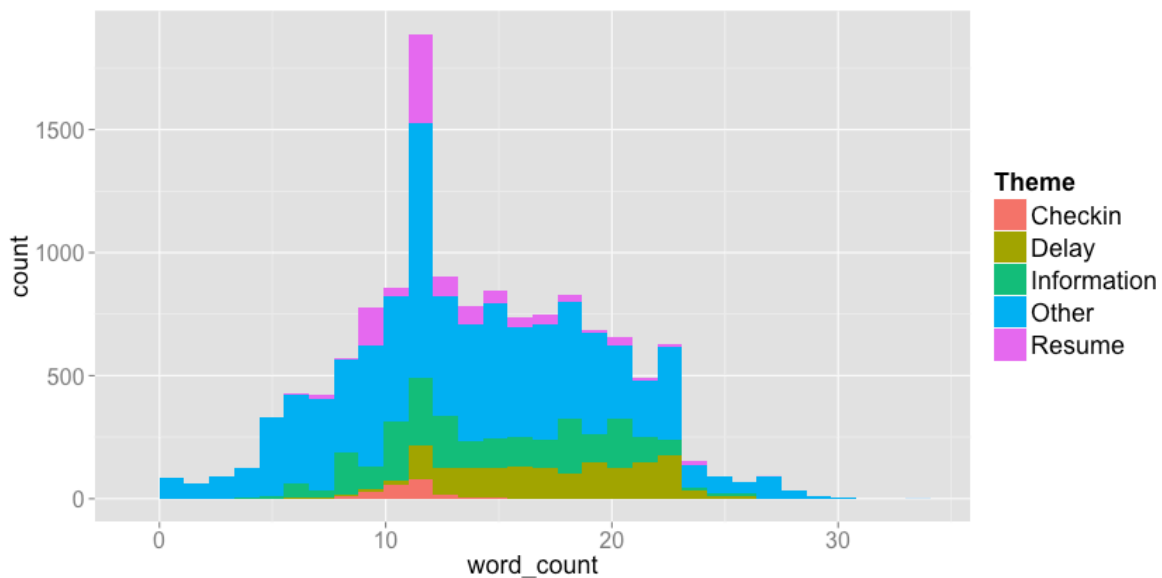


Figure 7-4: Word count distribution

We also distinguished the type of each Twitter account appeared in the data set base on the account's summary and content listed on the time line of the account. There were 12 types of accounts that we use in this study (Table 7-2).

Table 7-2: Distribution of Twitter accounts

Type	Account	Tweets	Type	Account	Tweets
User	4591	9757	Business	8	13
Light rail agency	8	1561	Events	8	20
Transportation relates	48	798	Other organization	1	1
Urban relates	14	42	Sport	4	22
Service	20	42	School	10	30
Government	18	64	News	129	1042

The 7 rail transit lines were clustered into groups based on their operation or construction status in the data collection period. The LAX line (Crenshaw line) was under construction. The Red line, The Blue line, and the Green lines were established lines in normal operation and no active extension activities. Segments of the Gold, Expo and Purple lines were in normal operation during the study period, but extensions of the Gold and Expo Lines were under construction and the extension of the Purple line was in the pre-construction phase.

7.3.4 Sentiment estimation

Sentiment analysis is one of the most active research areas in the natural language processing field. Sentiment analysis mainly studies opinions, which express positive or negative sentiments. The most important indicators of sentiments are sentiment words. For example, good is a positive sentiment word, and bad is a negative sentiment words.

Sentiment analysis is a form of data mining used to interpret textual and verbal information. By parsing human verbal communication through natural language processing, applying computational linguistics, and deploying text analytics, subjective information (i.e., attitudes, opinions, and intentions) in source materials can be identified and extracted for more intense processing and scrutiny.

Using a dictionary to compile sentiment words is a popular approach. Hu and Liu (Hu & Liu, 2004) collected a set of sentiment words (seeds) with known positive or negative orientations. We used this word collection for sentiment analysis. There are 4,783 positive words and 2,006 negative words. A search algorithm was applied to identify which words in a tweet match which sentiment words. Positive words were assigned a value of “+1”, while negative words were assigned a value of “-1”. If a word was not in the list of positive or negative words, its value was “0”. Sentiment score of a tweet was produced by summarizing the value of each word in a tweet. A tweet was defined as negative if its score was less than zero, neutral if equal to zero, and positive if more than zero.

7.3.5 Network analysis: “who retweet whom”

Retweeting is a popular behavior in the Twitter platform, described as when an individual engages in and distributes information by retweeting a message from other Twitter users. Retweeting thus has important implications for message amplification as it provides organizations and individuals with a wider audience to share information. This application is potentially important for transportation service providers because it can more directly align service information provided directly to users’ time use and schedule. In the transportation

field, there has been little attention dedicated to how Twitter information is passed among transit users depending on its content. We identified a retweet as a tweet that includes “RT @” or “via @”. There are 5,706 tweets in the data set which were retweeted.

By retweeting a tweet, all of the users’ connections are able to see the message, therefore the information reaches more people. To analyze the flows of information through retweeting, a graph of the relationships between tweet posters and retweeters was created, in which each edge connects two accounts that appear in the same tweets, one from the poster, and the other from the retweeter.

7.3.6 Topic modeling

Because more than a half of tweets in the data set were not assigned a theme (Delay, Resume, Checkin, or Information) and were therefore classified as Other, we applied topic modeling to identify the topics included in the remaining tweets. Topic models provide a method to infer the themes present in tweets, and to represent them as clusters of words that tend to appear in similar contexts. This technique provides more information about the tweets rather than manually checking or searching using keywords. The method allows detecting correlations that extend individual words because using only key words or the term frequencies (TF) neglects the information contained in word ordering (Gruen & Hornik, 2011).

In machine learning and natural language processing, topic modeling is a type of statistical modeling which uses probability of TF occurrences in documents (Gruen & Hornik, 2011). TF is a value indicating how frequently word ‘A’ appears in a specific document, while document frequency (DF) is the number of documents that word ‘A’ appears in, inverse

document frequency (IDF) is the inverse value of DF. To give the relative weight to TF more than to IDF, a log function is applied to the latter.

The TF-IDF is a type of weighting used in information retrieval and text mining to remove noise from data, which is obtained by multiplying the IDF and TF. We used TF-IDF to remove the most frequent and the scarcest words before extracting topics from tweets. The weighting indicates how important a certain word 'A' is in a specific document among all words in the document. The higher the value, the more important the meaning of word 'A' has in a specific document.

To extract the topics from the Other tweets, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) was used in an unsupervised fashion. LDA treated each tweet as a mixture of latent topics. Each topic was itself a probability distribution over words, and the words in the tweet were viewed as being sampled from this mixture of distributions. The LDA model (topic distributions and mixtures) was trained from all available tweets in the corpus (a collection of written texts) using the package "Topicmodels" in R (Gruen & Hornik, 2011).

To provide features for the model, we used an n-gram model. The n-gram is a contiguous sequence of n words they appear in texts which are used as features in representing vector space model (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014). Each feature corresponds to an n word, with the value of true if that feature is present and false otherwise. "N" in n-grams corresponds to the number of elements in a sequence. In this work, we explored the usage of bigrams—a pair of words or $n = 2$ —as features.

7.4 Results

7.4.1 Flow of information through networks

Nearly half of the 13,392 tweets collected between May 1st 2014 and October 27th 2014 regarding a rail transit line in Los Angeles were retweeted. The distribution of retweets related to checkin, delay, resume, and information is shown in Figure 7-5. The X-axis shows the time of day from early morning to evening, while the Y-axis represents the corresponding sum of frequency of retweet count for each tweet. The observed behavior from each type of rail transit line (constructing, in use, in use and extending, and in use and pre extending) and from each topic of tweets are different. We had more retweets about delay from the rail transit lines that operated normally only. If the rail transit line were extending, it had more Information retweets, particularly during the weekday evening period.

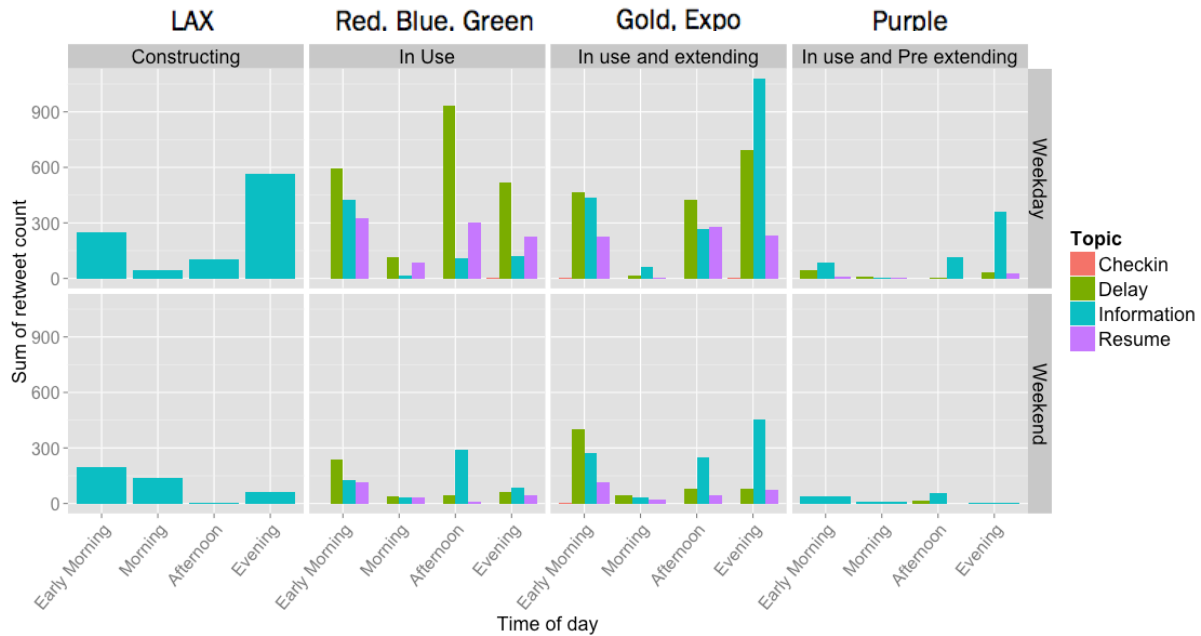


Figure 7-5: Retweets distribution, by rail line category and days of week status

To analyze the impact of the above factors relative to the number of retweets, a regression analysis was performed (Table 7-3). The results show that the number of retweets varied significantly by the content of the tweets, day of week, and time of day. Tweets that were posted on weekends had fewer retweets than those posted on weekdays. The same behavior happened for tweets that were posted in the morning. Tweets about delay had significantly higher retweets than other tweets. Although the result was not statistically significant for the impact of line's status to the number of retweets, it shows that if a line was in service and an extension was not yet initiated, it had fewer total number of retweets than other line (-108 tweets). If a line was in use and extending at the same time, it was associated with more twitter attention. The number of retweets for this kind of line was the highest (+108 tweets).

Table 7-3: Regression results

	Estimate	Std. Error
(Intercept)	27.376	81.845
Status In Use	77.403	68.616
Status In use and extending	108.452	68.619
Status In use and Pre extending	-107.854	70.127
Topic Delay	267.856***	51.636
Topic Information	263.983***	50.991
Topic Resume	141.232***	52.12
Timeofday Morning	-174.656***	47.986
Timeofday Afternoon	-35.862	46.554
Timeofday Evening	19.282	46.914
Dayofweek Weekend	-144.6***	33.844

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: 157.1 on 79 degrees of freedom
Multiple R-squared: 0.5001, Adjusted R-squared: 0.4368
F-statistic: 7.904 on 10 and 79 DF, p-value: 1.141×10^{-8}

Using our collected tweets, using graph visualization we found that individuals retweeted from other individuals the most, then transit agencies (Figure 7-6). This kind of analysis and visualization provides a quick overview of the type of relationships between posters and retweeters. Each edge represents a pair of a poster and a retweeter. An edge contains a source node, the poster's account type and a destination node (the end of the arrow), and the retweeter's account type. The results show that the focus of most retweets was the transfer of transit-related information from person to person. Therefore, in order to spread information or promote public transit, the agencies could utilize Twitter as a social marketing tool because of its incredibly powerful virality. The information flows also differed between content of tweets (Figure 7-7). While Checkin tweets had the fewest retweets, tweets including

such information as construction and announcements had the most retweets and the widest network of information flow.

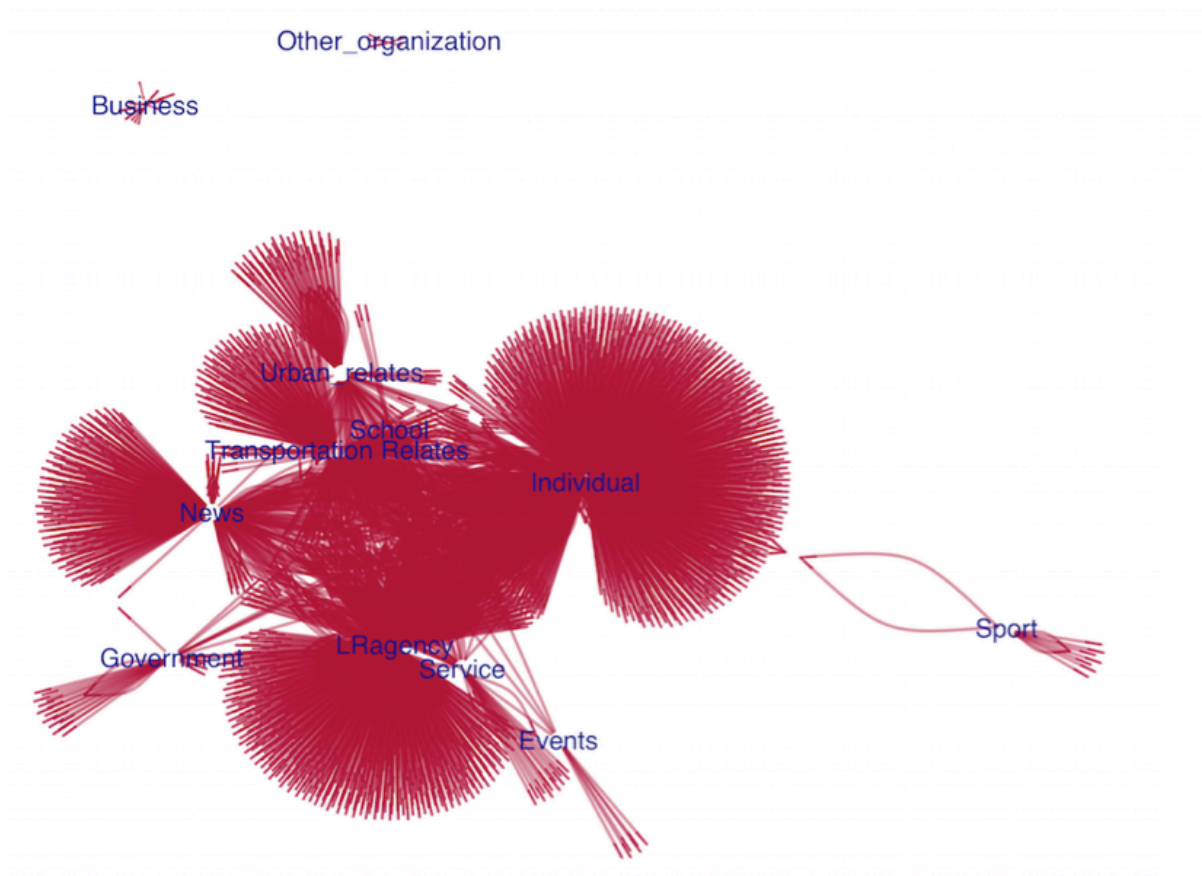
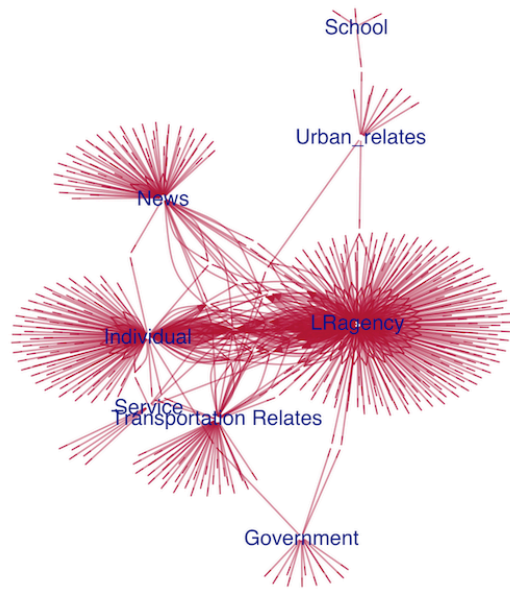
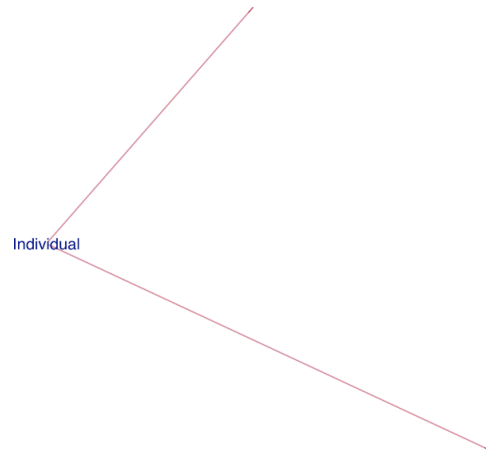


Figure 7-6: Who retweet whom

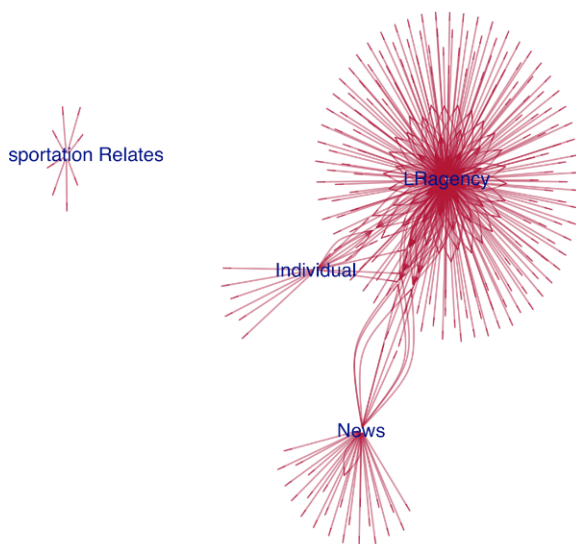


a) Information

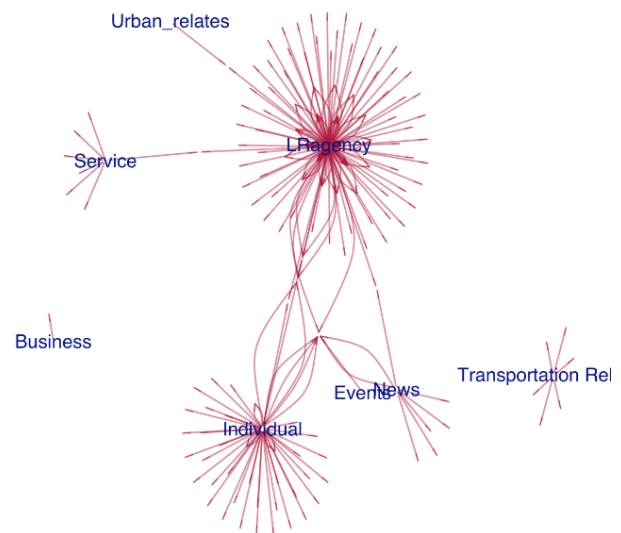
Business



b) Checkin



c) Delay



d) Resume

Figure 7-7: Who retweet whom by Theme

7.4.2 Sentiment analysis

Sentiment analysis mainly studies opinions, which express positive or negative sentiments. A tweet may contain information about habitual behavior (such as “check in” at a station or on a rail transit line), might show how angry or disappointed he/she was with the service, or may just be a retweeted of content from the rail transit agency about an incident. Using only the favorite count (the number of time a tweet is saved as a favorite) of each tweet to classify sentiment is likely insufficient. For example, a tweet marked with more “favorites” by users may not represent a positive sentiment. A tweet marked with fewer “favorites” is not necessary represent a negative sentiment. Therefore, sentiment extracted from content of a tweet may contain more precise information about a user’s preference about the rail transit line.

After translating emotion icons, emojis to words, and pre-processing the raw text of all tweets in the dataset, each was assigned an overall score based on the total score of each of its words. For example:

Original tweet:

“@ignoramity it shouldn't be. I take the blue line to HP here and there so SG shouldn't be too hard to get to. 😞”.

And here is the tweet after processing:

“it shouldnt be i take the blue line to hp here and there so sg shouldnt be too hard to get to positive”.

The average sentiment value for each rail transit tweet line is shown in Figure 7-8. This type of information can help transit providers and government and policy makers understand

the condition or status of each rail transit line from the commuters' perspective. For example, in the period of analysis (5 months), we found that the Red line was associated with the most positive tweet sentiments while the Green line has the most negative tweet sentiments.

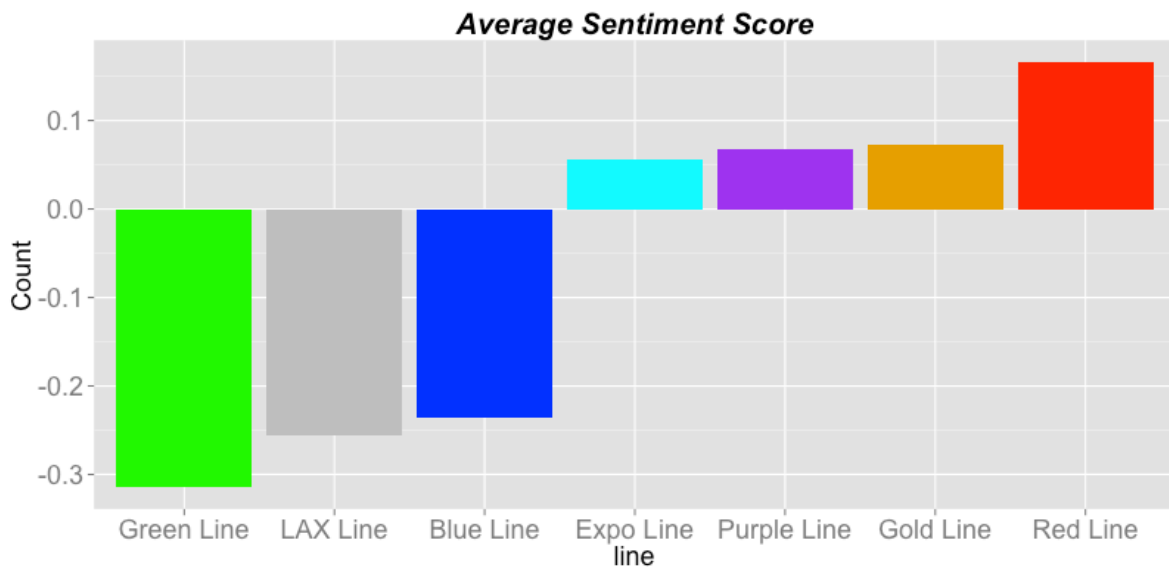


Figure 7-8: Average sentiment score by rail transit line

Also, for each rail transit line, we conducted sentiment analysis along a temporal dimension, to identify when riders had the most negative or positive opinions about the service. In this section, we show tweet distributions by sentiment and time of day for the Blue line and the Red line. In the morning, while the Blue line had the highest proportion of negative tweets (44%), this number in the Red line was the smallest proportion (18%), compared to the other segments: early morning, afternoon and evening (Figure 7-9). Figure 7-10 shows that, for any day of week, the Blue line always had a greater proportion of negative tweets than positive tweets (up to 2.5 times on Friday). Meanwhile, on the Red line, the proportion of negative tweets was equal (Wednesday and Friday) or slightly higher (Thursday) than the proportion of

positive tweets. In the other days of week, the proportion of positive tweets was always higher than the proportion of negative tweets (up to 3.9 times on Monday).

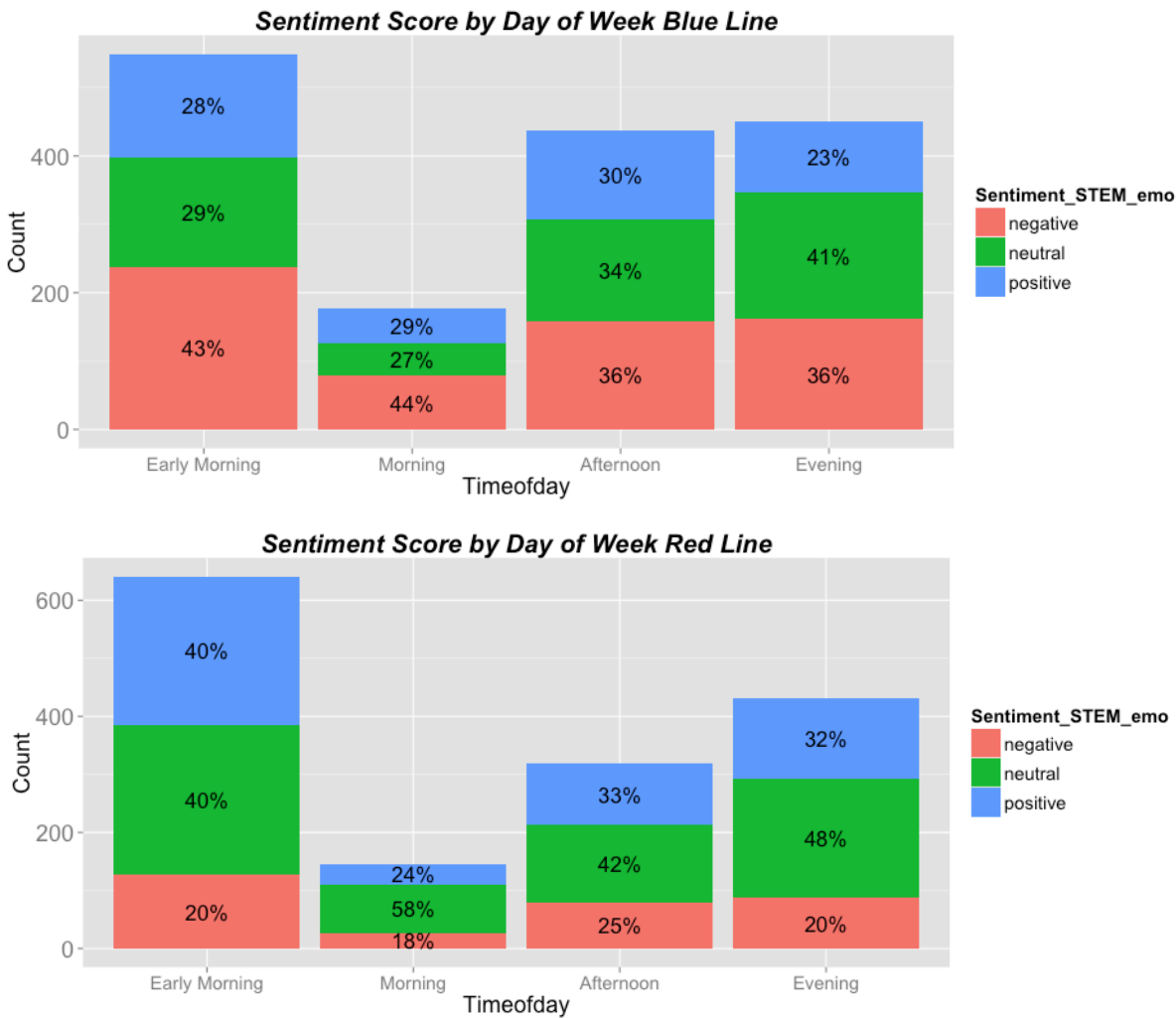


Figure 7-9: Sentiment by time of day

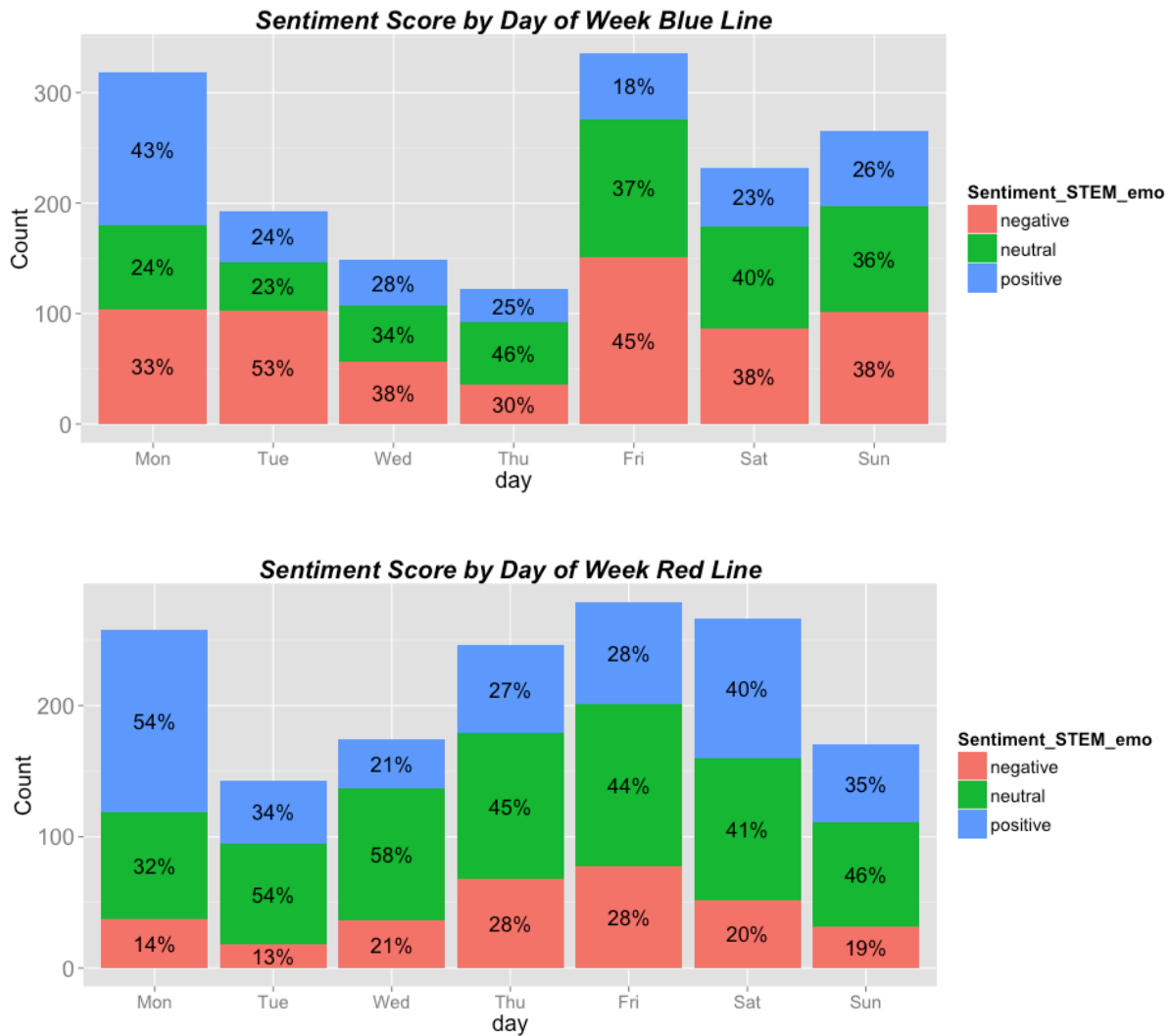


Figure 7-10: Sentiment by day of week

7.4.3 Topic modeling

We analyzed the topics of tweets for particular lines to help explain why a particular line had a negative or positive sentiment score. This technique can help clarify why people like/dislike a particular segment of service. First, word-clouds were used to reveal the high degree of relevance of user comments to the topic of concern. The word-clouds below illustrate the type

language model for predicting the next item in such a sequence in the form of a $(n - 1)$ -order Markov model. This approach allowed us to derive the number of connections a word has with other words and we graphed these patterns to visualize how words and tweets are connected as a network of terms. The colors and connections colors reflect the number of connections (degrees) each term has with other terms. Thus, we see definite terms that occur together (more connections-degrees) in tweets with larger labels and darker edges (lines). For example, Figure 7-11 shows that with the Blue Line, the frequency of words such as “kill”, “die”, “fatal”, “beat” are higher than the other words.

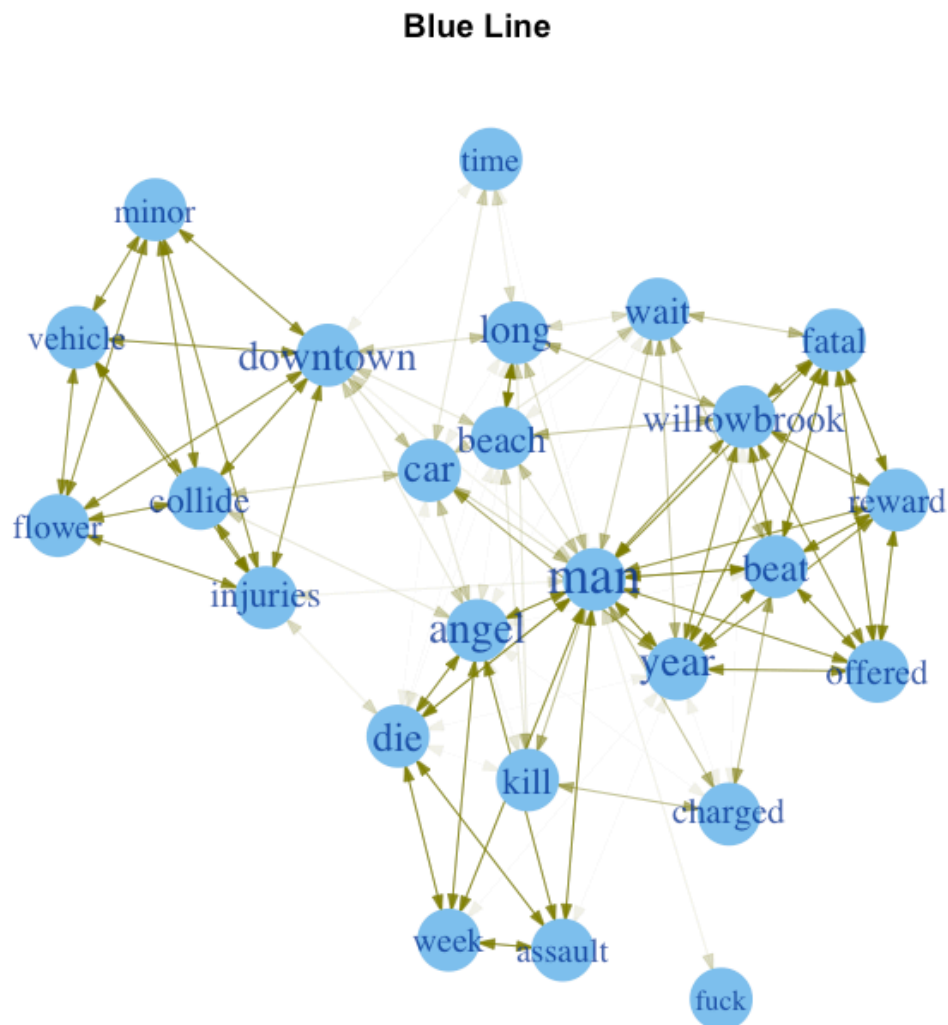


Figure 7-12: Network analysis of term for the Blue Line

Topic modeling also was used to infer the themes present in tweets, represented as clusters of words that tend to appear in similar contexts. We chose 10 as the number of topics for the model to learn. The LDA topics were constructed using the corresponding filtered versions of the tweets. Extremely scarce frequent terms were removed using TFIDF as a measurement; only words within interquartile range were used. The first 20 words (highest frequency) in each topic are shown in Table 7-4, which is based on tweets for the Blue line.

Because topic models only provide lists of n-grams, they do not automatically generate the title for each topic. To get the titles, we manually assigned subject labels to these topics based on the words contained in the topic. For example, the first topic in Table 7-4 can be named as “first time taking the light rail line”, the second topic is “motorist dead”.

Table 7-4: Topic model result for the Blue line

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	i was	blue line	blue line	blue line	blue line	blue line	blue line	long beach	metro blue	blue line
2	and train	line train	metro blue	to get	take the	line and	go to	pico station	line station	the metro
3	line at	the train	kill in	to be	line to	red line	need to	trip to	at metro	woman charged
4	with metro	to blue	offered in	collision with	i hate	im on	fatal beat	to long	los angel	charged in
5	a photo	wait for	line pico	green linemetro	station on	old man	off the	and i	man die	man at
6	have been	want to	the na	in collision	the train	year old	wait for	post a	a week	at blue
7	i take	in compton	look at	na minute	the way	blue linemetro	when i	a train	after assault	beat death
8	is a	on their	as fuck	take the	a year	imperialwilmington rosa	line lol	artesia blue	assault at	death of
9	next to	at am	go on	to see	man wait	pink between	a blue	i dont	week after	da say
10	the first	in fatal	is a	to take	the fatal	ride the	artesia blue	miss the	angel man	knuckle da
11	am on	line like	line always	first time	arrest for	you never	commute home	at in	die a	of yo
12	angeleno stuck	line pico	scene of	head to	women arrest	against rail	i need	grimm is	fatal beat	use brass
13	been wait	motorist dead	way home	line metro	at me	del amo	line station	hell as	beat of	brass knuckle
14	else was	never get	a fuck	on my	beat of	do you	ride the	holmes avenue	hate the	in long
15	escamilla from	you can	be offered	train at	linemetro blue	for blue	to make	line head	a great	on blue
16	everyone else	a minute	by bike	amo station	metro artesia	for car	beach and	line run	beat at	station use
17	felt like	a weekday	by dieter	check out	min btwn	get on	florenc area	line up	than the	yo man
18	green line	back to	from my	downtown la	na min	i just	friend in	now it	took the	line kill
19	have to	be offered	get a	in on	on time	i know	in hp	on my	in metro	line na
20	hot tonight	big blue	honor for	look for	reward to	it na	make more	red line	is a	my house

7.5 Discussion and conclusion

Expanding public transit has become an important strategy for reducing vehicle travel and associated emissions. Consistent with this trend, Los Angeles County's main transit agency has committed funds for several new rail transit lines and segments scheduled to open by 2019. With this rapid urban and infrastructure growth, government and agencies need to manage the services in the way that maximizes quality of life and comfort for the city's inhabitants. Gathering and analyzing users' feedback and opinion are one of the tasks.

The research presented in this chapter demonstrates examples of integration of semantic analysis with the real-time nature of Twitter, and presents potential uses for Twitter data in public transportation operations. While the Twitter data do not yield individual travel patterns, they do associate geographically specific locations with activities and sentiment. In particular, the data are more can represent the real-time reaction than most survey methods and provide extensive volume of collected data. They are potentially an important, and lower-cost data source than a survey data. This chapter presents an innovative social approach that has not been reported previously in the literature and provides valuable feedback to local authorities, so they can improve the commuting experience, both for individual commuters and from a system point of view.

Using text-mining techniques, we identified differentiated insights about comments, criticisms, responses, and shared information of transit users and providers

about rail transit in Los Angeles. The study found that information flow analysis, sentiment analysis, topic modeling, and its application are indeed a practical and applicable tool for measuring commuters' emotions and opinions. This can be useful for understanding trends among commuters and how their emotions changed according to the rail transit line they used, the time of day, and the day of the week. Such social media can become a platform for commuters to not only access information but also to share their feelings and opinions about their commutes.

In the future, to further analyze the topics, we could set up the higher number of topics to be discovered. The bigger the number is, the more fine-grained the model can capture specific patterns. The methods we developed in this study could also be applied to analysis for other areas of service operations, which could leverage technology to help government manage information and facilitate interaction in meaningful ways in order to achieve broader public participation than is possible through normal channels. Analysis of social media data can provide access to segments of the users that have not participated in traditional ways.

CHAPTER 8: CONCLUSION

As cities across California pursue Senate Bill 375's directive to reduce greenhouse gas emissions, transit-oriented development has emerged as a main strategy to encourage transit use. Significant light rail transit developments have occurred in Los Angeles during the last decade. The city will open six new rail transit lines before 2020 as the city moves further into its rail development with optimistic claims that the extended urban rail systems would increase transit ridership, attract people out of their cars, ease congestion, and improve environment. To evaluate such projects and monitor changes in mobility, travel behavior needs to be investigated dynamically. Observed behavior and attitudes change not spontaneously but gradually after a system changes, sometimes involving overreaction. An evaluation of a new system or system improvement may be imprecise or even erroneous if it is not based on an understanding of time lags involved in individuals' responses.

Today, collection of data is not only high-volume and speed, but also high-variety in content. The data sources include sensors, mobile devices, and social media. We proposed to take advantage of availability of data, both structured and unstructured, from passive data, questionnaires, to social media data to analyze the underlying patterns and trends, and use the results to support enhancements in planning and programming. Raw data has little value for planning policy evaluation purposes. Exploring the insights contained in the data requires processing. For example, identification of travel mode and trip purpose is extremely important in processing GPS

data. The goal of this dissertation is to develop statistical reasoning techniques for human activity patterns recognition and public opinion extraction.

First, a framework of activity recognition is built upon, and extending, existing research on GPS segmentation and machine learning. This framework is able to take into account complex relations between locations, activities, and transition periods between places, as well as knowledge of homes, workplaces, and other frequently visited places. By segmenting and labeling activities and significant places based on activities and places' features, this approach achieves high accuracy both in extraction and labeling. Further, the segmentation method is fast and accurate when performing the feasibility of transferring knowledge from labeled data to unlabeled data. The model can infer people's travel mode (such as walk, bike, bus, rail, or car) with or without general knowledge such as street maps, bus routes, and bus stops, which can reduce the cost of data acquisition and time for processing. The only initial input for the framework is speed.

Second, in this study, we demonstrated and evaluated several methods of automating prediction of trip purpose. Multiple machine learning algorithms were employed, validated, and tested to identify the most suitable model, i.e., the one that has the highest predictive power at both class and overall level. We take advantage of other data sets with available participant-derived trip purposes to train the model. Estimation results show that, overall, RF is the best model. The overall accuracy for prediction is 81.12% on the testing set. The model also can predict accurately shopping trips with the sensitivity of 92.1%, go home and studying trips reach 100%. An

interesting result is that the accuracy of the model even increased to 81.75% when we did not include demographic variables. In the literature, this phenomenon is referred to as the curse of dimensionality, where a high number of variables can lead to lower classification accuracy (Powell, 2011). The reason could be that the data also becomes increasingly sparse in the space it occupies when the number of variable increase. The other reason may be that the available training data is too small, there may be too few data objects to allow the creation of a reliable model for assigning a class to all possible objects (Janecek, Gansterer, Demel, & Ecker, 2008). This implies that without socio-demographic variables, the trips' characteristics alone can predict the purpose of trip accurately. This conclusion is important in practice, when applying to large-scale trip purpose imputation, where participants' personal information is unavailable. The model's results also provide the clue that land use type of a location is the strongest indicator for imputing purpose of trips.

Third, using the travel modes and trip purposes derived and inferred by the proposed framework, we aimed to complement and evaluate knowledge about 'how personal travel behavior changes over time. The temporal dimensions considered vary from time of day, to day of week, and wave-to-wave. This time scale can reveal the underlying local patterns such as routine or habit that typically are treated as latent factors in single-day activity research. Activity duration was taken into the study as a measurement of travel behavior. The study found that although activity duration averages were significantly different over days of week and waves, the effect of these two factors on activity duration is minor, while time of day contributed over one third of

the total variance in the duration. Previous multilevel studies have not explored these time variances at such a detailed level.

Finally, the dissertation demonstrated uses of Twitter data as a potentially important data source at a lower cost than a survey to understand the insights about comments, criticisms, responses, and shared information about rail transit in Los Angeles. While the Twitter data do not yield individual travel patterns, they do associate geographically specific locations, activities, and opinion. The study found that information flow analysis, sentiment analysis, topic modeling, and its application are indeed a practical and applicable tool for measuring commuters' emotions and opinions. This can be useful for understanding trends among commuters and how their emotions changed according to the rail transit line they used, the time of day, and the day of the week. Additionally, analysis of social media data can provide access to segments of the users that have not participated in traditional ways.

The methods we developed in this dissertation could also applied to analysis from other areas of service operations, which could leverage technology to help government manage information and facilitate interaction in meaningful ways in order to achieve broader public participation than is possible through normal channels. Data analysis of social media data can provide access to segments of the users that have not participated in traditional ways.

REFERENCES

- Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. John Wiley & Sons.
- Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, 52, 27–58. doi:10.1146/annurev.psych.52.1.27
- Ajzen, I. (2002). Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior1. *Journal of Applied Social Psychology*, 32(4), 665–683. doi:10.1111/j.1559-1816.2002.tb00236.x
- Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007). A model for enriching trajectories with semantic geographical information (p. 22). ACM.
- Andrade-Cetto, J., & Sanfeliu, A. (2006). A The Kalman Filter, 119–125.
- Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). *OPTICS: ordering points to identify the clustering structure*. *ACM SIGMOD Record* (Vol. 28, pp. 49–60). ACM. doi:10.1145/304181.304187
- Arentze, T. A., & Timmermans, H. J. P. (2009). A need-based model of multi-day, multi-person activity generation. *Transportation Research Part B: Methodological*, 43(2), 251–265. doi:10.1016/j.trb.2008.05.007
- Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5), 275–286.
- Auger, I., & Lawrence, C. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1), 39–54. doi:10.1016/S0092-8240(89)80047-3
- Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfuser, G., & Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2), 95–124. doi:10.1023/A:1014247822322
- Bamberg, S., Ajzen, I., & Schmidt, P. (2003). Choice of Travel Mode in the Theory of Planned Behavior: The Roles of Past Behavior, Habit, and Reasoned Action. *Basic and Applied Social Psychology*, 25(3), 175–187. doi:10.1207/S15324834BASP2503_01
- Beckx, C., Panis, L. I., Janssens, D., & Wets, G. (2010). Applying activity-travel data for the assessment of vehicle exhaust emissions: Application of a GPS-enhanced data collection tool. *Transportation Research Part D: Transport and Environment*, 15(2), 117–122. doi:10.1016/j.trd.2009.10.004
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285–297. doi:10.1016/j.trc.2008.11.004
- Breiman, L. (2001). Random Forests - Springer. *Machine Learning*.
- Bricka, S. G., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing

- differences in survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies*, 21(1), 67–88. doi:10.1016/j.trc.2011.09.005
- Burbidge, S., & Goulias, K. (2013). Active travel behavior. doi:10.3328/TL.2009.01.02.147-167;pageGroup:string:Publication
- California Department of Transportation. (2002). California statewide household travel survey, 2000-2001 : final report. Retrieved February 1, 2015, from
- Celli, F. (2009, August 28). An algorithm for Trajectories Classification. *Fabriziocelli.Altervista.org*. Retrieved December 4, 2013, from http://fabriziocelli.altervista.org/res/algorithm_classification.pdf
- Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part a: Policy and Practice*, 44(10), 830–840. doi:10.1016/j.tra.2010.08.004
- Cheng, J., & Greiner, R. (1999). Comparing Bayesian network classifiers, 101–108.
- Cho, G.-H., Rodríguez, D. A., & Evenson, K. R. (2011). Identifying Walking Trips Using GPS Data. *Medicine & Science in Sports & Exercise*, 43(2), 365–372. doi:10.1249/MSS.0b013e3181ebec3c
- Collins, C., Hasan, S., & Ukkusuri, S. (2013). A Novel Transit Riders’ Satisfaction Metric: Riders’ Sentiments Measured from Online Social Media Data. *Journal of Public Transportation*, 16(2).
- Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In Cha Zhang & Y. Ma, *Ensemble Machine Learning* (pp. 157–175). Springer.
- Deng, Z., & Ji, M. (2013). Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. In *Traffic and Transportation Studies 2010* (pp. 768–777–777). Reston, VA: American Society of Civil Engineers. doi:doi:10.1061/41123(383)73
- Diaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1), 3. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363357/>
- Doherty, S. T., & Mohammadian, A. (2011). The validity of using activity type to structure tour-based scheduling models. *Transportation*, 38(1), 45–63. doi:10.1007/s11116-010-9285-x
- Doherty, S. T., Noel, N., Gosselin, M.-L., Sirois, C., & Ueno, M. (2001). Moving beyond observed outcomes: integrating global positioning systems and interactive computer-based travel behavior surveys. *Transportation Research Circular*, (E-C026).
- Doherty, S., & Miller, E. (2000). A computerized household activity scheduling survey. *Transportation*, 27(1), 75–97. doi:10.1023/A:1005231926405
- Domarchi, C., Tudela, A., & Gonzalez, A. (2008). Effect of attitudes, habit and affective appraisal on mode choice: an application to university workers. *Transportation*, 35(5), 585–599. doi:10.1007/s11116-008-9168-6
- Duncan, S., Stewart, T. I., Oliver, M., Mavoa, S., MacRae, D., Badland, H. M., & Duncan, M. J. (2013). Portable Global Positioning System Receivers: Static Validity and Environmental Conditions. *Amepre*, 44(2), e19–e29. doi:10.1016/j.amepre.2012.10.013

- Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, 21(7), 1–16.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.
- Feinerer, I. (2010). Analysis and Algorithms for Stemming Inversion. *Airs*, 290–299. doi:10.1007/978-3-642-17187-1_28
- Feinerer, I. (2013). Introduction to the tm Package Text Mining in R. *Tiré De Http://Cran.R-Project.Org/Web/Packages/Tm/Vignettes/Tm.Pdf*, Consulté Le, 18.
- Feng, T., & Timmermans, H. (2013). Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies*.
- Feng, T., Moiseeva, A., & Timmermans, H. (2011). Processing of National Travel Survey GPS Pilot Data, 1–76.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: an introduction to theory and research*. Addison-Wesley Publishing Company.
- Gärling, T., & Axhausen, K. W. (2003). Introduction: Habitual travel choice. *Transportation*, 30(1), 1–11. doi:10.1023/A:1021230223001
- Gärling, T., Fujii, S., & Boe, O. (2001). Empirical tests of a model of determinants of script-based driving choice. *Transportation Research Part F: Traffic Psychology and Behaviour*, 4(2), 89–102. doi:10.1016/S1369-8478(01)00016-X
- Gärling, T., Kalén, T., Romanus, J., Selart, M., & Vilhelmson, B. (1998). Computer simulation of household activity scheduling. *Environment and Planning a Abstract*, 30(4), 665–679. doi:10.1068/a300665
- Golob, T. T., & Meurs, H. (1986). Biases in response over time in a seven-day travel diary. *Transportation*, 13(2), 163–181. doi:10.1007/BF00165546
- Gonzalez, P. A., Weinstein, J. S., barbeau, Sean J, Labrador, M. A., Winters, P. L., Georggi, N. L., & Perez, R. (2010). Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET Intelligent Transport Systems*, 4(1), 37. doi:10.1049/iet-its.2009.0029
- Good, I. J. (1967). The estimation of probabilities: An essay on modern Bayesian methods. *American Educational Research Journal*, 4.
- Goulias, K. (2003). Multilevel Statistical Models. In K. Goulias, *Transportation Systems Planning: Methods and Applications*. CRC Press.
- Gruen, B., & Hornik, K. (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), 1–30.
- Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2), 171–186. doi:10.1023/A:1010920819831
- Hannes, E., Janssens, D., & Wets, G. (2008). Destination Choice in Daily Activity Travel: Mental Map's Repertoire. *Transportation Research Record*, 2054(-1), 20–27. doi:10.3141/2054-03
- Hirsh, M., Prashkea, J. N., & Ben-Akiva, M. (1986). Dynamic Model of Weekly Activity Pattern. *Transportation Science*, 20(1), 24–36. doi:10.2307/25768203?ref=no-x-route:ec4270ecd1eb5f7de3ed248b1de56936
- Houston, D., Luong, T. T., & Boarnet, M. G. (2014). Tracking daily travel; Assessing

- discrepancies between GPS-derived and self-reported travel patterns. *Transportation Research Part C*, 48, 97–108.
- Houston, D., Wu, J., Yang, D., & Jaimes, G. (2013). Particle-bound polycyclic aromatic hydrocarbon concentrations in transportation microenvironments. *Atmospheric Environment*, 71, 148–157. doi:10.1016/j.atmosenv.2013.02.002
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews, 168–177.
- Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., et al. (2005). An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2), 105–108. doi:10.1109/LSP.2001.838216(410)
- Janecek, A., Gansterer, W. N., Demel, M., & Ecker, G. (2008). On the Relationship Between Feature Selection and Classification Accuracy. *Fsdm*, 90–105.
- Jones, P., Koppelman, F., & Orfeuill, J. P. (1990). Activity analysis: State-of-the-art and future directions. In P. Jones, *Developments in Dynamic and Activity-Based Approaches to Travel*. doi:10.1234/12345678
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 35–45.
- Kelly, P., Krenn, P., Titze, S., Stopher, P. R., & Foster, C. (2013). Quantifying the Difference Between Self-Reported and Global Positioning Systems-Measured Journey Durations: A Systematic Review. *Transport Reviews*, 33(4), 443–459. doi:10.1080/01441647.2013.815288
- Killick, R., Eckley, I. A., Jonathan, P., & Chester, U. K. (2011). Efficient Detection Of Multiple Changepoints Within An Oceano-graphic Time Series.
- Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500), 1590–1598. doi:10.1080/01621459.2012.737745
- Kitamura, R., Mokhtarian, P., & Laidet, L. (1997). A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation*, 24(2), 125–158–158. doi:10.1023/A:1017959825565
- Liaw, A., & Wiener, M. (2012, October). Breiman and Cutler's random forests for classification and regression. *Cran.R-Project.org*. CRAN.
- Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7(4), 815–840.
- Los Angeles County Metropolitan Transportation Authority. (2009). 2009 Long range transportation plan. *Media.Metro.Net*. Retrieved February 18, 2015, from http://media.metro.net/projects_studies/images/final-2009-LRTP.pdf
- Lu, Y., & Zhang, L. (2014). Trip Purpose Estimation for Urban Travel in the U.S.: Model Development, NHTS Add-on Data Analysis, and Model Transferability Across Different States. *Transportation Research Board 93rd Annual Meeting*.
- Lu, Y., Zhu, S., & Zhang, L. (2012). A Machine Learning Approach to Trip Purpose Imputation in GPS-Based Travel Surveys. Presented at the Innovations in Travel Modeling 2012.
- Ma, J., & Goulias, K. (1996). Multivariate Marginal Frequency Analysis of Activity and Travel Patterns in First Four Waves of Puget Sound Transportation Panel. *Transportation Research Record*, 1556(1), 67–76. doi:10.3141/1556-09

- Mai, E., & Hranac, R. (2013). Twitter Interactions as a Data Source for Transportation Incidents. Presented at the Transportation Research Record, Proceedings of the 2013 Annual Meeting.
- Matthies, E. (2003). One to bind them all: How the modified moral decision making model can be used for the integration of measures to promote pro-environmental travel mode *Crossing Boundaries—the Value of Interdisciplinary*
- McGowen, P., & McNally, M. (2007). Evaluating the potential to predict activity types from GPS and GIS data. Presented at the Proceedings of the 86th Annual Meeting
- Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis. *Stanford University, CS229(2011 Http://Cs229. Stanford. Edu/Proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. Pdf)*. Retrieved from <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>
- Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. W. (2014). Trip Purpose Identification from GPS Tracks. Presented at the Journal of Public Transportation.
- Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems, 14*, 841.
- Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P. (2013). Supporting large-scale travel surveys with smartphones – A practical approach. *Transportation Research Part C: Emerging Technologies*. doi:10.1016/j.trc.2013.11.005
- Nurul Habib, K., & Miller, E. (2006). Modeling Individuals' Frequency and Time Allocation Behavior for Shopping Activities Considering Household-Level Random Effects. *Transportation Research Record, 1985*(1), 78–87. doi:10.3141/1985-09
- Otto, S. (2010). The psychology of transport choice
- Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories (pp. 863–868). Presented at the the 2008 ACM symposium, New York, New York, USA: ACM Press. doi:10.1145/1363686.1363886
- Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). *Human Computing and Machine Understanding of Human Behavior: A Survey*. link.springer.com (pp. 47–71). Springer Berlin Heidelberg. doi:10.1007/978-3-540-72348-6_3
- Pas, E. I. (1986). Multiday samples, parameter estimation precision, and data collection costs for least squares regression trip-generation models. *Environment and Planning a Abstract, 18*(1), 73–87. doi:10.1068/a180073
- Pierce, B., Casas, J., & Giaimo, G. (2003). Estimating trip rate under-reporting: preliminary results from the Ohio Household Travel Survey. Presented at the Transportation Research Board 82nd Annual Meeting.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems, 14*(3), 130–137. doi:10.1108/eb046814
- Powell, W. B. (2011). *Approximate Dynamic Programming*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781118029176
- 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*, 6(2), 1–27. doi:10.1145/1689239.1689243
- Roider, O., Rudloff, C., & Ray, M. (2012). An Individualized Travel Information System for Optimizing Mode and Route Choice behavior of Commuters. *Transport Research Arena 2012*, 48(0), 1948–1957.
- Schönfelder, S. (2009). Rhythms, routines and variety seeking of activitytravel behaviour. Empirical evidence and policy implications. Presented at the 3rd EuroCities Datta Workshop, Lausanne.
- Schönfelder, S., & Axhausen, K. W. (2010). *Urban Rhythms and Travel Behaviour: Spatial and Temporal Phenomena of Daily Travel*. London: Ashgate Publishing, Ltd.
- Schuessler, N., & Axhausen, K. W. (2009). Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record*, 2105(-1), 28–36. doi:10.3141/2105-04
- Schüssler, N., Montini, L., & Dobler, C. (2011). *Improving Post-processing Routines for GPS Observations Using Prompted-recall Data*.
- Schweitzer, L. (2014). Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter. *Journal of the American Planning Association*, 80(3), 218–238. doi:10.1080/01944363.2014.980439
- Scott, A. J., & Knott, M. (1974). A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3), 507–512.
- Shen, L., & Stopher, P. R. (2013, September 9). A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C*. Retrieved October 16, 2013, from <http://www.sciencedirect.com/science/article/pii/S0968090X13001848>
- Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3), 853–860. doi:10.1016/j.eswa.2013.08.015
- Stopher, P. R. (1992). Use of an activity-based diary to collect household travel data. *Transportation*, 19(2), 159–176.
- Stopher, P. R. (2004). *GPS, Location, and Household Travel*.
- Stopher, P. R., Clifford, E., Zhang, J., & FitzGerald, C. (2008a). Deducing mode and purpose from GPS data, 1–13.
- Stopher, P. R., FitzGerald, C., & Xu, M. (2007). Assessing the accuracy of the Sydney Household Travel Survey with GPS. *Transportation*, 34(6), 723–741. doi:10.1007/s11116-007-9126-8
- Stopher, P. R., FitzGerald, C., & Zhang, J. (2008b). Search for a global positioning system device to measure person travel. *Transportation Research Part C*, 16(3), 350–369.
- Tarigan, A. K. M., Fujii, S., & Kitamura, R. (2012). Intrapersonal variability in leisure activity-travel patterns: the case of one-worker and two-worker households. *Transportation Letters: the International Journal of Transportation Research*, 4(1), 1–13. doi:10.3328/TL.2012.04.01.1-13
- Triandis, H. C. (1977). *Interpersonal behavior*. Brooks/Cole Pub. Co.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Icwsm*, 10, 178–185.

- Vernaza, P., & Lee, D. D. (2006). Rao-Blackwellized particle filtering for 6-DOF estimation of attitude and position via GPS and inertial sensors, 1571–1578.
- Verplanken, B., & Aarts, H. (1999). Habit, Attitude, and Planned Behaviour: Is Habit an Empty Construct or an Interesting Case of Goal-directed Automaticity? *European Review of Social Psychology*, 10(1), 101–134. doi:10.1080/14792779943000035
- Wand, M. P., & Jones, M. C. (1993). Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation. *Journal of the American Statistical Association*, 88(422), 520. doi:10.2307/2290332
- Williams, G. (2011). *Data Mining with Rattle and R The Art of Excavating Data for Knowledge Discovery*. link.springer.com. New York, NY: Springer New York. doi:10.1007/978-1-4419-9890-3
- Wolf, J. (2004). Applications of new technologies in travel surveys. Presented at the International Conference on Transport Survey Quality and Innovation, Costa Rica.
- Wolf, J., Bricka, S. G., & Ashby, T. (2004a). Advances in the Application of GPS to Household Travel Surveys (pp. 1–19). Presented at the Transportation Research Board, National Household Transportation Survey Conference.
- Wolf, J., Hoogendoorn, S. P., & Bachman, W. (2001a). Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record*, 1768(1), 125–134. doi:10.3141/1768-15
- Wolf, J., Loechl, M., Thompson, M., & Arce, C. (2001b). Trip rate analysis in GPS-enhanced personal travel surveys. Presented at the International Conference on Transport Survey Quality and Innovation, Kruger Park, South Africa.
- Wolf, J., SchöUnfelder, S., Samaga, U., Oliveira, M., & Axhausen, K. (2004b). Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. *Transportation Research Record*, 1870(1), 46–54. doi:10.3141/1870-06
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1), 67–82.
- Wu, J., Jiang, C., Houston, D., Baker, D., & Delfino, R. (2011). Automated time activity classification based on global positioning system (GPS) tracking data, 10(1), 101. doi:10.1186/1476-069X-10-101
- Yang, M.-H., Luong, T. T. B., & Recker, W. (2014). Extracting traffic patterns from loop detector data using multiple changepoints detection. Presented at the Transportation Research Board of the National Academies, Washington, D.C.
- Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W.-Y. (2008a). Understanding mobility based on GPS data. Presented at the UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing, ACM Request Permissions. doi:10.1145/1409635.1409677
- Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008b). Learning transportation mode from raw gps data for geographic applications on the web, 247–256.