博士論文（要約）

Methodologies for Automatic Detection of Transportation Mode using Data collected through Multiple Sensors

(複数のセンサー情報を用いた交通機関自動判別手法)

MUHAMMAD AWAIS SHAFIQUE

ムハマド アワイス シャフィク

Methodologies for Automatic Detection of Transportation Mode using Data collected through Multiple Sensors

(複数のセンサー情報を用いた交通機関自動判別手法)

by

MUHAMMAD AWAIS SHAFIQUE

ムハマド アワイス シャフィク

A Dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

東京大学大学院工学系研究科
社会基盤学専攻
Department of Civil Engineering
Graduate School of Engineering,
The University of Tokyo, Japan

September 2015

II

Dedicated to my wonderful parents,

without whom I wouldn't have

achieved so much,

And to my loving wife, who is

always there for me.

THESIS APPROVAL

Student Name	:	Muhammad Awais Shafique
Student I.D.	:	37-127196
Institute	:	The University of Tokyo
Department	:	Civil Engineering
Laboratory	:	Transportation Research and Infrastructure Planning
Title	:	Methodologies for Automatic Detection of Transportation Mode using Data collected through Multiple Sensors
Submission	:	September, 2015

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Research Supervisor:

Prof. Dr. Eiji HATO

_____

Advisory Committee Members:

Prof. Dr. Takashi OGUCHI

_____

Assoc. Prof. Dr. Takashi FUSE

_____

Assoc. Prof. Dr. Tomonori NAGAYAMA

_____

Assoc. Prof. Dr. Takuya MARUYAMA

_____

Chairman, Department of Civil Engineering:

Prof. Dr. Takeshi ISHIHARA

_____

Department of Civil Engineering
Graduate School of Engineering,
The University of Tokyo, Japan

Methodologies for Automatic Detection of Transportation Mode using Data collected through Multiple Sensors

(複数のセンサー情報を用いた交通機関自動判別手法)

Muhammad Awais Shafique

Ph.D. Thesis-Civil Engineering
September 2015

Supervisor: Prof. Eiji Hato

ABSTRACT

The construction, expansion, rehabilitation or discontinuation of any transportation infrastructure requires the collection of ground data, which will provide the basis for the planned operation. Similarly, policies implemented to meet certain targets, e.g. pollution control, traffic reduction, sustainability etc. also require extensive research, which is again based on data. The vitality of the data makes the collection all the more important. Conventionally, household trip data was collected by questionnaire surveys, which is still being practiced in many parts of the world even today. The collection method was later improved to conduct the surveys online. These methods have an inherent problem of dependence on the memory of the respondent. Due to this dependence, accuracy of data cannot be assured. Shortcomings include inaccuracies in recording the starting and ending times, underreporting due to missing short trips and non-response. To address the source issue, passive data collection methodologies have recently evolved.

This dissertation deals with the state of the art technique to gather and interpret travel related data. Data recorded by various sensors can be used to identify the mode of travel used by the person carrying the device. For this purpose, the most recent method is to use supervised learning algorithms like support vector machine (SVM), neural network (NN), Naïve Bayes (NB), decision trees (DT), adaptive boosting (AdaBoost), random forest (RF) etc. The current research also benefits from the use of supervised learning algorithms. The first part of the dissertation deals with the data collected by sensors integrated in a purpose-built wearable device named Behavioral Context Addressable Loggers in the Shell (BCALs). BCALs was employed to collect travel data from three cities of Japan namely Niigata, Gifu and Matsuyama. The accelerometer and GPS data was directly processed by the device and some basic features were yielded for the analysis. Overall accuracy of around 80 % was achieved to distinguish among four modes of transportation, i.e. walk, bicycle, car and train. The minimum accuracy per mode was reported to be 56 %. Upon further processing the data using moving window concept to extract a number of useful features from accelerometer data alone and using 70 % randomly selected data to train different algorithms, the classification results improved. Among the four classification algorithms tested, random forest performed best and for a final selection of 125 point moving window size, an overall accuracy of around 99 % was achieved, with minimum accuracy of 97 %.

To adopt the methodology for data collected from smartphones, only GPS data and acceleration data along 3 axes, collected by BCALs, was used as raw data. Resultant acceleration and average resultant acceleration was calculated from the individual

accelerations, using moving window concept for the latter feature. From GPS data, distance and time between consecutive readings were calculated. GPS coordinates were further used to extract driving and walking distance and time using Google Maps distance matrix API. The only personal attributes collected i.e. gender and age, were also included as features. Again 30 % data was tested using random forest and the overall classification accuracy turned out to be more than 99 %. Best feature selection revealed that only four variables were enough to achieve similar accuracy.

Sensors' data using smartphones was collected by participants in Kobe city. The data was used to compare the performance of various classification algorithms for travel mode detection. Boosted decision trees was most accurate closely followed by random forest but because random forest was much quicker and the difference in accuracy was not much, it was preferred and used in the studies to follow.

Again moving window concept was used to extract various features from accelerometer data, including standard deviation, skewness and kurtosis. Gyroscope readings were directly used as features. Moving window size of 10 minutes and 10 % learning data amount was selected. The collected data was scaled down to reduced data recording frequencies to acquire an energy-efficient solution. For six modes of transportation, the overall accuracy ranged from 99.96 % for 10 Hz data to 94.48 % for 0.2 Hz data but at the same time, the computational times ranged from 304.86 seconds to 3.53 seconds. The results suggested that a compromise needs to be agreed upon between the accuracy and computational time. Later, speed calculated from GPS data was used as raw data along with resultant acceleration calculated from

accelerometer data, to extract features. Because of imbalanced data, a modification was introduced to devise weighted random forest. The results were improved further by employing a 2-step post-processing method. Weighted random forest improved the accuracy to a maximum 13 % and post-processing further improved by a maximum 13 %. 10-fold cross-validation was used to validate the results. Down-sampling provided additional refinement, resulting in 98.42% overall accuracy for 0.067 Hz data. The last chapter studies the possibility of utilizing the mode choice model along with the machine learning algorithm. The combination yielded an astonishing 99% overall accuracy with no mode accuracy less than 90%.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful

Alhamdulillah. All praise and glory to **Almighty Allah (Subhanahu Wa Taalaa)** who gave me courage and patience to carry out this work. Peace and blessing of Allah be upon His last Prophet **Muhammad (Peace Be upon Him)**.

I would like to express my deepest gratitude and very special appreciation to my supervisor, **Prof. Dr. Eiji Hato**, for his enthusiastic supervision of my thesis. His helpful guidance, constructive criticism, valuable suggestions, consistent supervision, and encouragement helped me getting through the journey of my PhD study. I also would like to express my gratitude to **Prof. Dr. Hitoshi Ieda**, **Asst. Prof. Dr. Kiichiro Hatoyama**, **Prof. Dr. Makoto Shimamura** and **Asst. Prof. Dr. Hideki Yaginuma** for their useful advices and providing all facilities for my research in Transportation Research and Infrastructure Planning Laboratory.

I would like to thank the members of the reading committee **Prof. Dr. Takashi Oguchi**, **Assoc. Prof. Dr. Takashi Fuse**, **Assoc. Prof. Dr. Tomonori Nagayama** and **Assoc. Prof. Dr. Takuya Maruyama** for their important remarks about my thesis. I would also like to thank all my lab fellows from TRIP lab and BinN lab, who assisted me in my research and turned my stay into a remarkable experience.

I am grateful to my parents for their love, support, perfect understanding, and encouragement that inspired me and strengthened me throughout my research. I could never have completed this thesis and would never have got this far without them. To

my wife, who stood by my side in every tribulation and shared every single moment of joy and sorrow.

Finally, I would like to acknowledge the **Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT)** for providing me with the Monbukagakusho scholarship to pursue my PhD in Japan, and The University of Tokyo from giving me the admission.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 INTRODUCTION

1.1. Background

Household travel data comprises of the socio-economic and demographic characteristics of surveyed individuals, their household characteristics and the travel information of the individuals for a prescribed time varying from one day to several days and even weeks or months. The travel information contains the start and end locations, start and end times, mode of transportation, accompanied persons and the purpose of the trip. All this data is collected to understand the travel behavior and patterns of the people in the sample. The collected data is indispensable for the transportation officials and is utilized for efficient design and management of transportation infrastructure. Furthermore, this information provides the ground truth to become the base for introduction of new policies as well as revision or suspension of implemented policies. These policies, in turn are the vital part of transportation demand management.

Conventionally, the travel data has been collected by conducting travel surveys. These surveys include face-to-face interviews, paper questionnaires, telephone surveys, and computer-assisted surveys. All these methods have drawbacks, some of which are as follows.

1. They all depend on the memory of the respondents.
2. The recording of data is tedious for the respondents.
3. The methods are expensive and consequently have low update frequency

4. Seasonal variations are difficult to capture and are therefore usually ignored.
5. The response rate is low because of the tiresome nature of recording method applied.
6. Small trips are usually ignored or forgotten.
7. Perception of time varies from person to person and is also affected by the mode of travel.
8. Responses are usually biased.
9. Exact starting and ending times are not reported, rather approximate values are stated.

The above-stated shortcomings led to the introduction of passive data collection methods. These are the methods were information is collected automatically either by installing devices at fixed locations or by having the participants carry the devices along with them on their journeys. It started out with experiments on Global Positioning Systems (GPS) installed in the participants' vehicles to monitor their daily movement and route choice. Obviously this methodology lacked the ability to collect data pertaining to modes other than the personal vehicles. With the introduction of light GPS devices that can be carried around, the location data collection was expanded to include other modes as well. With technological advancements, multiple sensors were utilized to record data and infer the travel mode and trip purpose. These sensors included accelerometer, barometer, magnetometer, gyroscope etc. Purpose-built wearable devices housing various sensors were used by different researchers with the aim to collect such a data which will be most practical for identifying the mode of transportation.

Smartphones today, have a range of sensors integrated in them. This inclusion of various sensors has presented smartphones as a viable data collection device. The additional advantages lie in the increasing penetration rates of smartphones in different countries of the world. A number of researchers are studying the application of smartphones for travel data collection. The problems at hand are as follows,

1. The reported classification accuracies of past studies are not very high.
2. Few studies use simple features for classification.
3. The quick drainage of smartphone battery during data collection is a big issue, and limits the application of the methodology.

3.1. Research Objectives

As discussed earlier, conventional methods of travel data collection have a number of weaknesses. To address these issues, automatic travel data collection methods have been devised. However, the use of devices, including smartphones, is a developing topic. In view of this fact my research objectives are as follows,

1. Develop a methodology to use the sensors' data (GPS, accelerometer and gyroscope) collected by smartphones or any other device, to identify the travel modes.
2. The developed methodology should be battery-efficient and should be highly accurate, while using simple features to reduce the complexity of the approach.

3.2. Thesis Organization

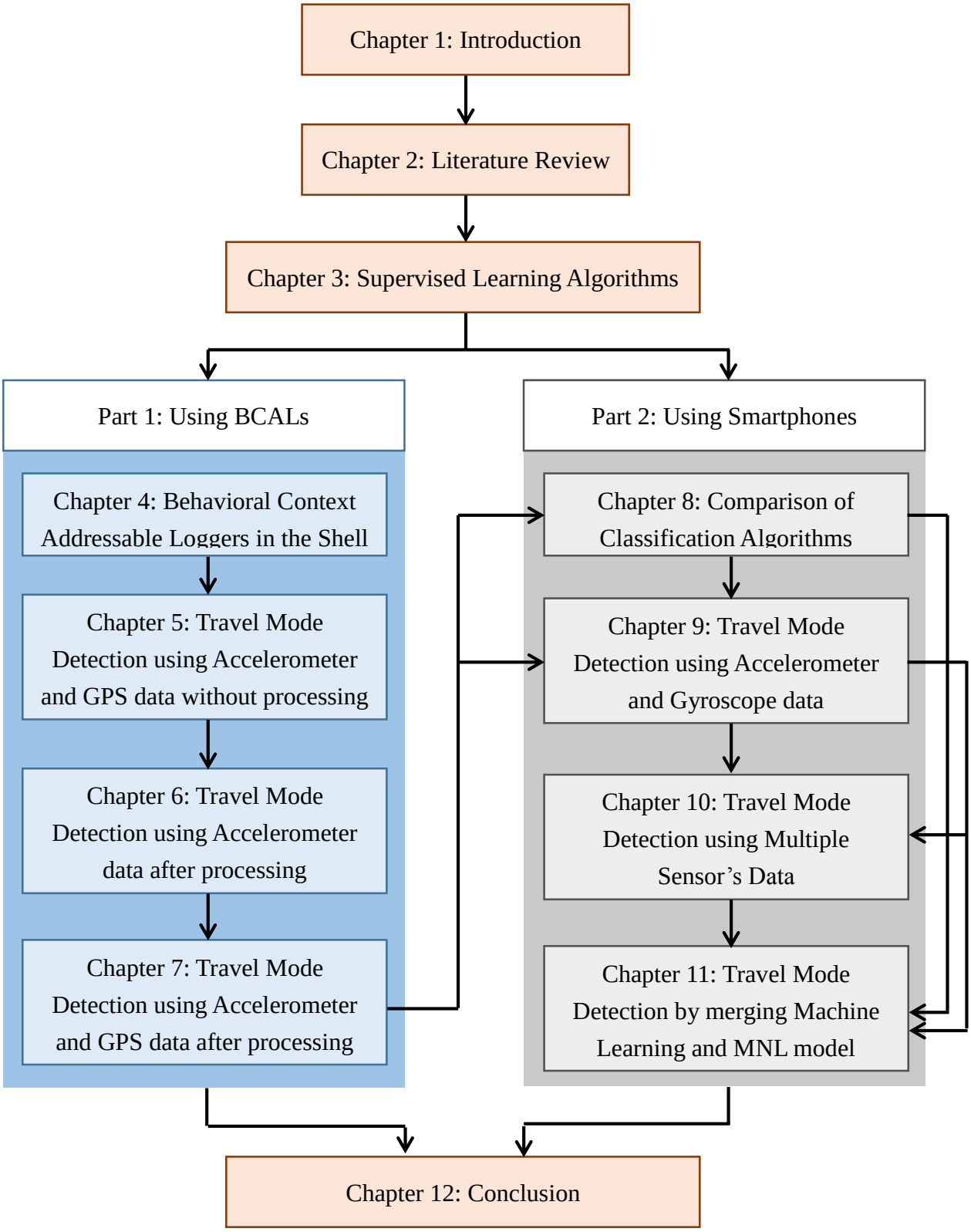


Figure 1-1: Thesis Organization

Chapter 2 LITERATURE REVIEW*

The contents of this chapter cannot be published, because this chapter is scheduled to be published in an academic journal. It is planned to be posted within next 5 years.

* This chapter is to be published as Shafique M.A. and Hato E. A Review of Automatic Travel Mode Detection Methods. *Transport Reviews.*

Chapter 3 SUPERVISED LEARNING ALGORITHMS*

The contents of this chapter cannot be published, because this chapter is scheduled to be published in an academic journal. It is planned to be posted within next 5 years.

* This chapter is to be published as part of Shafique M.A. and Hato E. Improving the Accuracy of Travel Mode Detection for low Data Collection Frequencies. *Transportation Research Part C.*

This chapter is presented in part as Shafique M.A. and Hato E. A Comparison among various Classification Algorithms for Travel Mode Detection using Sensors' data collected by Smartphones. *14th International Conference on Computers in Urban Planning and Urban Management, CUPUM 2015*, MIT, Cambridge, Massachusetts. July 2015.

Chapter 4 BEHAVIORAL CONTEXT ADDRESSABLE LOGGERS IN THE SHELL

2.1. Introduction

Behavioral Context Addressable Loggers in the Shell (BCALs) is a small, portable device developed by Hato (2010) for the purpose of travel-activity observation. This chapter introduces BCALs as it is the source of data collection used in section 1 of the thesis. The aim for the development of such a device was to introduce an instrument which is capable of estimating behavioral contexts without the requirement of any entry from the respondents. This way, the human error related to reliance on inaccurate memory of the events can be eliminated.

2.2. BCALs

Designed and developed by Hato (2010), the instrument (Figure 4-1) was equipped with multiple sensors and therefore was capable of collecting different kinds of data, listed in Table 4-1. Originally, the acceleration data was used to identify the travel mode, whereas the atmospheric pressure along with ultraviolet rays was utilized to judge whether the participant was indoors or outdoors, and if indoors then which floor level. Moreover, sound and temperature observations were intended to evaluate the surrounding environment.

Figure 4-1: BCALs equipped with various sensors

Table 4-1: Data acquired from BCALs (Hato, 2010)

Sr. No.	Data
1	X-axis acceleration (32 Hz)
2	Y-axis acceleration (32 Hz)
3	Z-axis acceleration (32 Hz)
4	Atmospheric pressure (32 Hz)
5	Angular velocity (32 Hz)
6	Ultraviolet ray (32 Hz)
7	Direction (32 Hz)
8	Sound (10 Hz)
9	PS location (latitude, longitude, acceleration, velocity, and direction) (1 Hz)
10	Elliptical error of GPS location measurement

2.3. Summary

Although, BCALs can record numerous types of data, but its use in this thesis is only restricted to GPS and accelerometer data.

Chapter 5 TRAVEL MODE DETECTION USING GPS AND ACCELEROMETER DATA WITHOUT PROCESSING*

5.1. Introduction

In this chapter as well as in the following two chapters, sensors' data collected by BCALs is used to identify the mode of transportation used by the participants. Data acquired from travel surveys provide basic information for the traffic modeling, service optimization and routing. This is immensely valuable for traffic planners, transportation authorities, public transport providers and researchers. Conventional data collection methods for travel surveys comprise telephone interviews, personal interviews, travel diaries, mail-back or web-based questionnaires, traffic counting on cross sections or intersections as well as analyses of transport schedule inquires. When collecting data on a large scale, most of these methods are expensive and time consuming. Consequently the update frequency of the data is very low. Moreover, nonresponse issues and underreported trips are well-known problems in surveys.

The second half of 1990s witnessed the introduction of Global Positioning System (GPS) devices to supplement the measurement of personal travel. One of the first household surveys employing GPS was conducted in 1997 in Austin, Texas, followed

* This chapter is published as Shafique M.A., Hato E. and Yaginuma H. Using Probe Person Data for Travel Mode Detection. *International Journal of Computer, Information, Systems and Control Engineering, 8*(10), 1482 – 1486. 2014.
This chapter is presented as Shafique M.A., Hato E. and Yaginuma H. Using Probe Person Data for Travel Mode Detection. *International Conference on Signal Processing, Pattern Recognition and Applications*, Osaka, Japan. Oct. 2014.

by many studies conducted to examine the application of GPS to determine travel behavior (Bohte and Maat, 2009; Murakami and Wagner, 1999; Stopher et al., 2008; Wolf et al., 2001). These studies aim at determining the possibility of GPS device, combined with Global Information Systems (GIS), to replace or supplement conventional methods. Devices are either wearable or mounted in private vehicles. Common problems with GPS are loss of signal in underground facilities, high energy consumption and the willingness of users carrying the device.

5.2. Methodology

5.2.1. Data Collection

This study comprises of the data collected from three Japanese cities namely Niigata, Gifu and Matsuyama. Figure 5-1 shows the locations of the three cities with reference to the location of Tokyo.

Niigata is a coastal city facing the Sea of Japan and the Sado Island. It is the capital and the most populous city of Niigata prefecture. The city includes of a number of wetlands. In the west of Tokyo lies the Gifu prefecture, having a capital with the same name. The area varies from buildup city center to orchards in the surrounding regions. Further south west of Gifu is the Ehime prefecture on the island of Shikoku. Its capital is Matsuyama. In addition to railway lines, Matsuyama also houses tram lines. Population of each city (as reported in 2010 census) is given in Table 5-1.

The data collected can be classified into two categories, Location Data and Trip Data.

Location data, collected by BCALs, comprised of features extracted from accelerometer and GPS whereas trip data, recorded by travel diaries, contained travel activity information. Table 5-2 shows the contents of both types of data. The location data was labeled with the help of information collected in the trip data.

The surveys were conducted during Jan to Feb 2011 in Niigata, Dec 2010 to Jan 2011 in Gifu and Nov 2010 to Jan 2011 in Matsuyama. Table 5-3 gives the amount of data used in the study discussed in this chapter. Moreover the distribution of data with respect to the mode of transportation is also shown.

Figure 5-1: Location of Tokyo, Niigata, Gifu and Matsuyama

Table 5-1: Population of cities surveyed

City	Population
Niigata	811,901
Gifu	413,136
Matsuyama	517,231

Table 5-2: Contents of Location and Trip Data

Data	Contents
Location Data	User ID, Recording date and time, Exercise intensity, No. of steps, Min, Max and Avg. accelerations in movement, crosswise and vertical directions, Resultant and Avg. resultant accelerations
Trip Data	User ID, Departure date and time, Arrival date and time, Means of transportation

Table 5-3: Distribution by Mode

Mode	Niigata	Gifu	Matsuyama
Walk	164,078 (71.49%)	83,645 (58.24%)	168,687 (62.62%)
Bicycle	3,214 (1.4%)	24,559 (17.09%)	15,404 (5.72%)
Car	61,785 (26.92%)	34,678 (24.14%)	81,653 (30.31%)
Train	425 (0.19%)	744 (0.52%)	3,631 (1.35%)
Total	229,502 (100%)	143,626 (100%)	269,375 (100%)

5.2.2. Mode Assignment

Before assigning the mode of transportation used, the location data for each city was scanned for any redundant entries, with reference to the trip data. Afterwards using the departure and arrival times listed in the trip data, the relevant data sets in the location data were assigned the corresponding mode of transportation. The data sets which were left unassigned were deleted.

5.2.3. Elementary Analysis

Using GPS, the location of the respondent is recorded after every 1 minute. Hence the distance covered per minute can be assessed. This feature can help distinguish among

different modes of transportation. Figure 4-3 shows a part of the trip done by a respondent in Niigata city. The person walked from point A to F after which he used car from point F to K. The points are taken approximately 10 minutes apart to clarify the picture. It can be observed that the two modes can easily be separated based on the distance covered from point to point.

The accelerometer sensor integrated in the smart phone calculates and records the acceleration along X, Y and Z directions. These accelerations can also assist in determining the mode of transportation used. Figure 5-2 shows the variation of average resultant acceleration along the same route (A-K). It is evident from figure 5-3 that some pattern exists among the modes when it comes to acceleration.

Figure 5-2: Route formed by GPS data

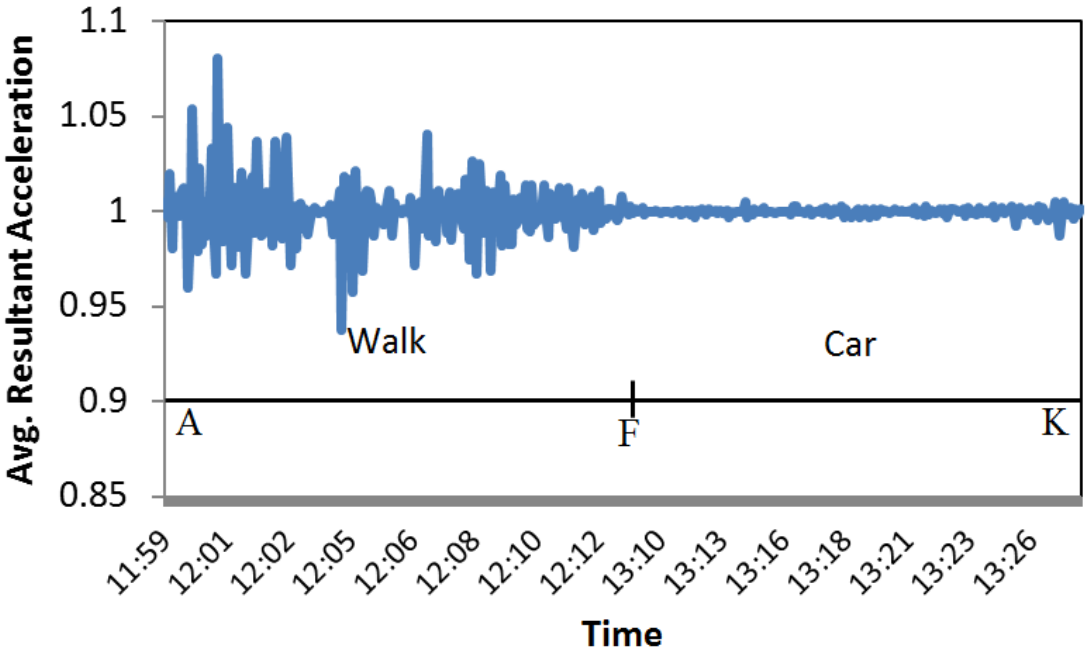


Figure 5-3: Acceleration along the trip

5.2.4. Classification

Support Vector Machine

The first choice for this type of classification problem was Support Vector Machine (SVM). SVM was trained by using the learning data for each city. After the training phase, the algorithm was used to predict the mode of transportation for the test data.

Adaptive Boosting

Next, Adaptive Boosting or AdaBoost was employed. It was used in two ways as shown in Figure 5-4. Firstly AdaBoost was trained for one mode (as it is a binary classifier) and then used to predict the test data. For example the mode was “Train” then the prediction resulted in “Train” and “Other”. Further the algorithm was trained for the second mode and prediction was done on the data sets predicted as “Other” from the first step. This procedure continued for all four modes. Consequently a small

amount of test data predicted as “Other” remained in the end. SVM was used to predict this data.

In the second method, AdaBoost algorithm was trained for each mode separately but the prediction was done for the entire test data. The prediction results were analyzed and the data sets having only one mode predicted were assigned that mode. Rest all were assigned as “Other”. The whole procedure was repeated for the remaining test data assigned as “Other”. In the end SVM was used to finish off the task.

Prediction Improvement

The prediction results, acquired by the application of SVM alone and by AdaBoost along with SVM, were further improved by adopting a simple method. By applying conditions on user ID, measurement date and measurement time, the data was divided into a number of trips. The mode is expected to remain same throughout one trip, so in light of this hypothesis the statistical mode of the results was determined for each trip and was then assigned to the respective trip.

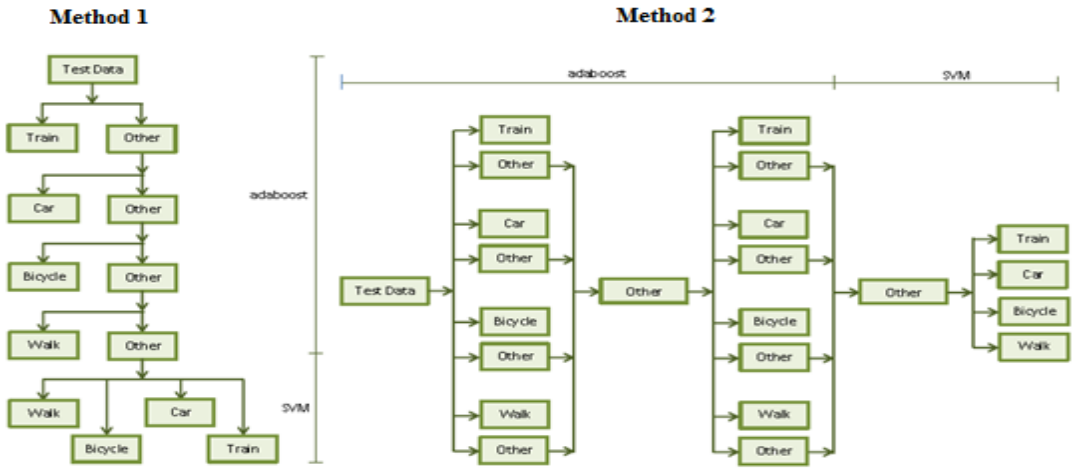


Figure 5-4: Methods for AdaBoost application

5.3. Results and Discussion

The prediction results acquired for the three cities, by applying the methodology discussed, are summarized in Table 4-5. The results suggest that the method of prediction improvement has a positive influence on the outcomes. Moreover the application of AdaBoost proved to be better than SVM alone. If the two methods of AdaBoost application are investigated then the second method involving repetition yielded comparatively better prediction results.

Table 5-4: Prediction Results

City	Mode	Normal			Improved		
		SVM	AdaBoost (1)	AdaBoost (2)	SVM	AdaBoost (1)	AdaBoost (2)
Niigata	Walk	81.77 %	81.76 %	82.38 %	92 %	91.39 %	92.49 %
	Bicycle	71.53 %	86.43 %	87.06 %	93.25 %	100 %	100 %
	Car	64.67 %	74.83 %	75.32 %	68.96 %	81.67 %	81.3 %
	Train	93.88 %	97.18 %	97.88 %	100 %	100 %	100 %
	All	77.05 %	79.99 %	80.57 %	85.83 %	88.91 %	89.6 %
Gifu	Walk	64.46 %	67.44 %	70.27 %	75.72 %	76.16 %	79.58 %
	Bicycle	72.09 %	73.89 %	74.13 %	86.05 %	90.76 %	89.83 %
	Car	68.65 %	75.84 %	76.28 %	75.64 %	83.76 %	86.23 %
	Train	87.5 %	96.91 %	97.04 %	84.95 %	84.95 %	84.95 %
	All	66.9 %	70.72 %	72.52 %	77.51 %	80.54 %	82.97 %
Matsuyama	Walk	67.75 %	68.65 %	68.43 %	79.54 %	79.99 %	79.76 %
	Bicycle	59.65 %	62.64 %	63.46 %	72.32 %	81.41 %	81.06 %
	Car	44.63 %	57.93 %	59.78 %	44.92 %	67.28 %	70.62 %
	Train	60.53 %	67.2 %	67.78 %	58.14 %	58.63 %	55.96 %
	All	60.18 %	65.04 %	65.51 %	68.34 %	75.93 %	76.74 %

When cities are compared, Niigata gives the best results followed by Gifu and then Matsuyama. This might be because of difference in infrastructure, climate change and variation in technologies introduced. Although Niigata has the highest population among the three cities compared and the precipitation level is also very high, yet the prediction results are much better, partly due to moderate climate present during the survey.

Chapter 6 TRAVEL MODE DETECTION USING ACCELEROMETER DATA AFTER PROCESSING*

6.1. Introduction

In this chapter, the accelerometer data collected by BCALs is used for mode classification. The data correspondes to the same survey used in the previous chapter. An accelerometer measures the acceleration of a device in three directions with respect to gravitational force. This means that when the device is placed on a flat surface, an acceleration of 1 g is detected in a downward direction, whereas zero acceleration is recorded in the other two directions. To improve the methodology, pre-processing is employed in order to extract features of high importance.

Mode determination will not just prove beneficial for the transportation sector, but will also pave the way for a new and effective means of advertising. For example, if a user's location and mode of transportation are known in real time, a message can be sent to his or her mobile phone advertising the nearest facilities available in connection with the mode detected. In addition, products relating to a particular mode used can be advertised directly to the user. In this way, the data accumulated can be used to implement a targeted customer-oriented advertising program (Figure 6-1).

* This chapter is published as Shafique M.A. and Hato E. Use of acceleration data for transportation mode prediction, *Transportation*, *42*(1), 163-188. 2015.

Figure 6-1: Concept of Customer-oriented advertisement

For the purpose of mode detection, the analyst can currently avail of different types of classification algorithms. Some of these are listed in Table 6-1, along with their advantages and disadvantages and the methodologies associated with each. A number of researchers have compared the various algorithms in comparative studies, some of which are summarized in Table 6-1. The four classifiers used in this study, namely, AdaBoost, SVM, decision tree and random forests, were selected based on the results derived from existing comparative studies. These algorithms have exhibited good performance in numerous studies, and their respective advantages and disadvantages can be seen in Table 3-1. The current study compares the algorithms with a view to ascertaining the one most appropriate for transportation mode detection.

Table 6-1: Some previous algorithm comparison studies

Study	Algorithms compared	Best Algorithm
Caruana and Niculescu-Mizil (2006)	Support Vector Machines, Neural Nets, Logistic Regression, Naïve Bayes, Memory-based Learning, Random Forest, Decision Trees, Bagged Trees, Boosted Trees, Boosted Stumps	Boosted Trees
Nick et al. (2010)	Naïve Bayes, Support Vector Machines	Support Vector Machines
Reddy et al. (2010)	Decision Trees, K-Means Clustering, Naïve Bayes, Nearest Neighbor, Support Vector Machines, Continuous Hidden Markov Model, Decision Trees with discrete Hidden Markov	Decision Tree with discrete Hidden Markov Model
Stenneth et al. (2011)	Naïve Bayes, Bayesian Network, Decision Trees, Random Forest, Multilayer Perceptron	Random Forest
Yu et al. (2013)	Decision Trees, AdaBoost, Support Vector Machines	Support Vector Machines

GPS data is not used in this study at all. Although GPS data has been shown to work well for mode detection, certain disadvantages are associated with it. The main difficulty is the drop in accuracy due to signal loss or degradation during warm or cold starts, and in 'urban canyons' (Gong et al., 2012; Schuessler and Axhausen, 2009; Stopher et al., 2008). Warm and cold starts happen when a GPS logger requires between five and thirty seconds more to find enough satellites for accurate location detection after being off (or underground) for a long period of time. In densely built central business districts (CBDs), satellite signals do not generally reach the GPS device directly but are bounced off tall buildings. This is known as the urban canyon

effect. The above drawbacks associated with GPS use tend to decrease the accuracy of the results extracted from GPS data. Furthermore, respondents' privacy concerns are also a problem in this area. If a smartphone is used as the data collection instrument, developing a methodology using acceleration data alone will not only address the above problems but will also extend the battery time of the smartphone during data collection, as the GPS sensor will not be activated.

6.2. Methodology

6.2.1. Data Collection

The data was collected from three cities in Japan, namely, Niigata, Gifu and Matsuyama. In Niigata, the surveys were conducted during January and February 2011 and involved 12 participants; in Gifu, they were conducted in December 2010 and January 2011 and involved 8 participants; and in Matsuyama, they were conducted in November 2010 and January 2011 and involved 26 participants. The data collected can be classified into location data and trip data.

The location data was recorded using Behavioral Context Addressable Loggers in the Shell (BCALs) (Hato, 2010). BCALs, are purpose-built wearable devices equipped with different sensors, in addition to a GPS and an accelerometer. They can record location as well as acceleration in three directions, a task that is now possible using modern smartphones. The BCALs observed the various sensors' readings at a frequency of 16 Hz or 16 readings per second, but the readings transmitted to the server were spaced out at an average of five seconds. Hence, the maximum, minimum and

average readings were calculated by the device for each five-second interval and then recorded by the server. The wearable devices were kept in the same position throughout the trip so that accelerations in different directions could be judged easily. The trip data was collected using paper-based travel diaries in which the respondents were asked to record the details of their everyday trips. Feedback calls were made to the respondents to correct any mistakes made during reporting. Again, this is a task that can be fulfilled using smartphones, a method used by many researchers. A simple application developed for the smartphone can be utilized to record the start and end of a trip, as well as the mode of transportation used.

The location data comprised GPS data and accelerometer data. The accelerometer data recorded was the minimum, maximum and average acceleration in movement, crosswise and vertical directions. Moreover, resultant acceleration and average resultant acceleration were also noted. The trip data covered the information regarding each trip, i.e., the date, start time, end time and mode used.

Table 6-2 presents the raw location data and the mode-assigned data (discussed in Section 6.2.2. Mode assignment) for each city. The table also shows the assignment of the data to various modes. Table 6-3 displays the trip share for each mode. Due to data limitations, the analysis was carried out for four modes only. Acceleration data relating to the bus as a fifth mode was either non-existent or so small that it was not treated separately but simply merged with the car travel data. Similarly, only one trip was recorded for Shinkansen (the high-speed train), and instead of adding a new mode, it was included with the train data.

Table 6-2: Amount of data collected through BCALs

Data¥City	Niigata	Gifu	Matsuyama
Raw Location Data	341,712	258,388	507,014
Mode Assigned Data	229,502 (100.00%)	143,626 (100.00%)	269,375 (100.00%)
Walk	164,078 (71.49%)	83,645 (58.24%)	168,687 (62.62%)
Bicycle	3,214 (1.4%)	24,559 (17.09%)	15,404 (5.72%)
Car	61,785 (26.92%)	34,678 (24.14%)	81,653 (30.31%)
Train	425 (0.19%)	744 (0.52%)	3,631 (1.35%)

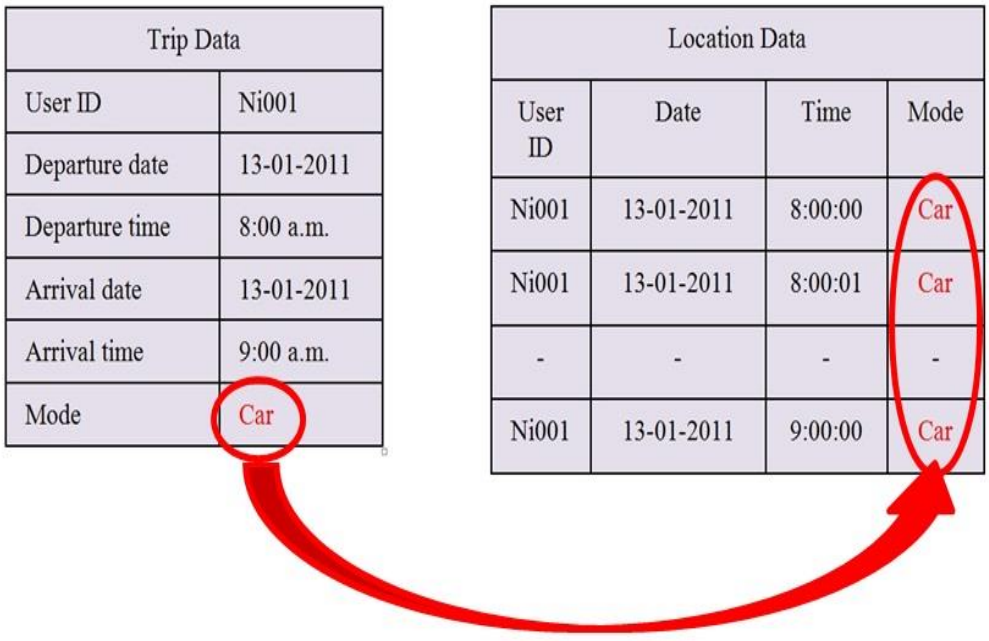
Table 6-3: Number of trips recorded

Mode¥City	Niigata	Gifu	Matsuyama
Walk	662	342	861
Bicycle	12	180	90
Car	306	219	386
Train	4	3	40
Total	984	744	1377

6.2.2. Mode Assignment

The location data was filtered in terms of the trip data. For example, if accelerometer data was recorded with respect to a user for a specific day, but the user had not registered any trips for that particular day in the trip data, then the accelerometer data recorded was of no use. Moreover, data sets with zero acceleration (‘rest’ position)

were also discarded. Using the departure and arrival times listed in the trip data, the corresponding data sets in the location data were assigned the respective mode of transportation, as shown in Figure 6-2. After the mode of transportation was assigned to the location data, the remaining data sets were disposed of. The reason some data remained unassigned is that the accelerometer data may have contained data sets recorded before the start of the trip or after the end of the trip. The remaining data was used in subsequent pre-processing and analysis.



Trip Data	
User ID	Ni001
Departure date	13-01-2011
Departure time	8:00 a.m.
Arrival date	13-01-2011
Arrival time	9:00 a.m.
Mode	Car

Location Data			
User ID	Date	Time	Mode
Ni001	13-01-2011	8:00:00	Car
Ni001	13-01-2011	8:00:01	Car
-	-	-	-
Ni001	13-01-2011	9:00:00	Car

Figure 6-2: Example of Mode Assignment

6.2.3. Elementary Analysis

A distinction between the modes was detected upon careful examination of the acceleration data. For instance, Figures 6-3 to 6-6 show part of the acceleration data for each mode. It can be observed that walking has maximum variability, followed by

cycling. This could be due to excessive movement by the traveler carrying the device. On the other hand, the car and train modes showed relatively small acceleration variability, probably due to the smooth travelling environment. Therefore a clear distinction can be perceived between the different modes by just inspecting the acceleration data.

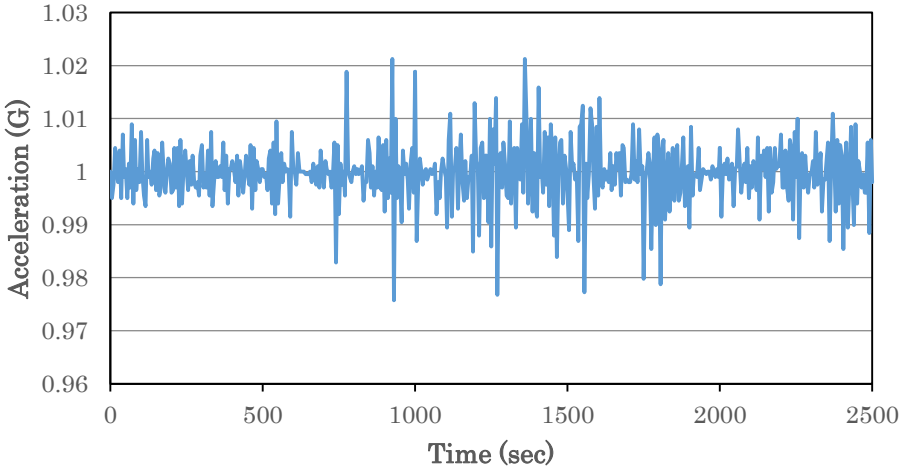


Figure 6-3: Average Resultant Acceleration for walking

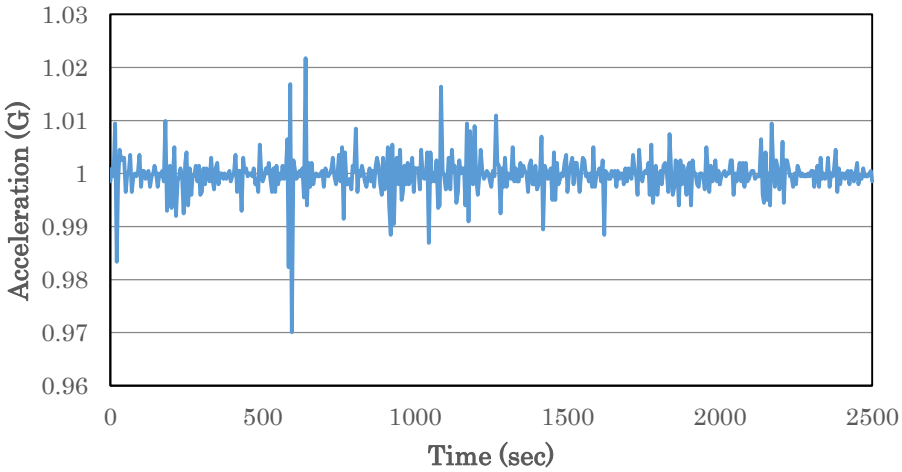


Figure 6-4: Average Resultant Acceleration for bicycling

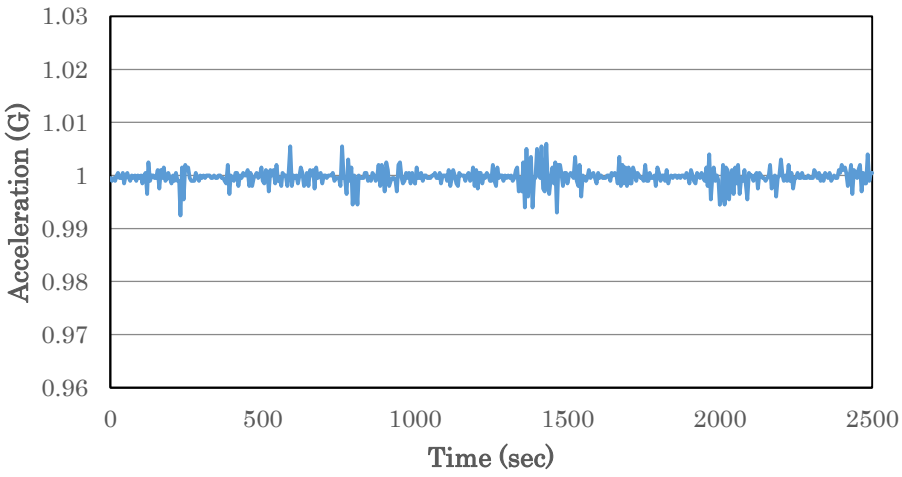


Figure 6-5: Average Resultant Acceleration for automobile travel

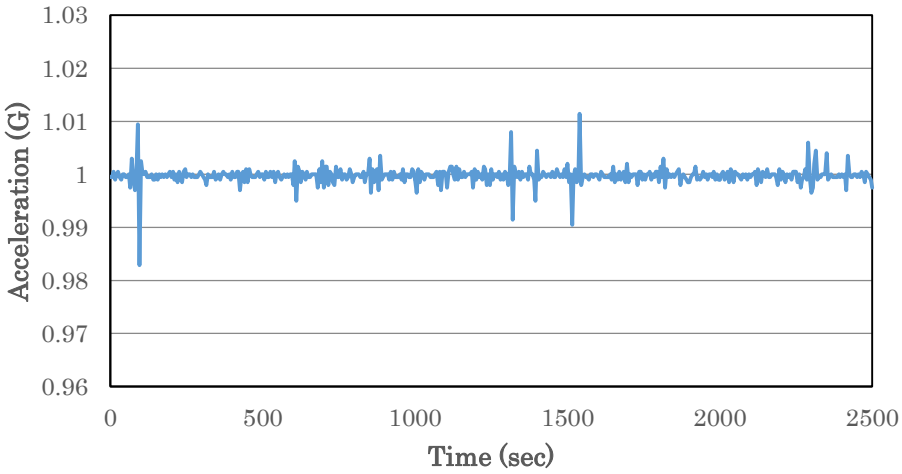


Figure 6-6: Average Resultant Acceleration for train travel

6.2.4. Pre-Processing

Pre-processing was applied in two stages. First, the moving average was calculated, followed by the differences between each mode. The moving average was calculated at 25 point, 50 point, 75 point, 100 point and 125 point in order to identify the trend most likely to maximize classification accuracy.

In this case, *x* denotes the various data entries for acceleration in any direction, *n* is the total number of data entries and *k* is the window size (25, 50, 75, 100 and 125) for calculating the moving average. At any position *i* within the data, the window will cover x_j entries to calculate the moving average. The window will keep the reference entry x_i at the center, except at the start and end of the data set. As the reference entry x_i moves closer to the start or end of the data set, the window will be suppressed. As a solution to this, the window was halved at the start and end of the data set, with the reference entry kept at one end of the window rather than placed in the center. The following equations 6.1 and 6.2 were formulated for the calculation of the *k* point average. Equation 6.2 was used only for average resultant acceleration.

$$(k\ point\ Avg)_i = \begin{cases} \frac{2}{k}\sum_{j=i}^{i+k/2} x_j & if\ i \le k/2 \\ \frac{1}{k}\sum_{j=i-k/2}^{i+k/2} x_j & if\ k/2 < i < n - k/2 \\ \frac{2}{k}\sum_{j=i-k/2}^{i} x_j & if\ i \ge n - k/2 \end{cases} \tag{6.1}$$

$$(k\ point\ Avg)_i = \begin{cases} \frac{2}{k}\sum_{j=i}^{i+k/2} |x_i - x_{i-1}|_j & if\ i \le k/2 \\ \frac{1}{k}\sum_{j=i-k/2}^{i+k/2} |x_i - x_{i-1}|_j & if\ k/2 < i < n - k/2 \\ \frac{2}{k}\sum_{j=i-k/2}^{i} |x_i - x_{i-1}|_j & if\ i \ge n - k/2 \end{cases} \tag{6.2}$$

Equation 5.1 shows that at the start of the data set, that is, when the reference position *i* had not yet exceeded the *k/2* mark, a window of size *k/2* was used to calculate the

average value, with the reference value at the start of the window. Similarly, at the end, the *k/2*-sized window was used, keeping the reference value at the end of the window. Between these two extremes, the window size was increased to *k*, with *k/2* before *i* and *k/2* after *i*.

In this way, moving averages were calculated for maximum, minimum and average accelerations in the movement, crosswise and vertical directions. Furthermore, moving averages were also calculated for resultant and average resultant acceleration (acc_{res}, $acc_{\mathrm{avg.res}}$). After the original values were replaced with the moving averages, the differences between maximum and minimum accelerations ($acc_{\max}, acc_{\min}$) were calculated for all three directions ($cross, vert, mov$), and their differences subsequently calculated. Moreover, the differences between average accelerations (acc_{avg}) along the three directions were also calculated. Equations 6.3 to 6.9 show the complete procedure used for the difference calculations.

$$D_d = acc_{\max.d} - acc_{\min.d} \qquad for\ d = cross, vert, mov \tag{6.3}$$

$$D_1 = D_{cross} - D_{vert} - D_{mov} \tag{6.4}$$

$$D_2 = D_{vert} - D_{mov} - D_{cross} \tag{6.5}$$

$$D_3 = D_{mov} - D_{cross} - D_{vert} \tag{6.6}$$

$$D_{a1} = acc_{avg.cross} - acc_{avg.vert} - acc_{avg.mov} \tag{6.7}$$

$$D_{a2} = acc_{avg.vert} - acc_{avg.mov} - acc_{avg.cross} \tag{6.8}$$

$$D_{a3} = acc_{avg.mov} - acc_{avg.cross} - acc_{avg.vert} \tag{6.9}$$

Figure 6-7 shows the entire pre-processing method. After pre-processing, the final features were as follows: maximum, minimum and average acceleration along the three directions; differences between maximum and minimum (D_x, D_y, D_z); their differences (D_1, D_2, D_3); differences between average accelerations (D_{a1}, D_{a2}, D_{a3}); resultant acceleration and average resultant acceleration. In addition, moving averages were calculated for all values.

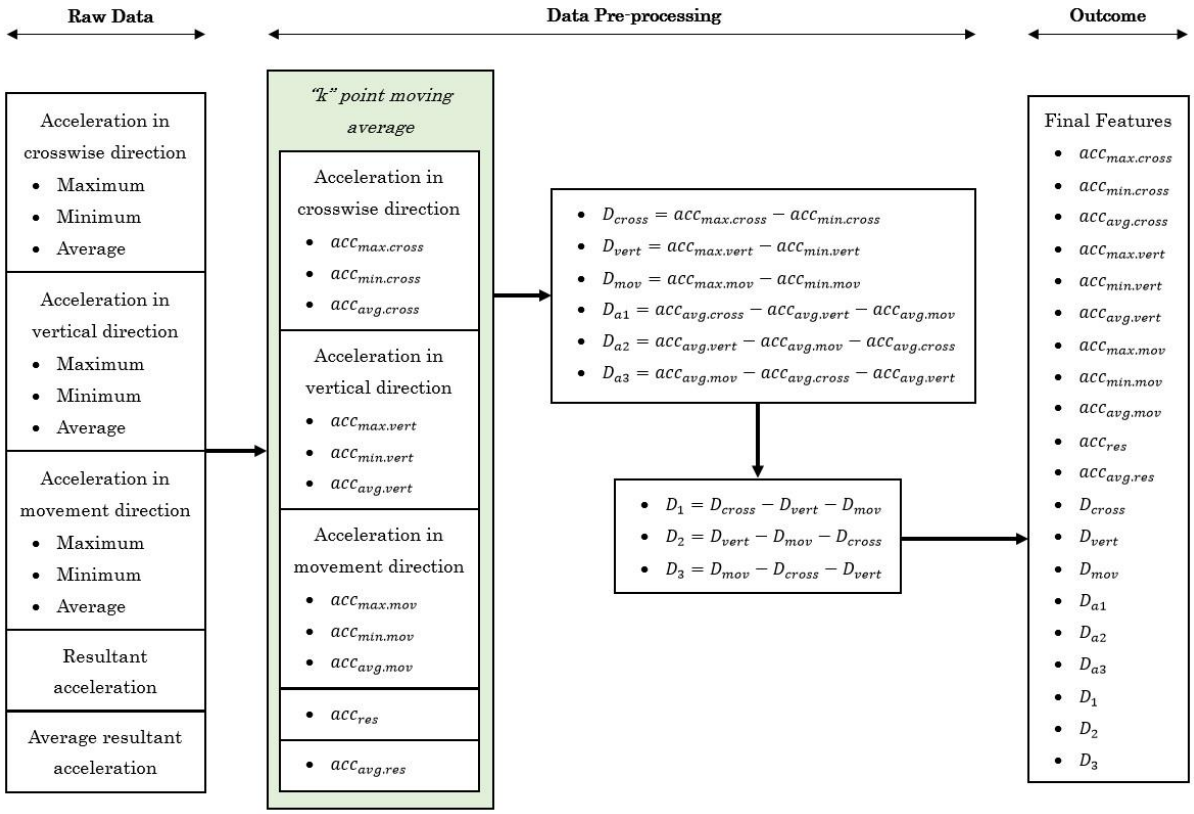


Figure 6-7: Pre-processing and feature extraction

6.2.5. Training and Testing Data Selection

As the data for each mode was different, the training data was randomly selected in the following two ways:

1) Equal number selection
2) Equal proportion selection

While equal number selection ensures that all the modes are equally represented in the training data set, the algorithm lacks sufficient training for the most frequently occurring mode in the test data set. Conversely, equal proportion selection ascertains that training is done proportionally for the test data set, but the modes are not represented equally in the training data set. This variation may affect the prediction.

Equal number selection

For each city, the mode with the least data was selected and the number corresponding to 70% of that data was calculated. The data equal to that number was then randomly selected from each mode to form the training data set, leaving the rest as a test data set. In this way, no matter how much difference was present between the modes, the training data always comprised equal numbers from each. Table 6-4 shows the amount of training data selected for each city.

Equal proportion selection

A total of 70% of data for each mode was randomly selected to form the training data and the remaining 30% was used to test the algorithms. This method yielded a much larger quantity of training data, which can be seen in Table 6-4.

6.2.6. Classifiers

In order to determine the classifier that most accurately predicts transportation mode, a comparison was made between (a) Support Vector Machines (SVM); (b) Adaptive

Boosting (AdaBoost); (c) decision tree using rpart, and (d) random forests. These classifiers were selected due to their frequent and established use in existing literature. The aim was to identify the best performing algorithm by carrying out a comparison between them.

6.3. Results and Discussion

The overall classification results of the classifiers for the different moving averages, as well as the two types of training data selection methods, are summarized in Figures 6-8 to 6-10. From the figures, it is evident that maximum prediction accuracy can be achieved by employing a 125-point moving average at the pre-processing stage. For the 125-point moving average, Table 6-5 shows the overall classification accuracies, while Table 6-6 gives the detailed results. The accuracy calculated can be considered producer accuracy.

Table 6-4: Amount of training data used for travel mode classification

City	Data Selection Method	Mode				
		Walk	Bicycle	Car	Train	Total
Niigata	Total	164,078	3,214	61,785	425	229,502
	Equal number	298	298	298	298	1,192
	Equal proportion	114,855	2,250	43,250	298	160,653
Gifu	Total	83,645	24,559	34,678	744	143,626
	Equal number	521	521	521	521	2,084
	Equal proportion	58,552	17,191	24,275	521	100,539
Matsuyama	Total	168,687	15,404	81,653	3,631	269,375
	Equal number	2,542	2,542	2,542	2,542	10,168
	Equal proportion	118,081	10,783	57,157	2,542	188,563

For example, if the prediction accuracy is 85%, this means that 85% of the known data carrying a certain class label (ground truth) is returned with the same label by the algorithm. The accuracies were calculated after creating confusion matrices and dividing the number of correct predictions for each mode by the total quantity of data in the test data set that is linked to that mode.

It can also be observed from the figures, as well as from the results listed in the tables, that the equal proportion method is better than the equal number method, but some of the detailed results show differently. For instance, in the equal proportion method, SVM and AdaBoost seem to perform well, with overall accuracy exceeding 85% in all cases. However, a breakdown of the accuracies at mode level reveals that the accuracy in terms of train transport prediction is very poor, in fact zero in case of Niigata and Matsuyama. This is because the amount of data corresponding to train transportation in the training data is relatively very small, which results in a zero prediction accuracy, even for the training data itself.

Random forest performs best in all cases. In particular, its accuracy is very high, at 99.8%, for the 125-point moving average using the equal proportion method. Even in the equal number method, the overall accuracy is greater than 91%, which is quite impressive. The next best performer is decision tree, followed by AdaBoost and then SVM.

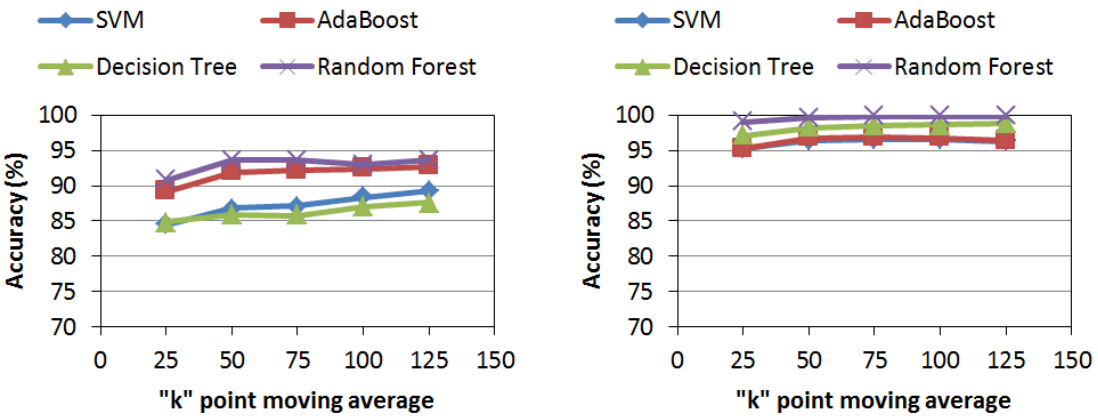


Figure 6-8: Prediction accuracy for Niigata city using (a) equal number method and (b) equal proportion method

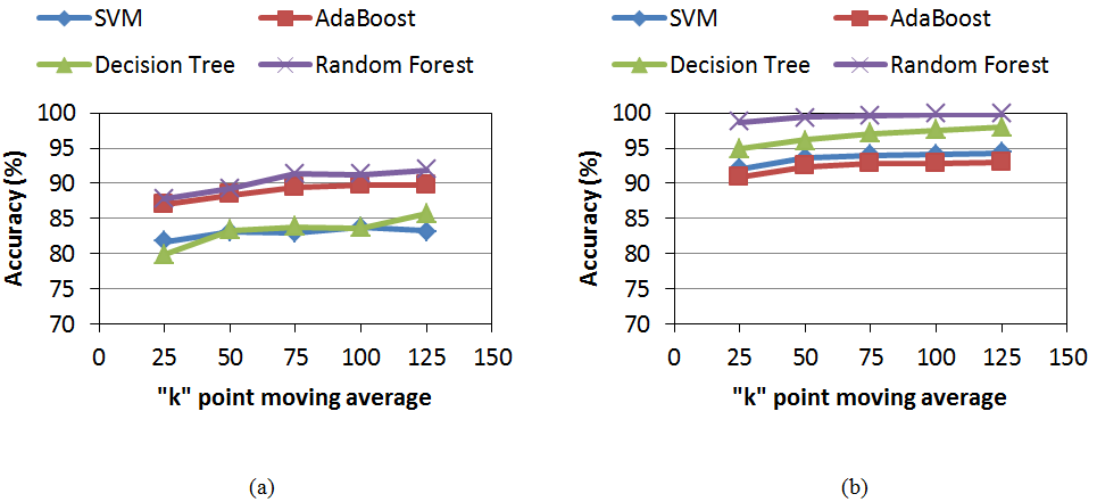


Figure 6-9: Prediction accuracy for Gifu city using (a) equal number method and (b) equal proportion method

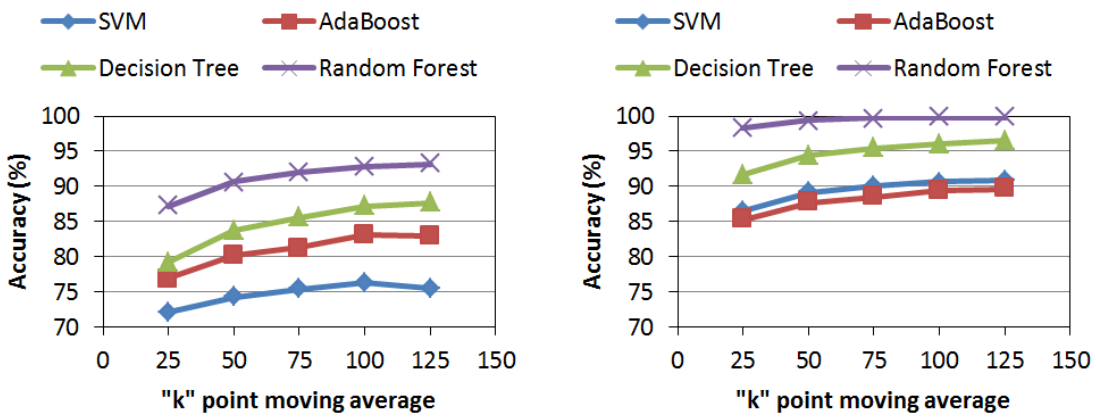


Figure 6-10: Prediction accuracy for Matsuyama city using (a) equal number method and (b) equal proportion method

The developed methodology was tested for three cities in order to establish the stability as well as the broader applicability of the approach. The results suggest that similar classification accuracy was achieved for the three cities. This is an indication that the approach is stable and might yield a good level of accuracy for other cities in Japan.

Table 6-5: Overall classification results at 125 point moving average

Selection method	City	SVM		AdaBoost		Decision Tree		Random Forest	
		Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
Equal number	Niigata	90.44	89.21	98.49	92.67	100.00	87.59	100.00	93.64
	Gifu	85.80	83.22	96.74	89.77	100.00	85.67	100.00	91.85
	Matsuyama	81.25	75.54	90.96	82.91	99.28	87.68	100.00	93.17
Equal proportion	Niigata	96.37	96.30	96.61	96.42	99.20	98.78	100.00	99.86
	Gifu	94.41	94.36	93.15	93.00	98.76	97.97	100.00	99.79
	Matsuyama	90.84	90.85	89.89	89.65	97.42	96.52	100.00	99.81

Table 6-6: Classification results at 125 point moving average

Selection method	City	Mode	SVM		AdaBoost		Decision Tree		Random Forest	
			Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
Equal number	Niigata	Walk	92.95	92.91	97.32	93.99	100.00	87.85	100.00	94.39
		Bicycle	90.94	90.05	98.99	96.54	100.00	91.50	100.00	97.81
		Car	77.85	79.27	97.65	88.96	100.00	86.69	100.00	91.43
		Train	100.00	100.00	100.00	100.00	100.00	97.64	100.00	100.00
		All	90.44	89.21	98.49	92.67	100.00	87.59	100.00	93.64
	Gifu	Walk	85.03	85.92	95.20	90.80	100.00	85.37	100.00	93.06
		Bicycle	90.98	87.16	96.93	89.82	100.00	83.36	100.00	90.88
		Car	74.09	73.78	94.82	87.16	100.00	87.92	100.00	89.57
		Train	93.09	94.62	100.00	100.00	100.00	99.55	100.00	100.00
		All	85.80	83.22	96.74	89.77	100.00	85.67	100.00	91.85
	Matsuyama	Walk	77.50	77.07	81.83	80.91	98.47	86.26	100.00	91.81
		Bicycle	87.25	86.71	94.26	91.10	99.76	93.40	100.00	98.27
		Car	71.75	70.32	88.28	85.56	98.94	89.60	100.00	95.11
		Train	88.51	87.51	99.49	99.54	99.96	97.52	100.00	100.00
		All	81.25	75.54	90.96	82.91	99.28	87.68	100.00	93.17
Equal proportion	Niigata	Walk	97.95	97.84	98.17	97.92	99.66	99.45	100.00	99.96
		Bicycle	77.96	79.88	84.76	86.41	94.13	92.22	100.00	99.38
		Car	93.81	93.70	93.55	93.37	98.32	97.44	100.00	99.64
		Train	0.00	0.00	30.54	37.01	89.60	85.04	100.00	99.21
		All	96.37	96.30	96.61	96.42	99.20	98.78	100.00	99.86
	Gifu	Walk	97.82	97.70	97.55	97.31	99.28	98.75	100.00	99.84
		Bicycle	88.10	88.04	82.18	81.49	97.91	96.73	100.00	99.78
		Car	91.46	91.57	90.93	91.43	98.23	97.14	100.00	99.71
		Train	56.05	56.05	63.92	61.88	92.90	89.69	100.00	98.65
		All	94.41	94.36	93.15	93.00	98.76	97.97	100.00	99.79
	Matsuyama	Walk	94.10	94.09	94.10	94.13	98.72	98.14	100.00	99.88
		Bicycle	64.50	63.75	51.79	49.12	89.58	87.06	100.00	99.70
		Car	93.12	93.29	91.47	91.17	96.97	95.95	100.00	99.80
		Train	0.00	0.00	20.38	19.47	80.45	74.38	100.00	97.43
		All	90.84	90.85	89.89	89.65	97.42	96.52	100.00	99.81

A careful examination of the results reveals that when using random forests, the prediction accuracy of the train transportation mode is the highest of all the modes, in

fact 100%, in the case of the equal number method. However, the same mode is predicted with the least accuracy relatively for the equal proportion method. This suggests that the prediction accuracy of the train mode can easily be improved by collecting more data so as to increase its representation in the training data set. Therefore, the optimum solution is to collect a comparable amount of data for each mode so that both selection methods will yield a training data set of a similar size.

This study highlights a limitation with respect to the SVM and AdaBoost algorithms. Minimal representation of the train transportation mode in the training data set following the equal proportion selection method resulted in the total misclassification of train data during prediction. This shows that the training of SVM requires equal or comparable representation from all classes, and the same is true for AdaBoost. On the other hand, no such constraints exist in the case of decision tree and random forests. A further observation was made regarding the computational time required by the algorithms. SVM and AdaBoost are very time consuming when it comes to large data sets like those used in this study, whereas decision tree and random forests outmatch them in this respect also.

However, the ideal scenario is to have a nearly equal amount of data for each contributing mode and then use the equal proportion method. In this manner, the strengths of both methods will be combined and yield even better prediction results. One of the limitations of this study relates to the fixed positioning of the data collection device while its carrier was travelling. The positioning should be flexible, especially in cases where purpose-built devices are replaced by smartphones. The newly

developed methodology needs to be modified and extended to incorporate varying placement of the device. Furthermore, the new approach should also be checked for additional modes. To this end, behavior models can also be incorporated into the analysis in order to enhance accuracy, and may be especially beneficial in the case of insufficient collected data.

Chapter 7 TRAVEL MODE DETECTION USING ACCELEROMETER AND GPS DATA AFTER PROCESSING*

The contents of this chapter cannot be published, because this chapter is scheduled to be published in an academic journal. It is planned to be posted within next 5 years.

* This chapter is to be published as Shafique M.A. and Hato E. Classification of Travel Data with Multiple Sensor Information using Random Forest. *Transportmetrica B.*

Chapter 8 COMPARISON OF CLASSIFICATION ALGORITHMS*

8.1. Introduction

Over the years, a lot of classification algorithms have been developed, and many among them, have been applied in the field of travel mode detection. For example, Neural Network (Byon et al., 2007; Gonzalez et al., 2008), Bayesian Network (Moiseeva and Timmermans, 2010; Zheng et al., 2008), Decision Tree (Reddy et al., 2010; Zheng et al., 2008), Support Vector Machine (Pereira et al., 2013; Zhang et al., 2011; Zheng et al., 2008), Random Forest (Shafique and Hato, 2015) etc.

The aim of the current chapter is to compare the performance of various classification algorithms for the purpose of travel mode identification. The comparison is done by taking two criteria into account, accuracy and computational time. Furthermore, the algorithms are not applied by taking the default values of the associated variables as it is. Rather, within each algorithm, a comparison is done with varying values of the variables involved.

---

* This chapter is to be published as part of Shafique M.A. and Hato E. Improving the Accuracy of Travel Mode Detection for low Data Collection Frequencies. *Transportation Research Part C.*

This chapter is presented in part as Shafique M.A. and Hato E. A Comparison among various Classification Algorithms for Travel Mode Detection using Sensors' data collected by Smartphones. *14th International Conference on Computers in Urban Planning and Urban Management, CUPUM 2015*, MIT, Cambridge, Massachusetts. July 2015.

8.2. Methodology

8.2.1. Data Collection

Smartphones were used by 50 participants from Kobe city, Japan, to collect travel data over a period of one month, while using seven different modes of transportation namely walk, bicycle, motor bike, car, bus, train and subway. The collected data consisted of GPS, accelerometer and gyroscope readings. Although, the sensors' data was recorded at an average frequency of 14 readings per second but for the current study, the frequency was scaled down to 1 reading per 5 seconds, because low data collection frequency is more energy-efficient. A big drawback with smartphones as data collection devices is the quick drainage of battery-time. Table 8-1 provides the amount of data and the number of trips for each mode, used in this study. As this study tends to provide a comparison among different algorithms, therefore GPS data was dropped because it requires much more processing, and only accelerometer and gyroscope data was used.

8.2.2. Feature Extraction

The raw data consisted of accelerometer data (accelerations in x, y and z directions) and gyroscope data (pitch and roll). Due to the different positions in which the smartphones were carried by each participant, the resultant acceleration was calculated from the individual accelerations and was used for feature extraction.

For the purpose of smoothening the data and reducing the effect of the outliers, the concept of moving window was used where the readings covered in a certain amount of time, known as the window size, were used to apply an operation (e.g. average, maximum etc.) at a certain data entry level and this window moved downwards as the calculations proceeded along the data column.

Using a window size of 5 minutes, maximum resultant accelerations, average resultant accelerations and maximum average resultant accelerations were calculated from the resultant acceleration values. Furthermore, standard deviation, skewness and kurtosis were also calculated. These calculated features along with the recorded features by gyroscope (pitch and roll) were used to train and test each algorithm. The training dataset was formed by randomly selecting 10% of data from each mode class and the rest was used to form the test dataset.

Table 8-1: Amount of data used in the study

Mode	Amount of data	No. of trips
Walk	146,973	442
Bicycle	9,098	10
Motor Bike	6,121	1
Car	13,981	31
Bus	10,666	21
Train	18,423	45
Subway	6,520	10

8.2.3. Classification Algorithms

The classification algorithms used in this chapter have already been introduced in chapter 3. Therefore, without repeating the introductions, the values of variables associated with each algorithm are only provided,

Support Vector Machines

For comparison, SVM was applied repeatedly using linear, RBF and polynomial kernels. For RBF kernel, gamma (γ) value was changed from 20 to 1E-06. Whereas for polynomial kernel, gamma (γ) value was changed from 0.1 to 1E-06 and degree (d) from 1 to 6. The default values of gamma and degree usually used were 4.7E-05 (1/data dimension) and 3 respectively.

Neural Networks

For neural networks, the number of units in the hidden layer or size was varied from 30 to 50 and maximum number of iterations (default 100) ranged from 100 to 500.

Decision trees

In case of simple decision trees, minimum number of observations for the split to take place was reduced from 20 (default) to 2. The complexity parameter (cp) was varied from 0.1 to 1E-05. In case of boosted decision trees, SAMME was applied with the complexity parameter ranging from 1E-02 to 1E-05.

Random Forest

Sampling was done with and without replacement, while the number of trees in the forest was varied from 100 to 200.

Naïve Bayes

There is no variable associated with Naïve Bayes, whose varying value can be checked for comparison within the algorithm.

8.3. Results and Discussion

Each algorithm was tested by manually varying the variables involved, rather than automatically tuning the algorithm to identify the most suitable values, because the aim was to observe the computational time for each change so as to gain an indicator (time) for the comparison of algorithms. All the calculations were performed on an Intel core i7 3.50 GHz with 32 GB RAM.

In case of SVM, the prediction accuracies (ratio of data of a certain class correctly labelled by algorithm to entire data of that certain class) for linear kernel and RBF kernel (with varying gamma values) are shown in Table 8-2, whereas the results for polynomial kernel are given in Table 8-3. All results for polynomial kernel are not shown in Table 8-3 because those variable values were skipped for which the entire data was labeled as walk. The results propose that both linear and polynomial kernels are not suitable for smartphone data. Using RBF kernel, the overall accuracy is maximum when gamma has a value of 10, but a gamma value of 1 gives equally good results with much less computational time. Furthermore, close inspection of the results suggest that gamma = 1 is actually yielding better results mode-wise. Because the amount of data for walk is more that 50% the entire data, therefore a slight increase in its prediction accuracy (in case of gamma = 10) made it look like a better option.

The results for neural networks are shown in tables 8-4 and 8-5. The overall prediction

accuracy improves as the number of weights is increased by increasing the size and maximum iterations. The maximum accuracy is achieved for size 50 and iterations 500, above which the algorithm is unable to perform due to too many weights. The complexity parameter in decision trees determines the pruning of the tree. The results shown in table 8-6 demonstrate that the maximum overall accuracy of the decision trees can be achieved for cp value of 0.0001. But if the decision trees are boosted, then the prediction accuracy jumps up by around 4% (Table 8-7).

Table 8-2: Prediction results for SVM (Linear and RBF kernels)

Mode	Prediction Accuracy (%)									
Walk	100.00	99.99	99.94	**98.73**	99.14	99.98	100.00	100.00	100.00	100.00
Bicycle	0.00	60.11	71.17	**77.20**	50.48	8.39	0.00	0.00	0.00	0.00
Motor Bike	0.00	70.86	80.52	**89.27**	62.84	1.29	0.00	0.00	0.00	0.00
Car	0.00	61.79	73.05	**76.52**	3.56	0.00	0.00	0.00	0.00	0.00
Bus	0.00	67.72	78.48	**82.25**	47.74	0.00	0.00	0.00	0.00	0.00
Train	0.00	55.48	63.70	**61.66**	18.25	0.00	0.00	0.00	0.00	0.00
Subway	0.00	49.23	57.63	**60.04**	34.76	5.61	0.00	0.00	0.00	0.00
Overall	69.40	87.85	90.83	**90.82**	78.08	69.96	69.40	69.40	69.40	69.40
Kernel	Linear	RBF								
Gamma	-	20	10	1	0.1	0.01	0.001	0.0001	0.00001	0.000001
Computational time (sec)	74.03	1311.99	1202.65	281.82	217.57	298.85	298.4	269.81	238.7	235.26

Table 8-3: Prediction results for SVM (Polynomial kernel)

Mode	Prediction Accuracy (%)									
Walk	99.93	100.00	99.52	100.00	99.60	100.00	99.49	99.99	99.58	99.99
Bicycle	2.92	0.00	43.93	2.60	36.79	2.72	42.88	2.72	36.48	2.72
Motor Bike	31.21	0.60	34.51	1.16	35.84	1.16	37.93	1.29	37.36	1.29
Car	0.00	0.00	0.78	0.00	4.20	0.00	6.32	0.00	7.38	0.00
Bus	0.00	0.00	2.41	0.00	4.20	0.00	8.61	0.00	4.57	0.00
Train	0.00	0.00	1.04	0.00	18.45	0.00	18.86	0.00	20.01	0.00
Subway	0.00	0.00	33.95	0.00	29.12	0.00	38.91	0.00	38.60	0.00
Overall	70.38	69.42	73.26	69.54	74.73	69.55	75.67	69.55	75.40	69.54
Degree	2		3		4		5		6	
Gamma	0.1	0.01	0.1	0.01	0.1	0.01	0.1	0.01	0.1	0.01
Computational time (sec)	85.44	75.98	113.81	75.16	88.2	76.86	111.73	76.35	99.62	73.45

Table 8-4: Prediction results for Neural Networks

Mode	Prediction Accuracy (%)													
Walk	97.77	97.48	96.00	96.27	95.61	96.17	96.98	95.24	96.71	94.87	95.35	96.01	96.01	96.18
Bicycle	36.53	54.79	55.30	58.70	63.57	57.71	65.71	65.56	66.90	52.31	53.97	48.88	55.45	59.32
Motor Bike	53.16	63.16	61.71	69.30	71.01	71.50	65.63	68.06	80.52	62.98	68.06	71.68	71.90	72.59
Car	18.81	23.46	26.14	23.30	42.26	28.81	23.17	21.51	20.59	22.75	31.47	26.54	37.37	30.02
Bus	0.04	0.67	36.21	42.56	47.20	53.40	0.34	37.14	58.97	41.32	44.20	52.54	52.64	40.19
Train	20.03	22.26	22.67	22.86	24.72	23.55	22.70	24.03	22.07	23.67	20.80	22.52	23.72	23.63
Subway	2.37	8.61	24.76	23.19	40.95	37.88	4.36	46.17	25.75	26.31	22.05	31.65	42.60	24.71
Overall	74.02	75.61	77.06	77.72	79.71	79.09	75.68	77.68	79.36	76.36	77.25	78.13	79.58	78.22
Size	30									40				
Max. iterations	100	150	200	250	300	350	400	450	500	100	150	200	250	300
Computational time (sec)	13.84	22.88	26.35	34.82	37.3	40.97	43.44	51.18	60.21	19.31	27.66	33.93	42.4	52.97

Table 8-5: Prediction results for Neural Networks (Cont.)

Mode	Prediction Accuracy (%)												
Walk	96.09	95.64	95.74	95.18	96.81	94.87	95.28	94.67	95.52	95.34	96.19	95.37	**95.47**
Bicycle	60.66	64.66	64.00	63.81	40.29	48.30	49.73	57.22	59.18	69.28	61.86	68.32	**68.94**
Motor Bike	72.80	76.38	75.25	74.60	65.78	67.77	71.21	71.15	71.88	76.47	73.33	77.51	**77.72**
Car	32.06	36.59	41.18	40.07	21.10	26.32	29.92	32.42	32.45	31.56	30.85	49.04	**40.92**
Bus	47.67	56.64	53.57	42.68	28.64	54.89	46.17	50.29	50.88	58.21	51.36	56.42	**60.26**
Train	22.96	22.93	24.80	27.16	20.92	27.15	27.24	29.62	28.09	29.61	22.74	29.49	**33.59**
Subway	29.75	34.73	41.65	41.96	26.69	17.50	9.13	34.71	37.42	25.00	28.61	42.48	**43.18**
Overall	78.82	79.69	80.22	79.40	76.30	77.27	77.27	78.53	79.21	79.71	79.01	81.31	**81.45**
Size	40				50								
Max. iterations	350	400	450	500	100	150	200	250	300	350	400	450	500
Computational time (sec)	53.35	62.39	71.32	74.12	21.25	34.52	43.09	52.4	58.7	68.07	78.43	86.03	94.46

Table 8-6: Prediction results for Decision Trees

Mode	Prediction Accuracy (%)				
Walk	100.00	99.18	97.01	**96.32**	95.84
Bicycle	0.00	37.02	85.75	**94.04**	94.16
Motor Bike	0.00	28.69	79.56	**93.26**	93.48
Car	0.00	0.00	61.72	**87.26**	88.18
Bus	0.00	44.56	69.88	**88.62**	88.97
Train	0.00	21.23	63.41	**85.57**	85.94
Subway	0.00	29.52	47.48	**84.15**	85.77
Overall	69.40	76.25	87.88	**93.84**	93.68
Complexity parameter	0.1	0.01	0.001	0.0001	0.00001
Computational time (sec)	0.21	0.87	1.28	2.23	2.44

Table 8-7: Prediction results for Boosted Decision Trees

Mode	Prediction Accuracy (%)			
Walk	88.05	99.68	**99.86**	99.81
Bicycle	68.80	96.89	**96.87**	96.62
Motor Bike	73.02	96.79	**98.00**	97.68
Car	48.74	94.28	**95.12**	94.93
Bus	60.63	92.22	**92.72**	92.35
Train	57.14	89.57	**90.74**	90.71
Subway	52.01	87.05	**87.71**	86.45
Overall	79.01	97.48	**97.84**	97.71
Complexity parameter	0.01	0.001	0.0001	0.00001
Computational time (sec)	94.48	123.83	191.15	214.78

In case of random forest, sampling without replacement provides slightly better results than with replacement (Table 8-8). Moreover, the increase in overall prediction accuracy is minimal with the increase in the number of trees beyond 100. In order to provide a specific value for the suitable number of trees, 150 will do as it provides high accuracy along with saving some computational time. A peek into the results of naïve Bayes, given in Table 8-9, reveals that its performance is least, in comparison to all the algorithms discussed.

A comprehensive comparison is provided in table 8-10. Here it can be seen that boosted decision trees provide the highest prediction accuracy but are not the most efficient classifier, as is evident from the computational time. Although, the accuracy achieved by random forest is slightly lower than by boosted decision trees, the computation is very quick making it a better option, especially when the data is huge. Decision trees are very quick but the prediction is not very accurate. SVM is the most time-consuming classifier, with accuracy even lower than decision trees. Neural network and Naïve Bayes come last in the list.

Table 8-8: Prediction results for Random Forest

Mode	Prediction Accuracy (%)									
Walk	99.81	99.82	99.84	99.82	99.82	99.82	99.82	99.81	99.83	99.83
Bicycle	95.48	95.68	96.02	95.95	95.92	96.01	96.06	96.08	95.79	96.09
Motor Bike	97.57	97.28	97.60	97.69	97.57	97.51	97.35	97.64	97.31	97.42
Car	93.10	93.40	93.54	93.37	93.40	93.42	93.75	93.49	93.67	93.72
Bus	90.74	91.09	90.82	91.33	91.18	91.17	91.64	91.43	91.30	91.42
Train	87.41	88.26	87.91	87.77	87.80	88.50	88.07	88.54	88.43	88.46
Subway	83.13	84.30	84.03	84.25	83.86	85.51	83.61	84.94	84.59	85.00
Overall	97.07	97.22	97.21	97.21	97.19	97.30	97.26	97.31	97.28	97.32
Replacement	True					False				
No. of trees	100	125	150	175	200	100	125	150	175	200
Computational time (sec)	3.92	4.4	5.14	5.96	6.71	3.38	4.02	4.85	5.65	6.34

Table 8-9: Prediction results for Naïve Bayes

Mode	Prediction Accuracy (%)
Walk	62.40
Bicycle	67.13
Motor Bike	57.30
Car	14.89
Bus	67.45
Train	3.34
Subway	4.02
Overall	52.64
Computational time (sec)	54.1

Table 8-10: Comparison of Classification Algorithms

Mode	Prediction Accuracy (%)					
	SVM	NN	DT	BDT	RF	NB
Walk	98.73	95.47	96.32	99.86	99.81	62.40
Bicycle	77.20	68.94	94.04	96.87	96.08	67.13
Motor Bike	89.27	77.72	93.26	98.00	97.64	57.30
Car	76.52	40.92	87.26	95.12	93.49	14.89
Bus	82.25	60.26	88.62	92.72	91.43	67.45
Train	61.66	33.59	85.57	90.74	88.54	3.34
Subway	60.04	43.18	84.15	87.71	84.94	4.02
Overall	90.82	81.45	93.84	97.84	97.31	52.64
Computational time (sec)	281.82	94.46	2.23	191.15	4.85	54.1

SVM = Support Vector machine

NN = Neural Network

DT = Decision Tree

BDT = Boosted Decision Tree

RF = Random Forest

NB = Naïve Bayes

This chapter provides an analysis of the performance of each algorithm by varying the associated variables and offers a comparison among the algorithms. The results suggest that random forest and boosted decision trees both provide good prediction accuracies but random forest is relatively very quick and thus is more suitable for identification of mode of transportation by employing the sensors' data collected by smartphones. If the detection is required very quickly, then decision trees can also be used but the accuracy will fall. This study will assist other researchers in selection of classification algorithm. Although, the conclusion drawn by this study holds good for the travel mode detection, for other problems similar study should be carried out to ascertain the suitable algorithm.

Chapter 9 TRAVEL MODE DETECTION USING ACCELEROMETER AND GYROSCOPE DATA*

The contents of this chapter cannot be published, because this chapter is scheduled to be published in an academic journal. It is planned to be posted within next 5 years.

* This chapter is to be published as Shafique M.A. and Hato E. Travel Mode Detection with varying Smartphone Data Collection Frequencies. *Transportation*.

Chapter 10 TRAVEL MODE DETECTION USING MULTIPLE SENSORS' DATA*

The contents of this chapter cannot be published, because this chapter is scheduled to be published in an academic journal. It is planned to be posted within next 5 years.

* This chapter is to be published as Shafique M.A. and Hato E. Improving the Accuracy of Travel Mode Detection for low Data Collection Frequencies. *Transportation Research Part C*.

Chapter 11 TRAVEL MODE DETECTION BY MERGING MACHINE LEARNING AND MNL MODEL*

The contents of this chapter cannot be published, because this chapter is scheduled to be published in an academic journal. It is planned to be posted within next 5 years.

* This chapter is to be published as Shafique M.A. and Hato E. Incorporating MNL model into Random Forest for Travel mode detection.

Chapter 12 CONCLUSION

12.1. Introduction

This chapter summarizes the research provided in the previous chapters. Furthermore, recommendations for further study are also mentioned.

12.2. Research Summary

Sensors like GPS and accelerometer are opening up a new horizon for introduction of technology to solve problems in the transportation sector. Travel data collection method can be revolutionized by employing devices carrying multiple sensors for passive data recording. This vast possibility is identified by researchers all over the world and a lot of research is being undertaken. The present dissertation is expected to contribute to the ongoing research. It presents a development procedure wherein each chapter refines the methodology by working on the lessons learned from the previous chapters and the development culminates into a final efficient, workable, sustainable and accurate methodology provided in the final chapter.

The main findings of this dissertation are,

- Travel mode can be identified with sufficient accuracy by using the data collected by sensors like GPS, accelerometer and gyroscope.
- When using accelerometer data, magnitude of resultant acceleration is a better variable than individual accelerations along the 3 axes, because it permits the

smartphones to be carried around in any position feasible for the participants.

- For feature extraction, moving window is better than 50 % overlapping window, as the amount of data is not decreased.
- Increasing the moving window size improves the prediction accuracy for longer trips, at the cost of smaller ones. A window size of 10 minutes is therefore a suitable trade-off, which is less than the average walking time of 16.15 minutes per trip.
- From resultant acceleration, additional features extracted like maximum resultant acceleration, average resultant acceleration, maximum average resultant acceleration, standard deviation, skewness and kurtosis prove to be quite valuable for classification. Similar features extracted from speed (calculated from GPS data) can further improve the classification accuracy.
- The prediction accuracy increases with increase in the proportion of data used to train the algorithm. Increasing the proportion of training data means that a large amount of data with known travel modes needs to be collected in order to identify a relatively small amount of unknown data. It is therefore necessary to limit the proportion of training data so that a small dataset can predict a larger one. 10 % training data is found to be at the threshold. Further decrease in the proportion results in a steep fall in accuracy.
- Among various classification algorithms mostly employed for travel mode detection, boosted decision trees and random forest provide the best results. Random forest is preferable due to less computational time required.
- Imbalanced data results in abnormally high prediction accuracy for the majority mode, whereas the minority modes show low accuracy primarily due to misclassification as the majority class.

- Weighted random forest can solve the problem of imbalanced data, to some extent.
- Down-sampling using mean as the threshold value can provide additional improvement for imbalanced data.
- A 2-step post-processing method introduced can further refine the results.
- Binomial logistic regression can be used to classify the travel modes by utilizing the features extracted from sensors' data.
- Data collection frequency directly affects the battery usage of the recording device. But at the same time, the classification accuracy drops with decrease in collection frequency. A compromise has to be made between accuracy and battery saving.
- MNL model used in combination with machine learning can greatly improve the mode detection accuracy.

12.3. Further Studies

For every research, there is always some scope of improvement. Although, the findings reported are concrete, many aspects remained to be investigated. Thus the current research can be improved and extended in a number of ways, some of which are mentioned below.

- Data from different cities and different countries need to be utilized so that the developed methodology will be applicable in every region.
- Excessive data is required to validate the developed methodology.
- Transferability of the methodology should also be checked and assured.
- Binomial regression analysis introduced in the current research should be

improved further.

- Detailed attributes should be used in MNL model so that fitness can be improved.
- Refinement in the methodology should be strived for so that accuracy can be further improved for low data collection frequency.

References

Abdulazim, T., Abdelgawad, H., Habib, K.M.N., Abdulhai, B., 2013. Using Smartphones and Sensor Technologies to Automate Collection of Travel Data. *Transportation Research Record: Journal of the Transportation Research Board* 2383(1), 44-52.

Anderson, J.A., 1972. A simple neural network generating an interactive memory. *Mathematical Biosciences* 14(3), 197-220.

Auld, J., Williams, C., Mohammadian, A., Nelson, P., 2009. An automated GPS-based prompted recall survey with learning algorithms. *Transportation Letters* 1(1), 59-79.

Axhausen, K.W., Schönfelder, S., Wolf, J., Oliveira, M., Samaga, U., 2003. 80 weeks of GPS-traces: Approaches to enriching the trip information. Citeseer.

Bachu, P.K., Dudala, T., Kothuri, S.M., 2001. Prompted recall in global positioning system survey: Proof-of-concept study. *Transportation Research Record: Journal of the Transportation Research Board* 1768(1), 106-113.

Ben-Hur, A., Weston, J., 2010. A user's guide to support vector machines. *Methods in molecular biology (Clifton, N.J.)* 609, 223-239.

Besag, J.E., 1972. Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 75-83.

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

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.

Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J., 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers,*

Environment and Urban Systems 36(6), 526-537.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, Pittsburgh, Pennsylvania, USA, pp. 144-152.

Breiman, L., 2001. Random Forests. *Machine Learning* 45(1), 5-32.

Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J., 1984. *Classification and regression trees*. Wadsworth International Group, Belmont, Calif.

Byon, Y.-J., Abdulhai, B., Shalaby, A., 2009. Real-time transportation mode detection via tracking global positioning system mobile devices. *Journal of Intelligent Transportation Systems* 13(4), 161-170.

Byon, Y.-J., Abdulhai, B., Shalaby, A.S., 2007. Impact of sampling rate of GPS-enabled cell phones on mode detection and GIS map matching performance, *Transportation Research Board 86th Annual Meeting*.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 161-168.

Cestnik, B., 1990. Estimating probabilities: a crucial task in machine learning, *ECAI*, pp. 147-149.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16(1), 321-357.

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.

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L.,

Rahimi, A., Rea, A., Bordello, G., Hemingway, B., 2008. The mobile sensing platform: An embedded activity recognition system. *Pervasive Computing, IEEE* 7(2), 32-41.

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

Clark, P., Niblett, T., 1989. The CN2 induction algorithm. *Machine Learning* 3(4), 261-283.

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

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.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29(2-3), 103-130.

Draijer, G., Kalfs, N., Perdok, J., 2000. Global positioning system as data collection method for travel research. *Transportation Research Record: Journal of the Transportation Research Board* 1719(1), 147-153.

Duda, R.O., Hart, P.E., 1973. *Pattern classification and scene analysis*. Wiley New York.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., Kerr, J., 2014. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Frontiers in public health* 2.

Ettema, D.F., Timmermans, H.J.P., Van Veghel, L., 1996. Effects of Data Collection Methods in Travel and Activity Research. *European Institute of Retailing and Service Studies*.

Feng, T., Timmermans, H.J., 2013. Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies* 37, 118-

130.

Figo, D., Diniz, P.C., Ferreira, D.R., Cardoso, J.M., 2010. Preprocessing techniques for context recognition from accelerometer data. *Personal Ubiquitous Comput.* 14(7), 645-662.

Frank, E., Trigg, L., Holmes, G., Witten, I.H., 2000. Technical note: Naive Bayes for regression. *Machine Learning* 41(1), 5-25.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm, *ICML*, pp. 148-156.

Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55(1), 119-139.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28(2), 337-407.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* 36(4), 193-202.

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

Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia-Social and Behavioral Sciences* 138, 557-565.

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N.L., Perez, R., 2008. Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones, *15th World congress on intelligent transportation systems*.

Greaves, S.P., 2006. A Panel Approach to Evaluating Voluntary Travel Behavior Change Programs-South Australia Pilot Survey, *Transportation Research Board 85th Annual Meeting*.

Grossberg, S., 1988. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks* 1(1), 17-61.

Hato, E., 2006. Development of MoALs (Mobile Activity Loggers supported by gps-phones) for travel behavior analysis, *Transportation Research Board 85th Annual Meeting*.

Hato, E., 2010. Development of behavioral context addressable loggers in the shell for travel-activity analysis. *Transportation Research Part C: Emerging Technologies* 18(1), 55-67.

Hemminki, S., Nurmi, P., Tarkoma, S., 2013. Accelerometer-based transportation mode detection on smartphones, *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, p. 13.

Hudson, J.G., Duthie, J.C., Rathod, Y.K., Larsen, K.A., Meyer, J.L., 2012. Using smartphones to collect bicycle travel data in Texas.

Huss, A., Beekhuizen, J., Kromhout, H., Vermeulen, R., 2014. Using GPS-derived speed patterns for recognition of transport modes in adults. *International journal of health geographics* 13(1), 40.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biological cybernetics* 43(1), 59-69.

Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced data sets: One sided sampling, *Proc. of the Int'l Conf. on Machine Learning*.

Kwapisz, J.R., Weiss, G.M., Moore, S.A., 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12(2), 74-82.

Langley, P., Iba, W., Thompson, K., 1992. An analysis of Bayesian classifiers, *AAAI*,

pp. 223-228.

Lee-Gosselin, M.E., Doherty, S.T., Papinski, D., 2006. Internet-based prompted recall diary with automated gps activity-trip detection: System design, *Transportation Research Board 85th Annual Meeting*.

Lester, J., Choudhury, T., Borriello, G., 2006. A Practical Approach to Recognizing Physical Activities, in: Fishkin, K., Schiele, B., Nixon, P., Quigley, A. (Eds.), *Pervasive Computing*. Springer Berlin Heidelberg, pp. 1-16.

Lewis, D.D., 1998. Naive (Bayes) at forty: The independence assumption in information retrieval, *Machine learning: ECML-98*. Springer, pp. 4-15.

Ling, C.X., Li, C., 1998. Data Mining for Direct Marketing: Problems and Solutions, *KDD*, pp. 73-79.

Maurer, U., Smailagic, A., Siewiorek, D.P., Deisher, M., 2006. Activity recognition and monitoring using multiple sensors on different body positions, *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*. IEEE, pp. 4 pp.-116.

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

McGowen, P., McNally, M., 2007. Evaluating the potential to predict activity types from GPS and GIS data, *Transportation Research Board 86th meeting, Jan*, p. 21.

Minsky, M., Papert, S., 1969. Perceptron: an introduction to computational geometry. *The MIT Press, Cambridge, expanded edition* 19, 88.

Mitchell, T.M., 1997. Artificial neural networks. *Machine Learning*, 81-127.

Moiseeva, A., Timmermans, H., 2010. Imputing relevant information from multi-day GPS tracers for retail planning and management using data fusion and context-sensitive learning. *Journal of Retailing and Consumer Services* 17(3), 189-199.

Murakami, E., Wagner, D.P., 1999. Can using global positioning system (GPS) improve trip reporting? *Transportation research part c: emerging technologies* 7(2), 149-165.

Nham, B., Siangliulue, K., Yeung, S., 2008. Predicting mode of transport from iphone accelerometer data. *Machine Learning Final Projects, Stanford University*.

Nick, T., Coersmeier, E., Geldmacher, J., Goetze, J., 2010. Classifying means of transportation using mobile sensor data, *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1-6.

Nitsche, P., Widhalm, P., Breuss, S., Maurer, P., 2012. A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. *Procedia-Social and Behavioral Sciences* 48, 1033-1046.

Pereira, F., Carrion, C., Zhao, F., Cottrill, C.D., Zegras, C., Ben-Akiva, M., 2013. The Future Mobility Survey: Overview and Preliminary Evaluation, *Proceedings of the Eastern Asia Society for Transportation Studies*.

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

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. *ACM Trans. Sen. Netw.* 6(2), 1-27.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65(6), 386.

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21(3), 660-674.

Sankaran, K., Zhu, M., Guo, X.F., Ananda, A.L., Chan, M.C., Peh, L.-S., 2014. Using mobile phone barometer for low-power transportation context detection, *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. ACM, pp. 191-205.

Santos, A., McGuckin, N., Nakamoto, H.Y., Gray, D., Liss, S., 2011. Summary of travel trends: 2009 national household travel survey.

Schapire, R.E., Freund, Y., 2012. *Boosting: Foundations and algorithms*. MIT press.

Schuessler, N., Axhausen, K., 2009. Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record: Journal of the Transportation Research Board* 2105(1), 28-36.

Sermons, M.W., Koppelman, F.S., 1996. Use of vehicle positioning data for arterial incident detection. *Transportation Research Part C: Emerging Technologies* 4(2), 87-96.

Shafique, M.A., Hato, E., 2014. Use of acceleration data for transportation mode prediction. *Transportation*, 1-26.

Shafique, M.A., Hato, E., 2015. Use of acceleration data for transportation mode prediction. *Transportation* 42(1), 163-188.

Shawe-Taylor, J., Cristianini, N., 2004. *Kernel methods for pattern analysis*. Cambridge University Press, Cambridge, UK ; New York.

Shen, L., Stopher, P., 2013. Should we change the rules for trip identification for GPS travel records, *Proceedings of the 36th Australasian Transport Research Forum ATRF*.

Shen, L., Stopher, P.R., 2014. Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews* 34(3), 316-334.

Shin, D., Aliaga, D., Tunçer, B., Arisona, S.M., Kim, S., Zünd, D., Schmitt, G., 2014. Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems*.

Stenneth, L., Wolfson, O., Yu, P.S., Xu, B., 2011. Transportation mode detection using mobile phones and GIS information, *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, Chicago, Illinois, pp. 54-63.

Stopher, P., FitzGerald, C., Zhang, J., 2008. Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies* 16(3), 350-369.

Stopher, P.R., 1992. Use of an activity-based diary to collect household travel data. *Transportation* 19(2), 159-176.

Stopher, P.R., 2009. The travel survey toolkit: where to from here. *Transport survey methods, keeping up with a changing world*, 15-46.

Stopher, P.R., Bullock, P., Horst, F., 2002. *Exploring the use of passive GPS devices to measure travel*. Institute of Transport Studies.

Tragopoulou, S., Varlamis, I., Eirinaki, M., 2014. Classification of movement data concerning user's activity recognition via mobile phones, *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*. ACM, p. 42.

Tsui, S.Y.A., Shalaby, A., 2006. Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*(1972), 38-45.

Wagner, D., 1997. Global positioning systems for personal travel surveys: Lexington area travel data collection test. *Report to the Federal Highway Administration, US DOT by Battelle Transportation Division*.

Wang, L.-X., 1994. *Adaptive fuzzy systems and control: design and stability analysis*. Prentice-Hall, Inc.

Widhalm, P., Nitsche, P., Brandie, N., 2012. Transport mode detection with realistic Smartphone sensor data, *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, pp. 573-576.

Wolf, J., 2000. Using GPS data loggers to replace travel diaries in the collection of travel data. Citeseer.

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

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board* 1768(1), 125-134.

Xia, H., Qiao, Y., Jian, J., Chang, Y., 2014. Using smart phone sensors to detect transportation modes. *Sensors* 14(11), 20843-20865.

Xiao, Y., Low, D., Bandara, T., Pathak, P., Lim, H.B., Goyal, D., Santos, J., Cottrill,

C., Pereira, F., Zegras, C., 2012. Transportation activity analysis using smartphones, *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*. IEEE, pp. 60-61.

Yang, J., 2009. Toward physical activity diary: motion recognition using simple acceleration features with mobile phones, *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*. ACM, pp. 1-10.

Yu, M.-C., Yu, T., Lin, C., Chang, E., 2013. Low power and low cost sensor hub for transportation-mode detection. *Studio Engineering, HTC, Tech. Rep*.

Yu, M.-C., Yu, T., Wang, S.-C., SC, D., Lin, C.-J., Chang, E.Y., 2014. Big data small footprint: The design of a low-power classifier for detecting transportation modes. *Proceedings of the VLDB Endowment* 7, 1429-1440.

Zhang, L., Dalyot, S., Eggert, D., Sester, M., 2011. Multi-stage approach to travel-mode segmentation and classification of gps traces, *ISPRS Workshop on Geospatial Data Infrastructure: from data acquisition and updating to smarter services*.

Zheng, Y., Liu, L., Wang, L., Xie, X., 2008. Learning transportation mode from raw gps data for geographic applications on the web, *Proceedings of the 17th international conference on World Wide Web*. ACM, Beijing, China, pp. 247-256.

Zhu, J., Zou, H., Rosset, S., Hastie, T., 2009. Multi-class adaboost. *Statistics and its Interface* 2(3), 349-360.

Zimowski, M., Tourangeau, R., Ghadialy, R., Pedlow, S., 1997. Nonresponse in household travel surveys. US Department of Transportation.

Zito, R., d'Este, G., Taylor, M.A., 1995. Global positioning systems in the time domain: how useful a tool for intelligent vehicle-highway systems? *Transportation Research Part C: Emerging Technologies* 3(4), 193-209.

Appendix

TRAVEL MODE DETECTION USING BINOMIAL LOGIT MODEL*

Introduction

The study introduced in this chapter is unique in sense that it explores the possibility of applying the binomial logistic model using the sensors' data. Although, multinomial logit models have been adopted for assessing the effect of a policy change on mode shift or the possibilities of introducing a new travel mode, but using the logit model to identify the travel mode from only the data collected by sensors is a novel approach. The binomial logit model is applied in a hierarchical manner to separate six modes namely walk, bicycle, car, bus, train and subway.

Binomial Logistic Regression

Binary logistic regression is a type of generalized linear models (GLM), which models how a binary response is dependent on a set of explanatory variables. The explanatory variables can be discrete, continuous or a combination. Binary response means that there can be only two possible outcomes, either success or failure. For example, a doctor wants to figure out the proportion of breast cancer patients in a given population. Naturally, every person's risk of being a patient of breast cancer will vary, depending

* This chapter was presented as Modelling of Accelerometer data for travel mode detection by hierarchical application of binomial logistic regression, *18th Euro Working Group on Transportation, EWGT 2015*, Delft, the Netherlands. July2015.

on a number of factors including age, lifestyle and eating habits. Consider these factors or predictor variables be represented by $X = (X_1, X_2, \dots X_k)$ with observed value $x_i = (x_{i1}, x_{i2}, \dots x_{ik})$ for a person i. Let Y be the binary response variable where $Y_i = 1$ if person i is a patient and $Y_i = 0$ if otherwise. The probability (π) that the person i is a patient can be formulated as follows

$$\pi_i = \Pr\left(Y_i = 1 \middle| X_i = x_i\right) = \frac{\exp(\beta_0 + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_i x_i)} \quad (1)$$

Or

$$\begin{aligned} logit(\pi_i) &= log\left(\frac{\pi_i}{1 - \pi_i}\right) \\ &= \beta_0 + \beta_i x_i \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \end{aligned} \quad (2)$$

Methodology

Data Collection and Processing

Probe person data was collected by participants in Kobe city, Japan by employing smartphones. The accelerometer embedded in the smartphones recorded accelerations along the three axes at a frequency of around 14 Hz. The number of trips made by each of the six modes is given in Table 1.

Table 1: Number of trips for each mode

Mode	No. of trips collected	No. of trips used
Walk	512	45
Bicycle	10	10
Car	31	31
Bus	26	26
Train	44	44
Subway	16	16
Total	639	172

As it is evident from the table that the number of trips for walk is about 80% of all the trips recorded, so to form a comparable scenario, 45 trips were randomly selected for analysis. The accelerations along the three axes were used to calculate the resultant acceleration. The resultant accelerations were averaged for each trip to get one average resultant acceleration value for each trip. Similarly, for each trip, the resultant acceleration values were used to calculate standard deviation, skewness and kurtosis. A dummy variable was also introduced to input the information that either the trip was made during a weekend or a weekday.

Application of Binomial Logistic Regression

Binomial logistic regression was applied in three different manners as follows,

- Ranking
- One against rest

- One against all

In ranking, the data was first split into motorized and non-motorized modes. Then the non-motorized modes were further divided into walk and bicycle, whereas the motorized modes were divided into on road and on track. In turn, the on road modes were split into car and bus, and likewise the on track modes were split into train and subway (Figure 1).

In case of one against rest, initially the data was split into mode walk and others. Then the data excluding walk was split into bicycle and others. Similarly, each mode was separated and with each turn the data kept on decreasing until only two modes were left in the last and the same split was made between train and subway as in ranking (Figure 2). The one against all method was essentially the same but this time the data was not decreased and for each mode the entire data was taken into consideration (Figure 3).

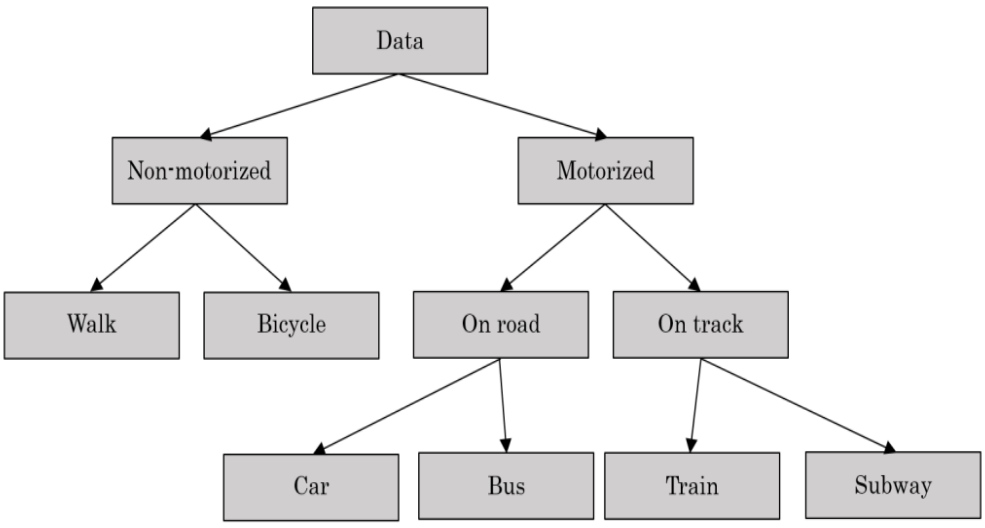


Figure 1: Ranking method of application

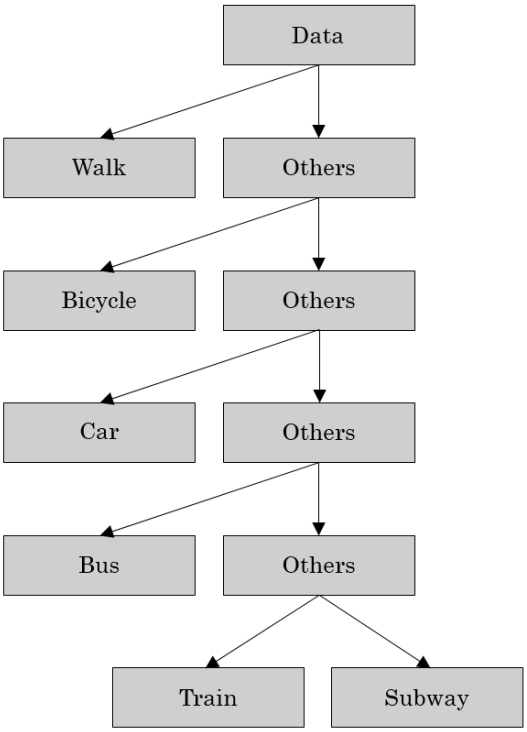


Figure 2: One against rest method of application

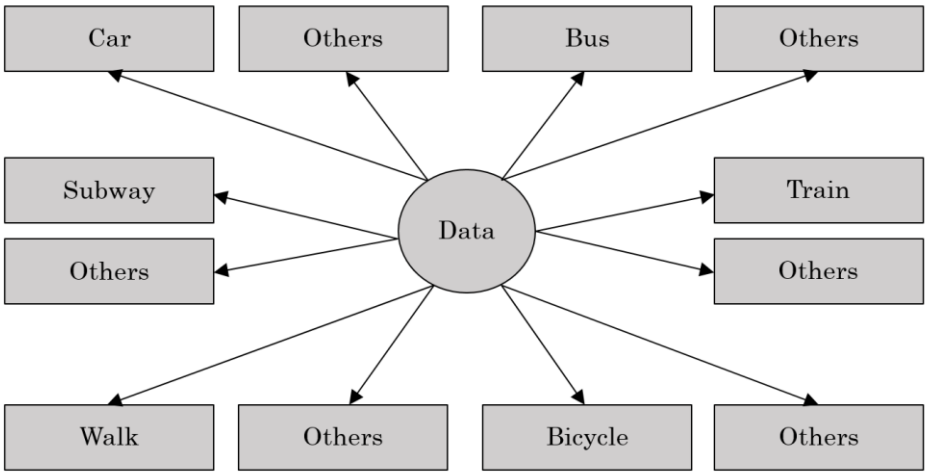


Figure 3: One against all method of application

Results and Discussion

After application of the regression model, the goodness of fit is calculated as the 1-pchisq using the residual deviance and corresponding degree of freedom. The results for each method are provided in Tables 2 to 4.

Table 2: Regression results for ranking method

Split	Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Goodness of fit
Motorized vs. non-motorized	Intercept	2.657	4.633	0.574	0.566	0.238
	Average Resultant Acceleration	-0.022	0.484	-0.044	0.965	
	Standard Deviation	-1.176	0.289	-4.068	0.000	
	Skewness	0.333	0.192	1.732	0.083	
	Kurtosis	-0.003	0.006	-0.508	0.612	
	Dummy (weekend)	-0.380	0.392	-0.971	0.332	
Walk vs. Bicycle	Intercept	-54.440	22.876	-2.380	0.017	0.975
	Average Resultant Acceleration	5.433	2.280	2.383	0.017	
	Standard Deviation	2.057	1.103	1.865	0.062	
	Skewness	0.262	0.525	0.499	0.618	
	Kurtosis	0.003	0.018	0.147	0.883	
	Dummy (weekend)	-0.014	0.981	-0.014	0.989	
On road vs. on track	Intercept	10.686	7.212	1.482	0.138	0.104
	Average Resultant Acceleration	-1.270	0.757	-1.678	0.093	
	Standard Deviation	1.581	0.472	3.348	0.001	
	Skewness	-0.013	0.014	-0.963	0.336	
	Kurtosis	-0.186	0.199	-0.935	0.350	
	Dummy (weekend)	1.508	0.491	3.070	0.002	
Car vs. Bus	Intercept	-40.506	16.581	-2.443	0.015	0.667
	Average Resultant Acceleration	4.155	1.717	2.420	0.016	
	Standard Deviation	-1.192	0.720	-1.656	0.098	
	Skewness	1.073	0.378	2.840	0.005	
	Kurtosis	-0.031	0.018	-1.729	0.084	
	Dummy (weekend)	3.592	1.309	2.744	0.006	
Train vs. Subway	Intercept	11.087	9.811	1.130	0.258	0.168
	Average Resultant Acceleration	-1.377	1.032	-1.335	0.182	
	Standard Deviation	0.926	0.807	1.147	0.252	
	Skewness	-0.075	0.411	-0.182	0.855	
	Kurtosis	0.038	0.029	1.317	0.188	
	Dummy (weekend)	-1.029	0.689	-1.494	0.135	

Table 3: Regression results for one against rest method

Split	Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Goodness of fit
Walk vs. Others	Intercept	12.710	5.895	2.156	0.031	0.960
	Average Resultant Acceleration	-1.108	0.605	-1.833	0.067	
	Standard Deviation	-0.914	0.407	-2.242	0.025	
	Skewness	0.347	0.325	1.069	0.285	
	Kurtosis	0.069	0.044	1.572	0.116	
	Dummy (weekend)	-0.450	0.446	-1.007	0.314	
Bicycle vs. Others	Intercept	-34.127	12.270	-2.781	0.005	0.986
	Average Resultant Acceleration	3.733	1.278	2.922	0.003	
	Standard Deviation	-0.200	0.506	-0.395	0.693	
	Skewness	0.428	0.252	1.697	0.090	
	Kurtosis	-0.002	0.006	-0.362	0.717	
	Dummy (weekend)	-0.705	0.529	-1.333	0.183	
Car vs. Others	Intercept	23.455	7.993	2.934	0.003	0.558
	Average Resultant Acceleration	-2.393	0.833	-2.872	0.004	
	Standard Deviation	1.744	0.566	3.080	0.002	
	Skewness	-0.802	0.237	-3.389	0.001	
	Kurtosis	0.014	0.013	1.093	0.274	
	Dummy (weekend)	-0.309	0.499	-0.619	0.536	
Bus vs. Others	Intercept	2.163	10.872	0.199	0.842	0.776
	Average Resultant Acceleration	-0.346	1.139	-0.304	0.761	
	Standard Deviation	1.264	0.607	2.083	0.037	
	Skewness	0.605	0.363	1.665	0.096	
	Kurtosis	-0.028	0.019	-1.453	0.146	
	Dummy (weekend)	3.826	1.248	3.065	0.002	
Train vs. Subway	Intercept	11.087	9.811	1.130	0.258	0.168
	Average Resultant Acceleration	-1.377	1.032	-1.335	0.182	
	Standard Deviation	0.926	0.807	1.147	0.252	
	Skewness	-0.075	0.411	-0.182	0.855	
	Kurtosis	0.038	0.029	1.317	0.188	
	Dummy (weekend)	-1.029	0.689	-1.494	0.135	

Table 4: Regression results for one against all method

Split	Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Goodness of fit
Walk vs. Others	Intercept	11.662	6.126	1.904	0.057	0.822
	Average Resultant Acceleration	-0.940	0.629	-1.494	0.135	
	Standard Deviation	-1.115	0.303	-3.674	0.000	
	Skewness	0.424	0.216	1.964	0.049	
	Kurtosis	0.003	0.012	0.240	0.811	
	Dummy (weekend)	0.284	0.449	0.633	0.527	
Bicycle vs. Others	Intercept	-25.708	12.299	-2.090	0.037	1.000
	Average Resultant Acceleration	2.936	1.287	2.281	0.023	
	Standard Deviation	0.170	0.550	0.308	0.758	
	Skewness	0.209	0.265	0.788	0.431	
	Kurtosis	0.001	0.009	0.072	0.943	
	Dummy (weekend)	-0.901	0.697	-1.293	0.196	
Car vs. Others	Intercept	19.443	6.905	2.816	0.005	0.991
	Average Resultant Acceleration	-1.963	0.720	-2.726	0.006	
	Standard Deviation	1.918	0.519	3.696	0.000	
	Skewness	-0.811	0.226	-3.584	0.000	
	Kurtosis	0.015	0.013	1.116	0.264	
	Dummy (weekend)	-0.357	0.471	-0.759	0.448	
Bus vs. Others	Intercept	-8.279	8.218	-1.007	0.314	1.000
	Average Resultant Acceleration	0.878	0.853	1.030	0.303	
	Standard Deviation	0.810	0.431	1.881	0.060	
	Skewness	0.303	0.218	1.386	0.166	
	Kurtosis	-0.017	0.007	-2.270	0.023	
	Dummy (weekend)	2.840	0.902	3.150	0.002	
Train vs. Others	Intercept	-0.359	5.172	-0.069	0.945	0.166
	Average Resultant Acceleration	0.107	0.534	0.200	0.842	
	Standard Deviation	0.295	0.284	1.038	0.299	
	Skewness	-0.155	0.209	-0.742	0.458	
	Kurtosis	0.031	0.018	1.751	0.080	
	Dummy (weekend)	-1.013	0.378	-2.678	0.007	

	Intercept	-15.610	9.016	-1.731	0.083	
	Average Resultant Acceleration	1.857	0.949	1.957	0.050	
Subway vs.	Standard Deviation	-0.322	0.400	-0.806	0.420	1.000
Others	Skewness	-0.117	0.237	-0.494	0.621	
	Kurtosis	0.005	0.016	0.311	0.756	
	Dummy (weekend)	0.385	0.586	0.658	0.511	

The results suggest that for every method adopted, problem arises when train is involved in the split. In ranking, the goodness of fit is less than 0.5 at 3 levels, motorized vs. non-motorized, on road vs. on track and train vs. subway. These three levels are dealing with train. Similarly, for one against rest and for one against all, the goodness of fit is less than 0.5 for train vs. subway and train vs. others respectively. This shows that the three methods are very much applicable for all modes except train. The reason might be hidden within the raw data. Figure 4 summarizes the resultant acceleration values collected for each mode. It is evident that the mode train is a bit difficult to model because its range of acceleration values cover all other modes.

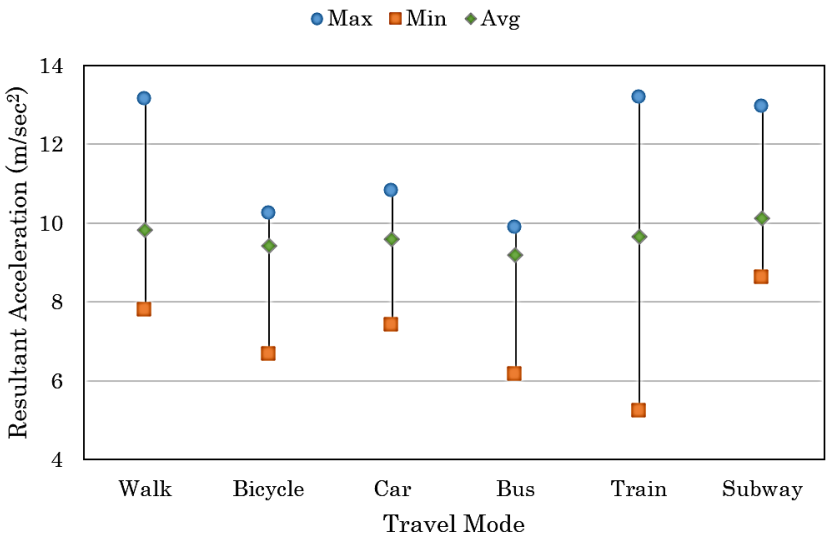


Figure 4: Summary of Resultant Accelerations for each mode

This chapter explores the possibility of a new problem solving technique for transportation mode classification. Binomial logistic regression can be successfully utilized to model the travel modes by using only the acceleration values and hence, can classify the data into various modes. The modeling of mode train is a bit problematic but the effect can be minimized by using the one against rest or one against all method of application, instead of ranking method. The one against all method provides best results where, apart from train, all the other modes show a goodness of fit value close to 1. This method should be improved further and ultimately it can complement the current methods used for travel mode identification.