Activity-Travel Pattern Analysis Based on Mobile Phone GPS Data

Lei GONG

Doctoral Dissertation

Activity-Travel Pattern Analysis Based on Mobile Phone GPS Data

by

Lei GONG

Submitted in Partial Fulfillment of the
Requirement for the Degree of
Doctor of Engineering

December 2015

Department of Civil Engineering
Nagoya University
Japan

Abstract

There is consensus that traditional person trip survey (PT survey) cannot avoid the disadvantages such as underreported trips, inaccuracies in times, surrogate reporting and sometimes confusion of appropriate trip purpose. The occurrence of Global Positioning System (GPS) technology, especially the popularity of smart phone with GPS censor, makes it possible to collect the PT data to overcome the disadvantages of traditional PT survey. However, the raw GPS data is a series of temporal trajectory points with coordinates (at least longitude and latitude). So efficient techniques and methodologies are needed to extract the PT data with a high accuracy. The objective of this thesis is to propose a series of methodologies to extract PT data, such as activity type (same meaning as "trip purpose" in this thesis), from continuous GPS trajectories and analyze the activity-travel pattern based on the extracted PT data.

The first step of extracting PT data from continuous GPS trajectories is to segment GPS points into separate trip and the activity engaged in at the trip end. A density-based algorithm is developed to distinguish the GPS points into moving points and stop points. It is an improved version of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. Two constraints from temporal and spatial are added as improvements to the original algorithm. Then a machine learning method, Support Vector Machine (SVM) is applied to distinguish the activity stop from all the identified stops.

After obtaining the distinguished activity stops, the next step is to identify the specific type of activity. Although some heuristic-rule-method has been applied to this field before, it is machine learning methods that avoid resetting the rules manually when data set changes, because machine learning methods can finish the task by automatically learning the inherent relationship between dependent variable and independent variables with efficiency. In the field of identifying activity types, it is still unknown which machine learning method has the best performance. That is why several machine learning methods are tested and compared in this research with the same dataset. It is found that classification trees outperforms SVM, Neural Networks (NN), and Discriminant Analysis (DA) from the perspective of time cost and accuracy.

Improving the accuracy of activity type identification are very necessary for the follow-up research, such as activity-travel pattern analysis and activity-based models. Due to this concern, several techniques of improving the accuracy of activity type identification are discussed. These techniques include data selection for training set and test set from seasonal data, and inclusion of effective additional variables.

One advantage of PT data obtained from GPS trajectories is gathering the multi-day information. Due to this, analyzing what kind of factors influence the activity-travel pattern during a long period become possible. Ordered logit models are used to estimate the impact of variables from time dimension, trip dimension and weather on trips and trip chains from a prospective of temporal heterogeneity. The number of trips in a trip chain, the number of trip chains in a day, and the number of trips are used as dependent variables individually in the analysis.

Since the accuracy of activity type identification will influence the analysis in the next step analysis, such as activity-travel pattern analysis and activity-based models. Until the activity type identification reaches 100%, it is necessary to consider the activity type identification accuracy in the follow-up study. An activity sequence generation model with consideration of accuracy of activity type identification is developed. The model can be decomposed as a series of discrete choice model in sequence and the activity identification accuracy can be converted as part of the constant value in the function. The results show the prediction of activity frequency forecasted by the activities identified with error and forecasted by the real activities are not significantly different from each other.

Thesis supervisor: Dr. Toshiyuki YAMAMOTO
Title: Professor of Institute of Materials and Systems for Sustainability, Nagoya University

Acknowledgement

I would like to express my deep gratitude to all the people who contributed to my thesis.

It is my great honor to be a doctoral student under the supervision of Professor Toshiyuki Yamamoto. During the doctoral period, I have learned so much valuable things that I have not learned before from Professor Yamamoto. As my academic advisor and thesis supervisor, he has provided me with extremely valuable suggestions and perceptive guidance. Professor Takayuki Morikawa, as my vice supervisor and a member of the committee, gave me great help in choosing current topic and constructive advice in the research progress as well as the thesis completion. Professor Kato and Professor Suzuki, as the committee members, offered me valuable comments in the process of midterm oral examination and the process of the thesis competition.

I would like to thank Professor Miwa and Lecturer Sato, for their kindness and precious comments on this research.

I also want to show my gratitude to Professor Ryo Kanamori, who provided me with the Hakodate data and great ideas of how to analyze data for the classification using random forests.

Dr. Cao Peng offered his map-matching algorithm to help me finish the job of matching the GPS trajectory points to the road network in Hakodate. Ms. Sun Xiaohui shared her knowledge of using R to do the estimation. Dr. Sugiarto provided his knowledge and techniques of modeling discrete choices in STATA to me. Mr. Araki in the master course helped me do the ground truth checking in the data set of Hakodate. Here, I also want to express my thanks to them for their dedicated work.

Ms. Kikata, Ms. Tsuda, and Ms. Kuroda also offered their great help in the preparation of my presentations in the conference and I want to thank to them.

And this research cannot be finished without the scholarship provided by the Ministry of Education, Culture, Sports, Science & Technology in Japan. It is not just a chance of finishing a thesis or earning a degree; it is opportunity to experience a wonderful life I longed for from my childhood.

Last, but never least, I would like to express my gratitude to my family, for their understanding and support.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. BACKGROUND

Person trip survey (PT survey) is a very important survey to obtain the person trip data for understanding the characteristics of current traffic demand and predict the future traffic demand in a city or a region or even a larger area. Person trip data usually contain the detailed information of each daily trip (like trip starting time, end time, trip purpose and travel mode, origin and destination, companionship), demographic information of the household and each family member.

With the technical progress such as telephone, computer etc. the PT survey has also experienced several stages, including paper-and-pencil-interview, computer-assisted telephone interview, and computer-assisted-self-interview (Wolf et al., 2001). These three stages can be treated as traditional PT survey, since they rely on the active reports from the survey participants. However, it is found that people's intention of attending the survey drops quickly in recent years. Besides, the active reports cannot avoid the disadvantages such as underreported trips, inaccuracies in times, surrogate reporting and sometimes confusion of appropriate trip purpose (McGowen and McNally, 2007). It is an urgent need to find some alternative survey method to ease the burden of participants' memorizing the detailed information of each trip in the daily travels.

The application of Global Positioning System (GPS) in the civil field not only provide the possibility of navigation, it is also a way of tracing the GPS holders. At first, the GPS devices were installed in the vehicle and electrified by the battery of the vehicle. At that time, it was only possible to trace the vehicle or the person when driving the vehicle. Later on, the GPS devices became smaller and finally can be attached into the cellphone as a censor. Nowadays, most of the smartphones have attached GPS censors, and it is possible to trace the persons in any travel mode going to anywhere in case the cellphone can receive the GPS signal. The tracing can collect temporal and spatial information of the cellphone holders and these information can be used to extract the personal trip data. Furthermore, since there is no need for the cellphone holders to remember and recall the detailed information of whole trips s/he does during the day, it makes possible to ease the burden of participants in the PT survey. The only thing participants need to do is to carry the cellphone which is a very common behavior for most of the people in the current society.

The collected information by GPS devices/censors usually include basic information, such as time stamp and coordinates (longitude, latitude& altitude), and additional information such as speed, acceleration, GPS signal quality, heading, NSAT (the number of satellites that a GPS device used to calculate its position), HDOP (horizontal dilution of precision, measuring how the satellites are arranged in the sky at the time of the record), and so on. The basic information is available to any type of GPS device/censor; additional information depends on the function of the GPS devices/censors.

Collected GPS information can be shown as a series of points with the features mentioned above, or a continuous trajectory on the map. Without successfully segmenting these trajectories into trips & activities at trip ends, and distinguish the trip purpose & trip mode of each trip, the collected GPS information cannot be used as PT data in travel demand analysis and travel behavior analysis.

As the passively collecting GPS data can last for several days, several weeks or even several months, the travel demand analysis and travel behavior analysis during a longer period can be done. And it can make travel demand analysis more precise and analyze how behavior changes according to the influence from other prospect, like weather.

1.2. PROBLEM STATEMENT

However, the GPS data cannot be used as PT data without any further processing, because these data are just a series of spatial points with time stamp and some other secondary information. From the continuous GPS trajectories to the PT data, there are several problems need to be solved.

The first problem is to segment trips and activities. Usually before this problem, there should be one more task of checking GPS error by GPS features (such as signal quality, speed, NSAT, HDOP, etc.) or other algorithms. But it is not the focus in this research. Here starts from segmentation of points. After showing the continuous GPS trajectories on the map (like in Figure 1.1) together with GIS (Geographic Information System) information, it is found that some GPS points gather together at some locations while some GPS points scatter along the road links or railway links. Referencing to the GIS information, it is found that the gathered GPS points are either near some POIs (Point of Interest, such as supermarket, shopping mall, hospital, ward office, police station, post office, school, sports stadium, etc.) or near intersections or bus stops, railway stations. So some of these gathered GPS points are demonstrating activities. Finding an appropriate algorithm to distinguish these GPS points can be a possible way to segment the trips and activities.

After distinguishing the GPS points which represent activities and which represent trips, it is still unknown what kinds of activities at trip ends they are and what types of travel modes used during the trips. So the second problem is to detect the activity type (or “trip purpose”, these two expressions mean the same thing in this thesis) and travel mode of each segmented trip. Some research already developed heuristic rules to do this task. However, these heuristic rules are usually proposed based on the features of local data. If the data set changes from current city to another city, the heuristic rules have to be changed due to the feature variation. So it is necessary to develop a methodology that can automatically understand the inherent features of the data set and can apply its understanding to the other part of the data set with a satisfactory accuracy efficiently. Machine learning is believed to be a suitable method to handle this kind of problem and several machine learning methods have already been applied on identifying travel mode. However, machine learning's application on activity type identification is rare and it still be unknown which type of machine learning method is the most suitable for identifying activity type, since there are several types of machine learning methods with advantages and disadvantages. And there is still no conclusion about the superiority based on comparisons of machine learning methods' application on activity type identification.

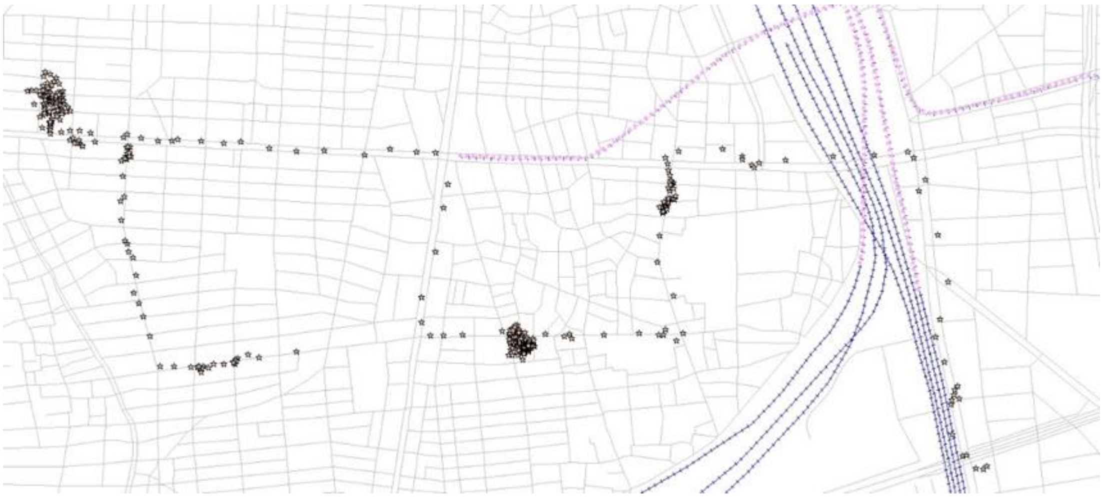


Figure 1.1 A series of trips in a trajectory of one person (Nagoya, 2008)

Different from traditional PT survey, GPS technology, as a passive way of collecting data, makes collecting PT data for a longer period possible. Due to this, travel behavior analysis, activity-travel pattern analysis, and activity-based model can be done using a longer period of data and it can make contributions to achieve more accurate analysis and models. What is more, fluctuation of some factors can only be observed during a longer period, otherwise it cannot be included in the analysis, such as weather, residence location change, household structure change, auto/bicycle/motorcycle possession change, and transportation policy etc.

1.3. OBJECTIVES

Based on the background and problems above, this thesis is focusing on detecting activity type in the GPS data and analyzing activity-travel pattern using continuous data. To be specific, the objectives of this thesis are listed as follows.

First, to develop an algorithm used for segmenting the trips and activities from continuous GPS trajectories with only the basic information of GPS (time stamp, longitude and latitude).

Second, to compare different machine learning methods and to find a suitable one for identifying activity types from the perspective of time cost and accuracy.

Third, to analyze the reasons of low accuracy in the process of machine learning method and to advance techniques to improve the accuracy.

Fourth, to analyze activity-travel pattern from a longer period with GPS data focusing on weather's influence.

Fifth, to develop an activity sequence generation model considering the accuracy in the process of identifying activity type identification from GPS trajectories.

1.4. STRUCTURE

The whole structure of this thesis can be found in Figure 1.2.

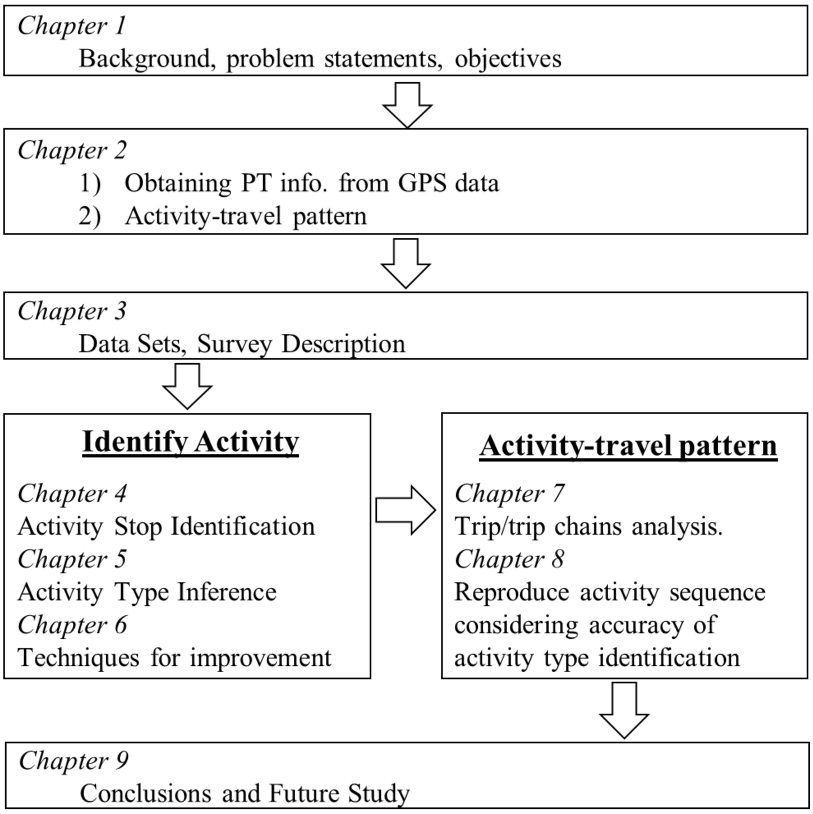

Chapter 1
Background, problem statements, objectives
Chapter 2
1) Obtaining PT info. from GPS data
2) Activity-travel pattern
Chapter 3
Data Sets, Survey Description
Identify Activity
Chapter 4
Activity Stop Identification
Chapter 5
Activity Type Inference
Chapter 6
Techniques for improvement
Activity-travel pattern
Chapter 7
Trip/trip chains analysis.
Chapter 8
Reproduce activity sequence considering accuracy of activity type identification
Chapter 9
Conclusions and Future Study

Figure 1.2 Structure of this thesis

This thesis contains 9 chapters. Chapter 2 reviews and summarizes the techniques used for obtaining PT data from GPS data and results of trip chain analysis in the prior research. Description of the data sets used in this research is shown in chapter 3. It is followed by the methodologies of activity location identification from continuous GPS trajectories in chapter 4. In this chapter, first, an improved density-based algorithm is used to identify all the stops in the trajectory, including non-activity stop and activity stop. Then a support vector machines (SVMs) for binary classification is applied to distinguish activity stop from non-activity stop. Chapter 5 tests and compares several machine learning methods to identify activity types from the perspective of time cost and accuracy. These machine learning methods include classification tree, SVMs, discriminant analysis and neural network. Then several techniques of improving the identification accuracy of activity type are advanced and tested in chapter 6. These techniques include data selection for training set and test set among seasons, and effect from additional variable dimensions. In chapter 7, a daily activity-travel pattern analysis is done using GPS data in a duration of 8 months focusing on the weather's influence. Analyzed activity-travel pattern include the number of trip chains in a day, the number of trips in a trip chain, and the number of trips. It is followed by an activity sequence generation model developed considering the identification accuracy of activity type from GPS data. Finally, conclusions and future research are drawn in chapter 9.

1.5. REFERENCE

McGowen, P., and McNally, M. (2007). Evaluating the potential to predict activity types from GPS and GIS data. In *Transportation Research Board 86th Annual Meeting*, Washington D.C.

Wolf, J., Guensler, R., andBachman, W. (2001). Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data. In *Transportation Research Board 86th Annual Meeting*, Washington D.C.

Chapter 2. Literature Review and Contribution of This Research

2.1. OBTAINING PERSON TRIP DATA FROM GPS DATA

GPS technology was used in person trip (PT) survey since mid-1990, and this technology achieved its popularity because of the improvement of accuracy and portability of GPS device. Although GPS data could provide precise spatiotemporal information of vehicular or personal movements, the travel mode (in the case of personal movements with wearable GPS devices) and activity type are unable to be obtained from the GPS directly. In addition, the GPS data error identification and the trip segment from the continuous GPS data are the prerequisites to travel mode identification and activity type inference. In this part, methodologies and input variables utilized to identify error, segment trip, infer activity type as well as identify travel mode in the existing researches are summarized.

2.1.1. Introduction

Household trip data are crucially infrastructural for traffic demand analysis in transportation system planning. The methods used for personal trip data collection experienced the stages of original paper-and-pencil interview (PAPI), computer-assisted telephone interview (CATI), and computer-assisted-self-interview (CASI) (Wolf et al., 2001). Although the computer assisted surveys tried to help the participants understand the questions and recall the trips they had during the day (Hato et al., 2006), the involvement of computer technology still could not solve the inherent disadvantages of the method of actively collecting PT data. These disadvantages include underreported trips, inaccuracies in times, surrogate reporting and sometimes confusion of appropriate trip purpose (McGowen and McNally, 2007).

Because GPS data is capable of providing accurate data including location, time, speed, heading and the measures of data quality (Stopher et al., 2008b), in the middle of 1990s, researchers started to investigate the possibility of obtaining the trip data from the GPS data and test the accuracy of the GPS data (Zitto et al., 1996; Wagner, 1997; Murakami et al., 1999; Sermons and Koppleman, 1998). At first, the GPS devices were installed in the vehicle and electrified by the battery on the vehicle. So it is applicable only to observe the travel behavior of persons when driving vehicles. This problem got solved in the early 2000s when the size and weight of GPS devices getting smaller and lighter with a detached battery (Stopher et al., 2008a). The smaller and lighter GPS devices, namely wearable GPS data logger (2nd generation, 3rd generation), appeared in the pilot study or personal trip survey in the UK (GeoLogger was used in 2002), and Australia (NEVE StepLogger was used in 2003; Starnav was used in 2005) (Stopher et al., 2008a). These wearable GPS devices still have the demerits such as respondents forgetting to bring the devices, GPS signal unavailability in the building, underground, in the tunnel or "urban canyons" areas.

Since GPS sensor was attached to smart phone, some researches also started to use GPS data obtained from smart phones to derive personal trip data. These researches either combined a web-based diary system or Geographic Information System (GIS) to get the confirmed information of travel modes and activity type passively or actively (Hato et al., 2006; Itsubo et al., 2006; Byon et al., 2007; Reddy et al., 2008; Zheng et al., 2008; Gonzalez et al., 2010; Zhang et al., 2011; Lee et al., 2013; Pereira et al., 2013). Furthermore, the assisted GPS system, named AGPS, are getting widely contained in the smartphones, e.g. iPhone. This technology can receive satisfied GPS signal inside buildings, vehicles, as well as "urban canyons" in cities where tall buildings and other edifices block GPS signals. This system has improved and owned higher receiver sensitivity (Moiseeva and Timmermans, 2010). It means that GPS data with higher accuracy can be collected with smart mobile phones. If the technologies of automatically deriving personal trip data can also be achieved with satisfactory accurate results, GPS data collection through smart phone may become the primary method of collecting personal trip data in the future at lower cost and with minimum burden on participants.

The collected features of GPS data may vary depending on the types of GPS devices. They generally include: valid code marking, date, time, latitude, longitude, altitude, NSAT (the number of satellites that a GPS device used to calculate its position), HDOP (horizontal dilution of precision, measuring how the satellites are arranged in the sky at the time of the record), speed, and heading (Wolf et al., 2001; Stopher et al., 2005, 2008a; Gong et al., 2012). Although the path, time, speed, and acceleration could be obtained precisely from the GPS raw data (assuming the GPS data with high accuracy), start & end of trip, travel mode, and activity type cannot be derived from the GPS raw data directly without further data processing or other assisted information.

In this section, the summarization is done in the perspective of data error recognition, trip segmentation, travel mode identification and activity type identification.

2.1.2. Data error recognition

Although the GPS data can avoid manual mistake of PT data, e.g. inaccurate time and under-reporting trips, some systematic errors may exist in the GPS raw data. As a result, the GPS raw data need to be examined to ensure the accurate data used in the next steps. Table 2.1 shows the features of GPS data used for error recognition in the major existing researches. Detailed information of these methods is interpreted as follows.

Table 2.1 Summary of data error recognition in existing researches

Year	Authors	GPS Devices	Records' Features Used for Error Recognition
2005 2008a	Stopher et al.	Wearable GPS devices including GeoLogger, StepLogger and Starnav	NSAT, HDOP value, speed
2006	Tsui and Shalaby	Wearable GPS devices	NSAT, HDOP value, speed, heading, path
2009	Bohte et al.	---	Duration, speed, number of trackpoint per trip

Stopher et al. (2005, 2008a) utilized 2 rules to remove the invalid data. The first rule is data with less than 4 satellites (for 3-D use) or HDOP of 5 or more were removed. The second one is the record with a speed over 250 km/h were removed.

Tsui and Shalaby (2006) used Data Filtering Module to identify the systematic errors. In this module, 4 successive filters are utilized to ensure the records correct. The first and second filter use the factors of NSAT and HDOP respectively. The records with NSAT fewer than 3 (for 2-D use) and HDOP higher than 5 are not considered in the next step. The third filter treats records with 0 directional heading and 0 speed as errors when GPS data trace is plotted on a map. The last filter is to remove multipath error in "urban canyons" areas, causing GPS signal to jump around the area and form a data cloud instead of clear traces.

Bohte et al. (2009) set up the following rules to delete unreliable records of GPS raw data. The first one is that trackpoints whose distance to previous one was less than 10 m if these trackpoints collected in the same building. The second one is to remove the trackpoints with a speed higher than 200 km/h. The next rule deletes the trackpoints with a speed less than 5 km/h and a time gap with previous trackpoint of at least 1 minute. The final one is to delete the trips with less than 4 trackpoints.

2.1.3. Trip segmentation

The GPS trajectories contain the trips and activities at trip ends. In existing research, trip segmentation, activity identification, stop and move identification usually share the same meaning. Some research just simply use one or several features of GPS data to segment the trips while others develop more complicated criteria based on features of GPS data and additional information, like GIS data to do the job. Here, detailed review can be found according to these two ways.

2.1.3.1. Simply judging from features in GPS data

This method was usually used for GPS devices installed on vehicle in the early usage stage of GPS. At that time, GPS device shares the power of vehicle and starts/goes off simultaneously with engine. So it usually means a trip end when the engine stops for certain time. As a result, most of the researches use a certain dwell time with or without other conditions to detect activities. Table 2.2 shows a summary of the features used for activities detection in the existing researches under two situations: GPS signal available situation and GPS signal lost situation.

Table 2.2 Summary of trip end identification methods in existing researches

Year	Authors	Signal Available						Signal loss
		S (m/s)	D (sec)	L / L(°)	H	PD	Other	D (sec)
2001	Wolf et al.	0	≥120	---	---	---	---	---
2004	Axhausen et al.	---	≥300	---	---	---	---	---
2006	Tsui and Shalaby	0	≥120	---	---	---	---	≥120
2002 2005 2008ab	Stopher et al.	0	≥120	≤0.00005	UC or 0	---	VC	---
2009	Bohte and Maat	---	≥180	---	---	---	---	---
2009	Schuessler and Axhausen	≤ 0.01	≥120	---	---	≥15	---	≥900
2012	Gong et al.	---	≥200	---	---	---	In 50m	---

Abbr: S: Speed. D: Duration. L/L: Change in Latitude or Longitude. H: Heading. PD: Point Density. VC: Visual Check on Map. UC: Unchanged.
Note: Requirement in each cell in the same line should be satisfied simultaneously in the situation of GPS signal is available (except for Schuessler and Axhausen 2009 where the judging criteria is when either S & D or PD get satisfied).

Under the situation of GPS signal is available, different threshold of dwell time is set, such as 120s (Wolf et al., 2001; Tsui and Shalaby, 2006; Stopher et al., 2002, 2005, 2008ab; Schuessler and Axhausen 2009), 180s (Bohte and Matt, 2009), 200s (Gong et al., 2012) or even 300s (Axhausen et al. 2004). Actually, this threshold varies mainly depending on the local characteristics of activities. In addition, some researches simultaneously included the "0" speed or approximately "0" as another necessary condition (Wolf et al., 2001; Tsui and Shalaby, 2006; Stopher et al., 2002, 2005, 2008ab; Schuessler and Axhausen, 2009). Furthermore, the change in latitude or longitude, the change of heading, and the density of track points of GPS data were also treated as necessary conditions in some researches (Stopher et al., 2002, 2005, 2008ab; Schuessler and Axhausen, 2009). Besides the features mentioned above, visual checking on map (Stopher et al., 2002, 2005, 2008ab) and the required boundary in which trackpoints satisfied dwell time threshold (Gong et al., 2012) are also included as the detecting rules in some research.

For the situation when there is no available GPS signal, the dwell time between two successive trackpoints were utilized as a judging criterion to detect potential trip in some research (Tsui and Shalaby, 2006; Schuessler and Axhausen, 2009).

2.1.3.2. Deriving criteria based on features in GPS and other additional information

In this category, some research (Agamennoni et al. 2009; Alvares et al. 2007; Ashbrook and Starner 2003; Kami et al. 2010; Leclerc et al. 2013; Tran et al. 2011; Xie et al. 2009; Zhou et al. 2007) have attempted to identify stops for activity in a single step, while others (Andrienko et al. 2013; Mizuno et al. 2013; Palma et al. 2008; Yan et al. 2010; Zimmermann et al. 2009) have used a two-step procedure: identifying all stops in the first step; and refining the identification to isolate activity stops in the second step.

These methods can be further categorized into the following five groups: centroid-based methods, speed-based methods, duration-based methods, density-based methods and hybrid methods. A brief review of these is given below.

A **centroid-based method**, specifically, a variant of the k-means clustering algorithm has been applied (Ashbrook and Starner, 2003) to obtain the locations that are significant for the subject. The points are divided into k clusters by iteratively calculating the mean of points (or centroid of points) as the new temporary center point within a given radius of the current temporal center point until the center point converges. However, the number of stops, k, has to be known beforehand. It is nearly impossible to know how many stops there are in a trajectory.

Speed-based methods. Agamennoni et al. (2009) defined a scoring function involving speeds to reflect the significance of a vehicle's current location. The scoring function defined the significance of the current location by comparing current speed with two thresholds of speed in a mining environment. Mizuno et al. (2013) used speed and change rate of average speed as input features for SVMs to obtain moving and stopping points. Nevertheless, speed-based methods need to know speed, which is not always available to all GPS devices or modules. Besides, some limitations arise in situations such as the subject moving in parking lot or stuck in traffic or in bad weather conditions.

Duration-based methods are the most popular method of identifying stop location. Palma et al. (2008) and Tran et al. (2011) applied a modified DBSCAN algorithm based on a minimum stop duration instead of a minimum number of points in a neighborhood when defining core points. The difference in the two papers is that distance along the trajectory was used for distance calculations by Palma et al. (2008) whereas the straight-line distance between two points was used for distance calculations by Tran et al. (2011). Alvares et al. (2007) and Xie et al. (2009) identified stops by judging stop duration and whether the GPS point intersect with the geometry of a spatial location. The difference is that Alvares et al. (2007) used a given threshold stop duration to map the trajectory to possible activities whereas Xie et al. (2009) utilized a matching table containing minimum and maximum elapsed times for each possible type of activity. One problem of duration-based methods is how to decide the optimal duration threshold because the result is very sensitive to the setting of this threshold.

Density-based methods. Kami et al. (2010) proposed a fast algorithm for probabilistically extracting significant locations from raw GPS data based on data point density. This algorithm eases the difficulty in parameter setting and works well even if there are a variety of noise levels in input data. Zimmermann et al. (2009) utilized an interactive density-based clustering algorithm, in which the density was defined on the basis of both the spatial and the temporal properties of a trajectory. Zhou et al. (2007) used a simplified mechanism of expanding clusters in DBSCAN. According to the simplified mechanism, any two clusters with shared points can be joined together as one cluster. Density-based methods require data to be collected at more frequent intervals. Moreover, since density-based methods use the concept of spatial point clustering, adjustments are needed when applied to GPS trajectory situations, which are different from those with spatial points with no direction of movement and time stamps.

Hybrid methods use two of the variables such as speed, duration, density, etc. together. Andrienko et al. (2013) extracted stops with a user-specified minimum duration and a diagonal spacing that is less than a user-specified distance threshold. Leclerc et al. (2013) also used duration but with an additional distance criteria for judging whether points are in a stop location or not. Yan et al. (2010) used a speed threshold and minimal stop duration to distinguish trajectories into stop episodes and move episodes. The speed threshold is dependent on the moving object and the location of the moving object. Hybrid methods might improve accuracy to some extent, but it is hard to completely avoid the demerits of the other methods mentioned above.

2.1.4. Travel mode detection

Figure 2.1 demonstrates the main methodologies with the input variables for travel mode detection applied in the existing researches. Methodologies from three categories are mainly utilized with the GPS data and other assisted data.

Many researches focus on the machine learning technology, which can learn inherited mapping relationship between target values and input variables from training data set and use the learned knowledge to automatically deal with other data set which share the same characteristics as the former. A lot of methods in this category, including Multi-Layer Perceptron Neural Network (Byon et al., 2007; Gonzalez et al., 2010), Decision Tree (Patterson et al., 2003; Zheng et al., 2008; Reddy et al., 2008), Bayesian Network (Zheng et al., 2008; Moiseeva and Timmermans, 2010), Support Vector Machine (Zheng et al., 2008; Zhang et al., 2011; Pereira et al., 2013), and Conditional Random Field (Zheng et al., 2008), have been applied in detecting travel modes. Most of the input variables come from the GPS data itself. In addition, some researchers compared several types of machine learning methods (Patterson et al., 2003; Zheng et al., 2008) and got the most accurate method based on their data set and local personal trip characteristics.

Another category for travel mode detection belongs to the Probability method. Fuzzy logic rules (Tsui and Shalaby, 2006; Schuessler et al., 2009) and probability matrix (Stopher et al., 2008a) are utilized to predict the probability of each mode based on the features of GPS data and respondent information. The mode with the largest probability will be decided as the estimated mode.

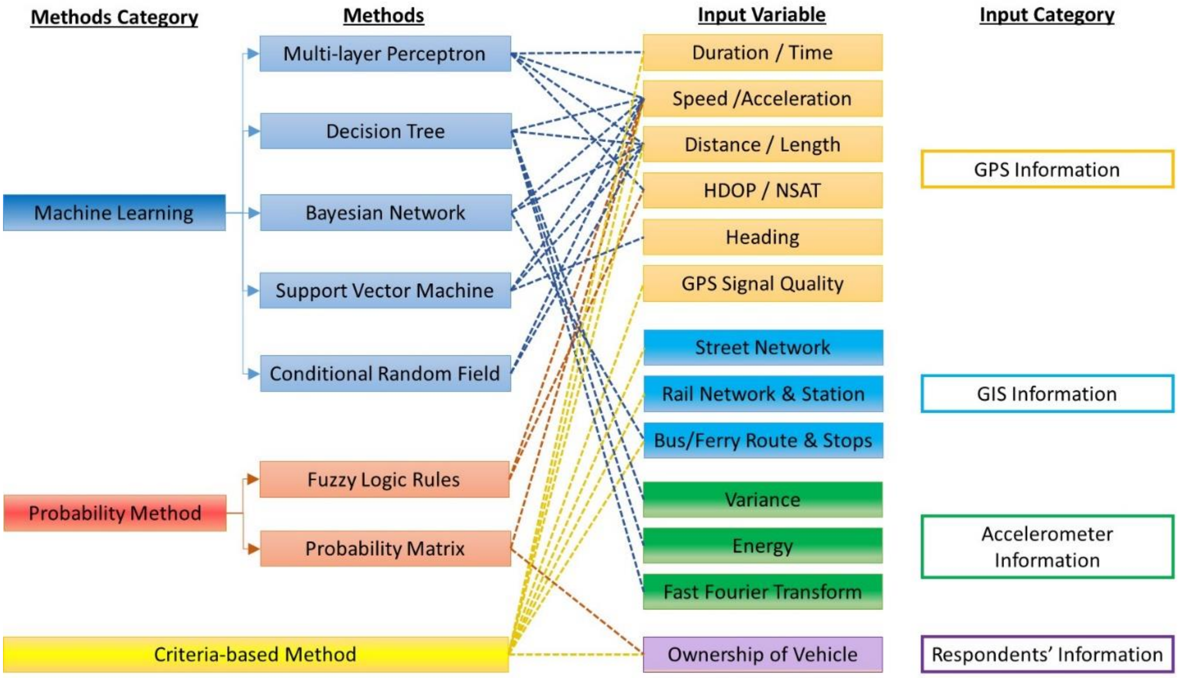


Figure 2.1 Methodologies for mode detection in existing researches and inputs variables

Furthermore, the criteria-based method which judges the features of each segment of trip according to a series of rules, is also utilized in some researches (Stopher et al., 2005, 2008b; Bohte et al., 2008; Chen et al., 2010; Gong et al., 2012). The input variables used in these methodologies come from 3 dimensions, the features of GPS, GIS, and demographic information of participants respectively.

Detailed description of input variables utilized in each method and the corresponding accuracy rate are listed in Table 2.3.

Table 2.3 Summary of methods of transportation mode detection utilized in existing researches

Year	Authors	Steps/ Methods	Input variables	Accuracy
2003	Patterson et al.	Bayesian Model with Expectation Maximization	Velocity, standard deviation of the velocity in the previous 60sec. bus routes & stops	84%
2005 2008b	Stopher et al.	Criteria-based Method	85th speed, 85th acceleration, maximum speed, maximum acceleration, ownership of bicycle, GIS file (including rail line, ferry route, bus route & stops, intersection), GPS signal quality	95%
2006	Tsui and Shalaby	Fuzzy Logic Model	Average speed, 95th speed, positive median acceleration, GPS data validity ratio in segment	91%
2007	Byon et al.	Multi-layer Perceptron neural network	Speed, acceleration, average HDOP, average NSAT	80%
2008	Reddy et al.	Decision tree with a first-order Hidden Markov Model	variance, energy, and sum of FFT(Fast Fourier Transform) coefficients between 1~5 Hz from accelerometer; speed from GPS	98.8%
2008a	Stopher et al.	Two steps; probability matrix used in first step	Ownership of bicycle, average speed, maximum speed, most frequent speed, distance of trip, street & public transport network in GIS	95%
2008	Zheng et al.	Decision Tree (DT), Support Vector Machine	Length, mean velocity, expectation of velocity, covariance of velocity, top three velocities &	74% (DT) 59% (SVM)

		(SVM), Bayesian Net (BN), Conditional Random Field (DRF)	accelerations of the segment of trips	70% (BN) 47% (DRF)
2009	Bohte and Matt	Criteria-based Method	Average speed, maximum speed, public transport network in GIS	70%
2009	Schuessler et al.	Fuzzy Logic Approach	median of the speed distribution, 95th speed, 95th acceleration	NA
2010	Gonzalez et al.[1]	Multi-Layer Perceptron Neural Network	For all points case: average & maximum speed, estimated horizontal accuracy uncertainty, Percent Cell-ID Fixes, standard deviation of distances between stop locations and average dwell time For critical points case: average & maximum acceleration, average & maximum speed, ratio of the number of critical points over the total distance/time of the trip, total distance, and average distance between critical points	88.6% (all points) 91.2% (critical points only)
2010	Chen et al.	Criteria-based Method	Travel time, speed, network of public transport and bus route & stops in GIS, GPS signal quality	79.1%
2010	Moiseeva and Timmermans	Bayesian belief network	distance to the railway track, average and maximum acceleration, average speed, maximum deviation from the average speed and accumulated distance for the threshold period of the time (3min in this research)	NA
2011	Zhang et al.	Two-stage approach; Support Vector Machine used in 2nd stage	Mean speed, maximum speed, mean heading changes,	93%
2012	Gong et al.	Criteria-based Method	Average speed, 85th speed, duration, distance to the rail/subway stations and bus stops, 85th speed, 95th acceleration	82.6%
2013	Pereira et al.	support-vector machine	data of GPS and accelerometer	NA

Note: 1. two sets of data base, namely all GPS points and critical GPS points, were used in the research. The input variable varies depending on the data set.

2.1.5. Activity type identification

Figure 2.2 illustrates the information of categorized methodologies for identifying activity type in the existing researches. These methods can be grouped into 3 categories.

The most widely used method for activity type identification is the rules-based method (Wolf et al., 2001; Stopher et al., 2005, 2008ab; Bohte and Matt, 2009; Chen et al., 2010; Pereira et al., 2013). Methods in this category match the selected information from GPS, GIS and respondents with a series of pre-defined heuristic rules to distinguish the activity type.

The second category is probabilistic method (Axhausen et al., 2004; Chen et al., 2010). The probability of each purpose is calculated based on the different values of information from GPS, GIS and respondents' information. The estimated activity type is decided by the calculated probability.

The third category is the machine learning (McGowen and McNally, 2007; Deng and Ji, 2010). Assisted with the information from GIS, respondents and transportation mode, different kinds of activity type can be separated by the learned knowledge from the training data set.

Detailed description of input variables in each method and the corresponding accuracy can be found in Table 4.

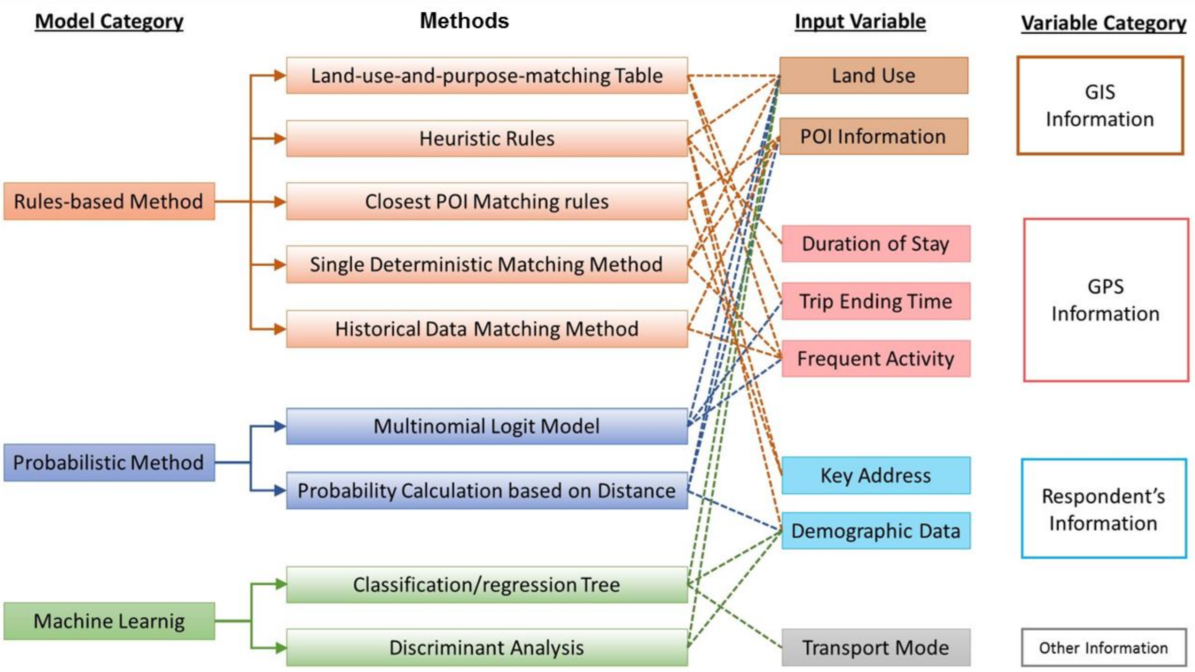


Figure 2.2 Methodologies for activity type inference in existing researches and inputs variables

Note: **POI (Point of Interest)** is the specific trip attraction points, such as restaurants, banks, petrol stations, business locations, mode interchange area etc.

Demographic Data include socio-demographic data and socioeconomic data of the participants.

Table 2.4 Summary of methods of activity type inference in the existing researches

Year	Authors	Methods	Input variables besides coordinates	Accuracy
2001	Wolf et al.	Land-use-and-purpose-matching table	Land use, trip ending time, duration of stay	93%
2004	Axhausen et al.	Probability Calculation based on Distance	Socio-demographic data, land use, distance to POI/ land use polygon	NA
2005 2008ab	Stopher et al.	Heuristic rules	Land use, duration, occupation and address of home/ school/ workplace/ frequently used grocery store of the respondents	NA
2007	McGowen and McNally	Category model (discriminant analysis and classification /regression trees model)	Land use, demographic data	73% and 74%
2009	Bohte and Matt	Closest POI matching rules	Home/work address, locations of POI	43%
2010	Chen et al.	Low-density area: Single deterministic matching method High-density area: Multinomial Logit model	Business listings, frequently visited locations, land use Trip ending time, activity frequency, land use	NA
2010	Deng and Ji	Decision tree methods	Trip ending time, speed, mode, trip distance, trip duration, occupation, income, family structure, age, land use	87.6%
2013	Pereira et al.	Historical data matching rules	Previous validation, points of interest, mode interchange	NA

Note: Previous validation is the historical travelling data of the respondents.

2.2. ACTIVITY-TRAVEL PATTERN ANALYSIS

A sequence of trips is one of the results when people realize the needs of activities. People organize these trips influenced by various factors and try to obtain the maximum utility during the process of organizing trips. The number of trips in a trip chain, the number of trip chains in a day, and the total number of trips are three variables related to how people organize their trips (or travel demand) during the day. And it becomes a hot topic of identifying and analyzing key factors influencing variability of trip organizing. Existing research has already revealed the influencing factors from the dimensions of gender (McGuckin and Murakami, 1999), occupation (Bayarma et al. 2007; Chu, 2004), land use together with density (Dharmowijoyo et al. 2015; Noland and Thomas, 2007; Schmöcker et al. 2010; Susilo and Maat, 2007), weather (Liu et al. 2014; Liu et al. 2015a), etc. These researches have been summarized in Table 2.5.

Table 2.5 Summary of the existing research on trip (chain) variations

Year	Authors	Topic	Main findings	Data set	Period for data collection
1999	McGuckin and Murakami	Examine trip-chaining behavior of adult men and women on weekday	Women tend to make more trips and chain more trips to the trip to and from work.	1995 Nationwide Personal Transportation Survey (conducted by telephone, using a computer-assisted telephone interviewing system)	24-hour period
2007	Ye et al.	Relationship between mode choice and the complexity of trip chaining patterns	Causal structure in which trip chain complexity precedes mode choice performs best for both work and non-work tour samples.	2000 Swiss Microcensus travel survey (trip diary)	one-day
2007	Noland and Thomas	Relationship between patterns of trip chaining and urban form (density)	Lower density lead both to a greater reliance upon trip chaining and to tours that involve more stops.	US Department of Transportation's 2001 National Household Travel Survey.	24-hour period
2007	Golob et al.	Relationship between trip chaining travel activity and age	After age 64, travel demand shifts from car to car passenger and then to public transport in complex trip chains	a pooled (2002–2004) cross-section of the Sydney travel survey	24-hour period
2010	Schmöcker et al.	The trip chaining complexity of individuals in London	Older people on average make more complex tours	London Area Travel Survey	one-day
2011	Currie and Delbosc	Trip chaining behavior of Melbourne residents	Complexity of chains was Found to be larger for rail and tram than for car-based trips	Victorian Activity Travel Survey from 1994 to 1999	one-day
2011	Andrews et al.	Examine perceptions, motivations and decisions relating to use of free bus passes	Elderly people tend not to do complex trip chaining and but to have more flexible time-space constraints.	On-board bus survey of 487 concessionary bus pass holders (by questionnaire)	period of one bus trip

2014	Liu et al	Explores the interactions between time allocation, travel demand and mode choice under different weather conditions	Trade-offs between routine and leisure activities under abnormal weather conditions.	Four Swedish National Travel Survey (NTS) datasets, covering respectively from 1998 to 2001, 2003 to 2004, 2005 to 2006 and 2011.	one-day
2014	Chen and Mahmassani	Impact of rainfall precipitation on activity decisions.	Travel behavior may differ under rainfall	2000 Bay Area Travel (BATS) survey, through land-line telephone	two-day period (one weekday and one weekend)
2015 a	Liu et al	Explore the influence of weather on individuals' trip chaining complexity.	The 'ground covered with snow' condition is the most influential factor on the complexity	Swedish National Transport Survey(NTS) Data including four data bases, each covering certain period of time: 1994–2001, 2003–2004, 2005–2006 and 2011	one-day
2015 b	Liu et al	influence of weather on mode choice decision in different seasons and regions	The impacts of weather differ in different seasons and different regions	Swedish National Transport Survey(NTS) Data including four data bases, each covering certain period of time: 1994–2001, 2003–2004, 2005–2006 and 2011	one-day

In terms of the relationship with the weather variables, rainfall and snow are two of the most significant variables influencing the complexity of trip chaining. As mentioned in the introduction section, one demerit of these existing research is that the data used are almost one-day data. Weather variability is achieved by collecting data in various locations, i.e. cross-section data (no matter at the same time or not) with spatial variability which is used to represent the weather's variability. GPS data in a longer period provide the possibility to observe the weather's variability in the same city. It would be interesting to analyze the weather's influence with time-series data, and compare the weather's influence of special heterogeneity with temporal heterogeneity.

2.3. SUMMARY

This part carefully summarized and categorized the methodologies utilized for GPS data error recognition, trip identification, travel mode detection and activity type identification based on GPS data in the existing researches. Also, the input variables in each method and corresponding accuracy are summarized in the tables for further convenient comparison. Compared to probability method and criteria-based method, machine learning is often applied in detecting travel mode. On the other hand, rules-based methods are more popular than probabilistic method and machine learning as the tool for inferring the activity type. It would be necessary to apply more machine learning methods to identify activity type and find the suitable one.

Organizing trips in a day suffers the influence from the field of gender, land use, weather, etc. GPS data collected in a longer period provide the possibility to check the weather's influence in a local place. It has already been revealed that weather's influence on trip chaining by considering weather's variation from the perspective of spatial heterogeneity. And it is necessary to check weather's influence on trip chaining behavior from the perspective of temporal heterogeneity.

2.4. REFERENCE

Agamennoni, G., Nieto, J., and Nebot, E. (2009). Mining GPS data for extracting significant places. *IEEE International Conference on Robotics and Automation*. ICRA'09. IEEE. 855-862

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., and Vaisman, A. (2007). A model for enriching trajectories with semantic geographical information. *Proc. 15th annual ACM international symposium on Advances in geographic information systems*. ACM. 22.

Andrews, G., Parkhurst, G., Susilo, Y. O., and Shaw, J. (2012). The grey escape: investigating older people's use of the free bus pass. *Transportation Planning and Technology*, 35(1), 3-15.

Andrienko, G., Andrienko, N., Fuchs, G., Raimond, A. M. O., Symanzik, J., and Ziemlicki, C. (2013). Extracting semantics of individual places from movement data by analyzing temporal patterns of visits. Proc. T*he First ACM SIGSPATIAL International Workshop on Computational Models of Place* (COMP'13).

Ashbrook, D., and Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5), 275-286.

Axhausen, K.W., Schonfelder, S., Wolf, J. Oliveira, M. and Samaga, U. (2004). Eighty Weeks of GPS Traces: Approaches to Enriching Trip Information. *Proceedings of the 83rd Annual Meeting of the Transportation Research Board*, January 2004, Washington D.C.

Bayarma, A., Kitamura, R., and Susilo, Y. (2007). Recurrence of daily travel patterns: stochastic process approach to multiday travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, (2021), 55-63.

Bohte, W. and 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*. pp.285-297

Byon, Y.J., Abdulhai, B. and Shalaby, A. (2007). Impact of sampling rate of GPS-enabled cell phones on mode detection and GIS map matching performance. *Proceedings of the 86th Annual Meeting of the Transportation Research Board*, January 2007, Washington D.C.

Chen, C., Gong, H. Lawson, C. and 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*. pp.830-840

Chen, R. B., and Mahmassani, H. S. (2015). Let it rain: Weather effects on activity stress and scheduling behavior. *Travel Behaviour and Society*, 2(1), 55-64.

Chu, Y. L. (2004). Daily stop-making model for workers. *Transportation Research Record: Journal of the Transportation Research Board*, (1894), 37-45.

Currie, G., and Delbosc, A. (2011). Exploring the trip chaining behaviour of public transport users in Melbourne. *Transport Policy,* 18(1), 204-210.

Deng, Z. and Ji, M. (2010) Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. *Traffic and Transportation Studie*s. pp. 768-777.

Dharmowijoyo, D. B., Susilo, Y. O., and Karlström, A. (2015). Day-to-day variability in travellers' activity-travel patterns in the Jakarta metropolitan area. *Transportation*, 1-21.

Golob, T. F., and Hensher, D. A. (2007). The trip chaining activity of Sydney residents: a cross-section assessment by age group with a focus on seniors. *Journal of Transport Geography*, 15(4), 298-312.

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

Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J and Labrador, M.A. (2010). Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. *Intelligent Transport System*s, IET Volume: 4, Issue: 1. pp. 37-49

Hato, E., Shinji, I. and Mitani, T. (2006). Development of MoALs (Mobile Activity Loggers supported by GPS-phones) for Travel Behaivor Analysis. *Proceedings of the 85th Annual Meeting of the Transportation Research Board*, January 2006, Washington D.C.

Itsubo, S. and Hato, E. (2006). A study of the effectiveness of a household travel survey using GPS-equipped cell phones and a WEB diary through a comparative study with a paper based travel survey. *Proceedings of the 85th Annual Meeting of the Transportation Research Board*, January 2006, Washington D.C.

Kami, N., Enomoto, N., Baba, T., and Yoshikawa, T. (2010). Algorithm for detecting significant locations from raw GPS data. *Discovery Science*. Springer Berlin Heidelberg. 221-235

Leclerc, B., Trépanier, M., and Morency, C. (2013). Unraveling the Travel Behavior of Carsharing Members from Global Positioning System Traces. *Transportation Research Record: Journal of the Transportation Research Board*, 2359(1), 59-67.

Lee, Y.S. and Cho, S.B., 2014. Activity recognition with android phone using mixture-of-experts co-trained with labeled and unlabeled data. *Neurocomputin*g, 126, pp.106-115.

Liu, C., Susilo, Y. O., and Karlström, A. (2014). Examining the impact of weather variability on non-commuters' daily activity–travel patterns in different regions of Sweden. *Journal of Transport Geography,* 39, 36-48.

Liu, C., Susilo, Y. O., and Karlström, A. (2015a). Measuring the impacts of weather variability on home-based trip chaining behaviour: a focus on spatial heterogeneity. *Transportation,* 1-25.

Liu, C., Susilo, Y. O., and Karlström, A. (2015b). The influence of weather characteristics variability on individual’s travel mode choice in different seasons and regions in Sweden. *Transport Policy*, 41, 147-158.

McGowen, P. and McNally, M. (2007). Evaluating the Potential to Predict Activity Types from GPS and GIS Data. *Western Regional Science Association 46th Annual Meeting,* November 2007, Newport Beach, CA

McGuckin, N., and Murakami, E. (1999). Examining trip-chaining behavior: Comparison of travel by men and women. *Transportation Research Record: Journal of the Transportation Research Board*, (1693), 79-85.

Mizuno K., Kanamori R., Sano S., Nakajima S. and Ito T. (2013). Identifying move and stop in GPS data with Support Vector Machines, *Conference of Infrastructure Planning and Management* (CD-ROM), JSCE,

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

Murakami, E. and Wagner, D.P. Can Using Global Positioning System (GPS) Improve Trip Reporting? *Transportation Research-C*, 7(3/4), 1999, pp. 149-165.

Noland, R. B., and Thomas, J. V. (2007). Multivariate analysis of trip-chaining behavior. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 34(6), 953.

Palma, A. T., Bogorny, V., Kuijpers, B., and Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. *Proc. 2008 ACM symposium on applied computing*. ACM. 863-868

Patterson, D.J., Liao, L., Fox, D. and Kautz, H. (2003). Inferring high-level behavior from low-level sensors. In UbiComp 2003: *Ubiquitous Computing* (pp. 73-89). Springer Berlin Heidelberg.

Pereira, F., Carrion, C., Zhao, F., Cottrill, C.D., Zegras, C and Ben-Akiva, M. (2013). The future mobility survey: overview and preliminary evaluation. *Proceedings of the 10th International Conference of Eastern Asia Society for Transportation Studies*. Taipei Taiwan

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

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

Schuessler, N. and Axhausen, K. W. (2009). Processing GPS raw data without Additional Information. *Proceedings of the 88th Annual Meeting of the Transportation Research*

Board, January 2009, Washington D.C.

Schmöcker, J. D., Su, F., and Noland, R. B. (2010). An analysis of trip chaining among older London residents. *Transportation*, 37(1), 105-123.

Stopher, P., Bullock, P. and Jiang, Q. (2002). GPS, GIS and personal travel surveys: an exercise in visualisation, *25th Australasian Transport Research Forum Incorporating the BTRE Transport Policy Colloquium*, October 2002, Canberra

Stopher, P. R, Jiang, Q. and FitzGerald, C. (2005). Processing GPS data from travel surveys. Proceedings of 2nd *Int. Colloquium on the Behavioral Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications*. June 2005. Toronto, Canada.

Stopher, P., FitzGerald, C. and Zhang, J. (2008a). Search for a global positioning system device to measure person travel. *Transport Research Part C*. pp.350-369

Stopher, P., Clifford, E., Zhang, J. and FitzGerald, C. (2008b). Deducing mode and purpose from GPS data. Working paper ITLS-WP-08-06. Institute of Transport and Logistic Studies, the Australian Key Center in Transport and Logistic Management, the Univiersity of Sydney

Susilo, Y. O., and Maat, K. (2007). The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation*, 34(5), 589-609.

Tran, L. H., Nguyen, Q. V. H., Do, N. H., and Yan, Z. (2011). *Robust and Hierarchical Stop Discovery in Sparse and Diverse Trajectories* (No. EPFL-REPORT-175473).

Tsui, S. Y. A, Shalaby, A. S. (2006). An Enhanced System for Link and Mode Identification for GPS-based Personal Travel Surveys. *Proceedings of the 85th Annual Meeting of the Transportation Research Board*, January 2006, Washington D.C.

Wagner, D.P., (1997). Report: *Lexington Area Travel Data Collection Test: GPS for Personal Travel Surveys*. Final Report for OHIM, OTA, and FHWA.

Wolf, J., Guensler, R., andBachman, W. (2001). Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data. *Proceedings of the 80th Annual Meeting of the Transportation Research Board*, January 2001, Washington D.C.

Xie, K., Deng, K., and Zhou, X. (2009). From trajectories to activities: a spatio-temporal join approach. *Proc. 2009 International Workshop on Location Based Social Networks*. ACM. 25-32

Yan, Z., Parent, C., Spaccapietra, S., and Chakraborty, D. (2010). A hybrid model and computing platform for spatio-semantic trajectories. *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg. 60-75.

Ye, X., Pendyala, R. M., and Gottardi, G. (2007). An exploration of the relationship between mode choice and complexity of trip chaining patterns. *Transportation Research Part B*: Methodological, 41(1), 96-113

Zimmermann, M., Kirste, T., and Spiliopoulou, M. (2009). Finding stops in error-prone trajectories of moving objects with time-based clustering. *Intelligent Interactive Assistance and Mobile Multimedia Computing*. Springer Berlin Heidelberg. 275-286

Zitto R., D’este G. and Taylor P. Global positioning systems in the time domain: how useful a tool for intelligent vehicle-highway systems? (1995) *Transportation Research-C*, 3(4). pp. 193-209.

Zhang, L., Dalyot, S. Eggert, D. and Sester, M. (2011). Multi-stage approach to travel-mode segmentation and classification of GPS traces. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XXXVIII-4/W25, 2011. ISPRS Guilin 2011 Workshop, October 2011, Guilin, China

Zheng, Y., Liu, L., Wang, L. and Xie, X. (2008). Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of WWW 2008*, (Beijing China, April 2008), ACM Press: pp.247-256.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems* (TOIS), 25(3), 12.

Chapter 3. Data Set Description

3.1. INTRODUCTION

In this chapter, the data sets used in thesis are introduced. In general, mobile phone GPS data collected in two areas in Japan are utilized in this thesis. One is Nagoya metropolitan area; the other one is Hakodate city, Hokkaido. These two data sets are chosen for the following reasons. Data collected in Nagoya area include the GPS data collected for 5 weeks on average by 80 participants. Data collected in Hakodate city include the GPS data collected for 4 months in summer and 4 months in winter by 2 persons (totally 20 persons attended the survey; the other 18 persons' data are not included because they are now under checking of ground truth). Data in Nagoya area is suitable to analyze activity type identification since demographic information of multiple-persons is available and the contribution of demographic variables can be analyzed. Data in Hakodate city is suitable to analyze weather's influence on machine learning since collecting data during 8 months can catch weather's fluctuation.

In each subsection, a general introduction of the data set will be given. It is followed by a descriptive analysis of the data set.

3.2. DATA SET I

3.2.1. General information

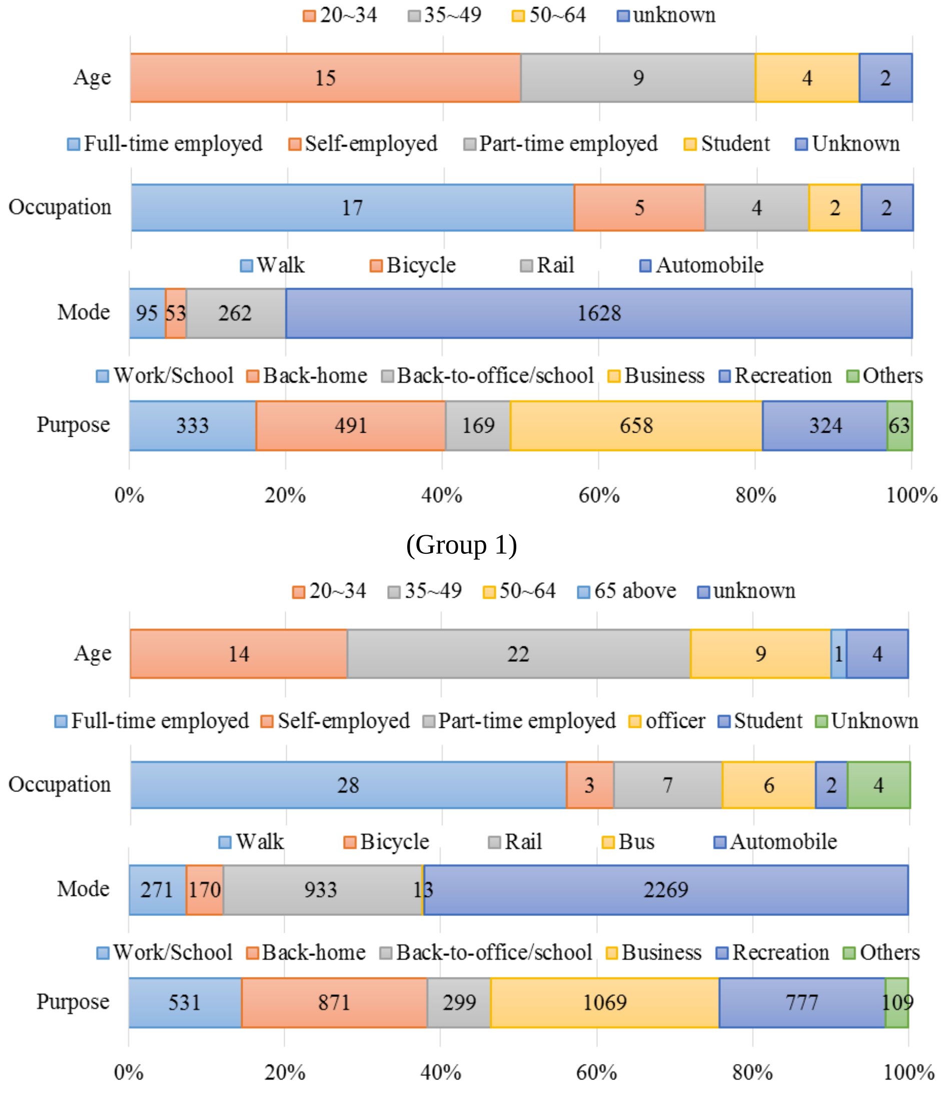
The GPS data utilized in the research were collected from 30 participants (1st group) and 50 participants (2nd group) in the Nagoya area of Japan over a period of five weeks in 2008. These two groups of participants are independently chosen. Each participant was assigned a mobile phone with GPS censor able to record and send GPS information to the server every 10 seconds. The GPS information sent back includes longitude, latitude, time stamp, signal quality, etc. Sometimes the GPS module returns GPS information with intervals longer than 10 seconds, such as in the case of communication delay because of tunnels, subways, etc. The signal quality feature of the GPS data is used to identify this kind of signal loss, and points with low signal quality are excluded from the data set for analysis. Overall, 97.4% of the GPS communication intervals in the data set are less than 20 seconds.

Speed and acceleration are not available in this dataset. On the other hand, volunteers were required to annotate the information by inputting the start, end, mode and purpose of each trip through an application installed on the smart phones. Additionally, socio-demographic information about each volunteer was collected through a questionnaire, including home and workplace addresses, occupation, yearly income of household, possession of a driving license, daily primary transportation mode and so on.

About the demographic information of the participants, the first group lacks 2 persons' data and the second group lacks 4 persons' data. And regarding the driving licenses, all the participants in the first group own the driving licenses while the second group does not include this information.

3.2.2. Descriptive analysis

Figure 3.1 illustrates the basic aggregated statistical analysis of this dataset. Almost all participants are in the age group 20–65, the working age in Japan, and almost all have a full time or part time job. This means these volunteers are active trip makers. Automobile, walking and rail are the main modes of transport; business, returning home and going to work are the main trip purposes in this dataset.



(Group 1)

(Group 2)

Figure 3.1 Aggregated statistical analysis of dataset I of Nagoya

Table 3.1 shows the mean value and standard error value of travel time and activity duration at the trip end in the data set Nagoya area. Travel time of "work/school" and "back home" are almost the same, around 40 minutes; travel time of the other four types of trip varies from 22 minutes to 28 minutes. Regarding activity duration, the largest is staying home and it is followed by “work/school”. This result agrees with the common knowledge. Activity duration of non-mandatory activity types are much less than the activity of “back home” or “work/school”.

Table 3.1 Statistical results of travel time and activity duration at trip end

Activity type	Travel time (min)		Activity duration (min)	
	Mean	StDev	Mean	StDev
Back-home	39	40	670	308
Back-to office/school	28	26	163	149
Business	25	26	60	123
Work/school	40	23	405	261
Others	22	18	51	91
Recreation	24	25	73	99

Following Figure 3.2 and Figure 3.3 show the distributions of travel time and the activity duration at the trip end respectively.

Regarding travel time, around 70% of the trips with the purpose of “back-home” or “work/school” are less than 50 minutes while around 70% of the trips with the purpose of “business”, “back-to office/school”, “others” or “recreation” are less than 30 minutes.

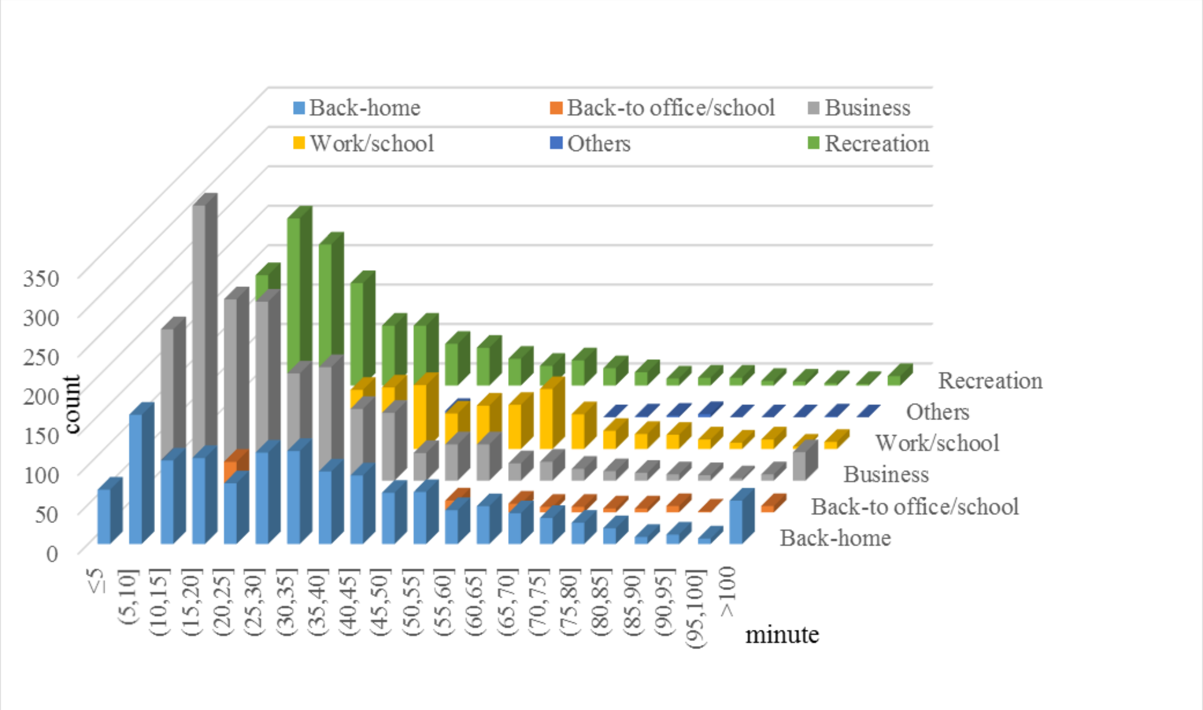


Figure 3.2 Distribution of travel time by activity type

With regard to activity duration at trip end, around 64% of the activities with “back-home” are more than 600 minutes. On the contrary, around 65% of the “business” activities are in less than 30 minutes, around 63% of the “others” activities are less than 30 minutes, and around 65% of the “recreation” activities are less than 60 minutes.

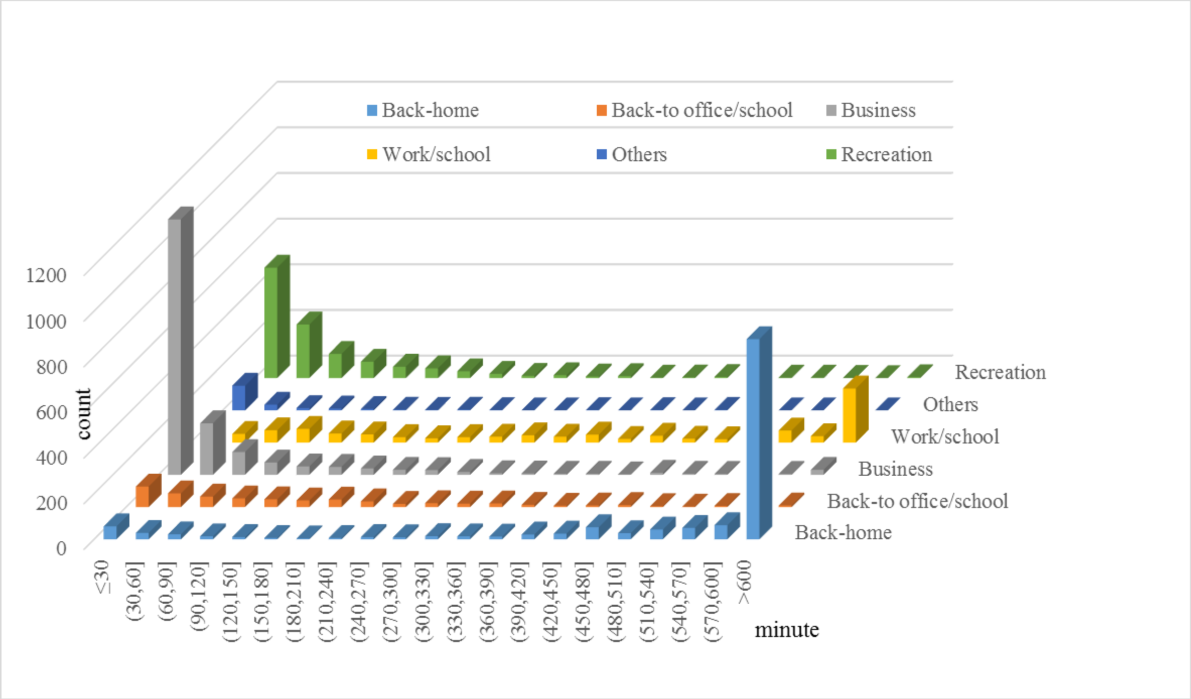


Figure 3.3 Distribution of activity duration by activity type

3.3. DATA SET II

3.3.1. General information

GPS data collected in Hakodate city are over a non-contiguous period of 8 months, including winter from December 2012 to April 2013 and summer from June 2013 to October 2013. GPS data were collected by 20 participants. Each volunteer was assigned a smart mobile phone fitted with a GPS module. An android application was used to collect GPS information (at 30 second intervals) together with travel plans, destinations, trip purposes and travel modes, etc. as input information (Sano et al. 2013).

3.3.2. Descriptive analysis

By now, data of only 2 participants out of these twenty persons are used, since the other 18 participants' data are under check for ground truth. Although the participants have already input their travel plans on the survey days, some errors have been found in these plans. As a result, it is not appropriate to directly use these input plans as ground truth for activity type identification. And it is why data of only two participants are used in this thesis. After the checking of ground truth of other 18 participants, the effect of additional GIS features and weather effect on the performance of machine learning in Chapter 6, as well as the ordered logit model estimation in Chapter 7 should be recalculated and updated.

The general demographic information of these two participants are as follows. One is male and has a habit of using more diverse travel modes. Another one is female who is more vehicle-dependent.

Table 3.2 Demographic information of the two volunteers in this paper

Person ID	Gender	Age	Holding driving license	Driving frequency	Working places
Participant A	Male	60s	Yes	Sometimes per month	Several ones
Participant B	Female	40s	Yes	Almost everyday	Single one

Composition of travel modes and activity type of these two participants in summer and winter can be found in the following figure. For participant A, season has an obvious influence on his mode choice, since trips by bicycle only occur in summer and in winter there are more trips by walk. As for participant B, since she is vehicle-dependent, there is no obvious seasonal effect on her travel mode choices.

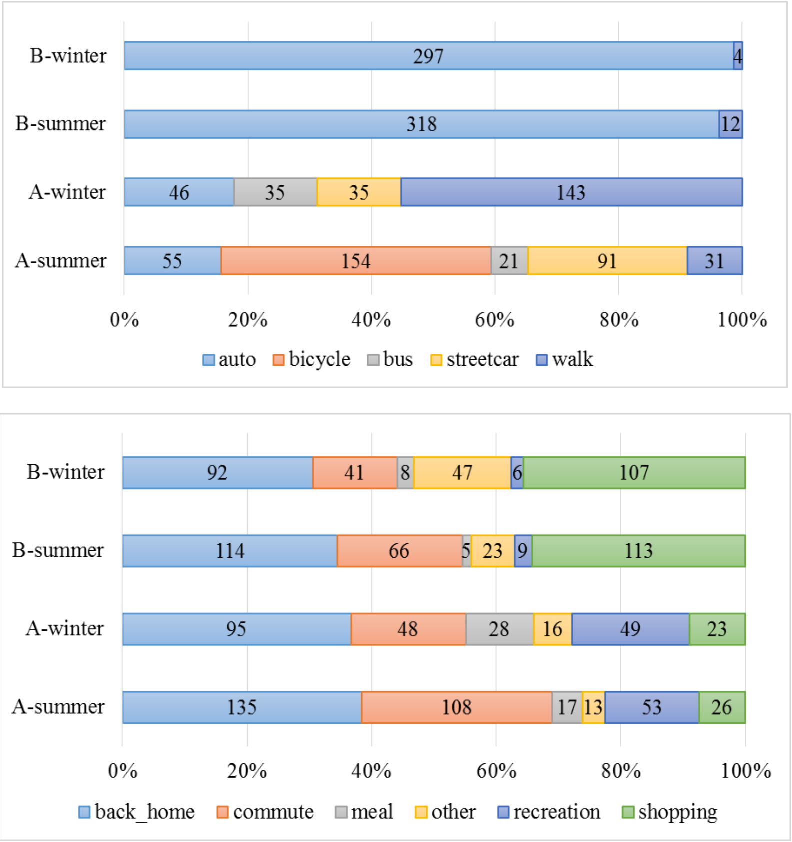
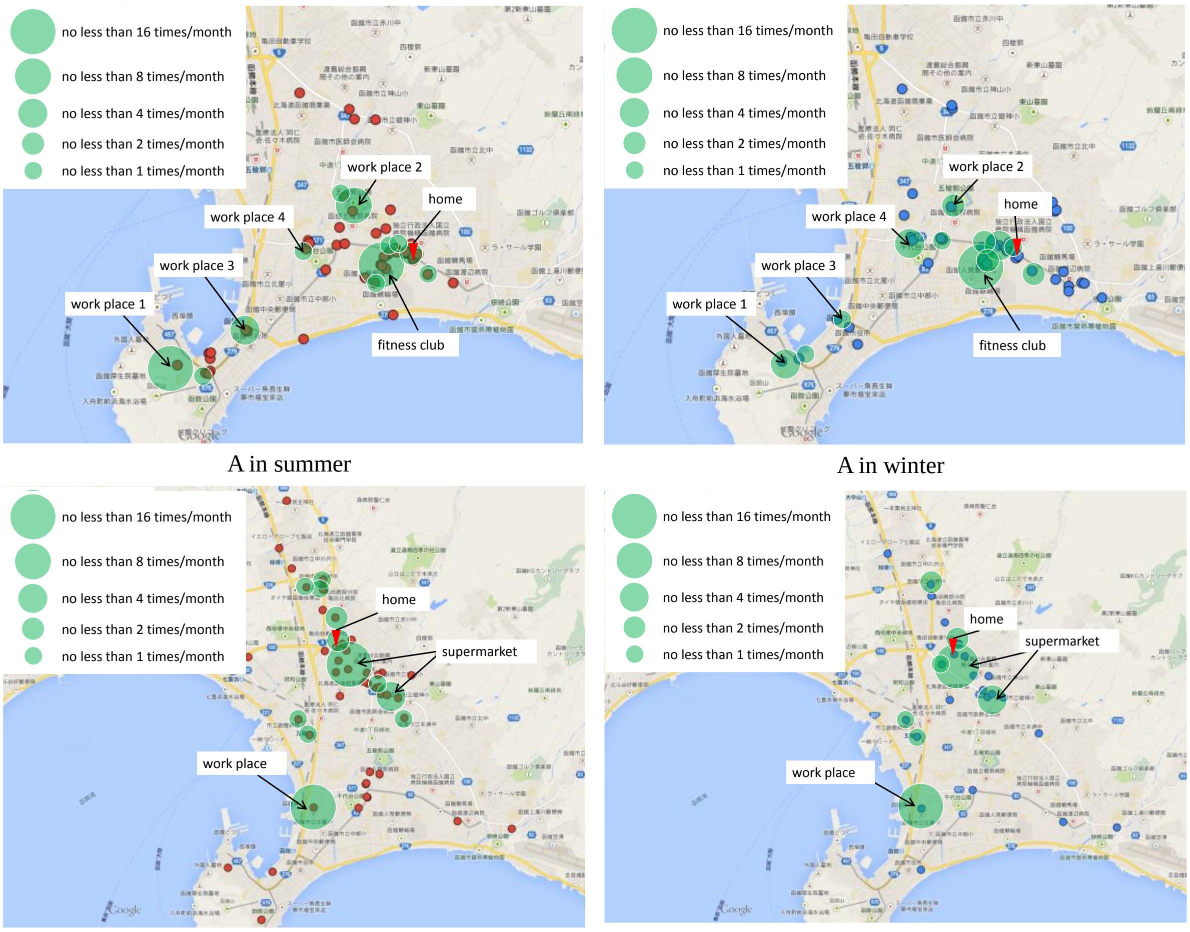


Figure 3.4 Composition of travel mode and activity type in Hakodate data

Since data collected in Hakodate last four month in summer and four month in winter, in this section, a focus of variance of travel pattern among seasons are shown. Figure 3.4 shows all destinations that participants visited during the investigation period with the corresponding visiting frequency.

Participant A particularly often visits to offices and fitness club. There are several offices and visiting frequency of each office differs greatly. Participant B often visits the office and supermarkets. Contrary to participant A, participant B has an extremely steady lifecycle because she goes to the office and supermarkets and seldom change frequency in summer and winter.


no less than 16 times/month
no less than 8 times/month
no less than 4 times/month
no less than 2 times/month
no less than 1 times/month
work place 2
home
work place 4
work place 3
work place 1
fitness club
no less than 16 times/month
no less than 8 times/month
no less than 4 times/month
no less than 2 times/month
no less than 1 times/month
work place 2
home
work place 4
work place 3
work place 1
fitness club
no less than 16 times/month
no less than 8 times/month
no less than 4 times/month
no less than 2 times/month
no less than 1 times/month
home
supermarket
work place
no less than 16 times/month
no less than 8 times/month
no less than 4 times/month
no less than 2 times/month
no less than 1 times/month
home
supermarket
work place
Google

A in summer

A in winter

B in summer

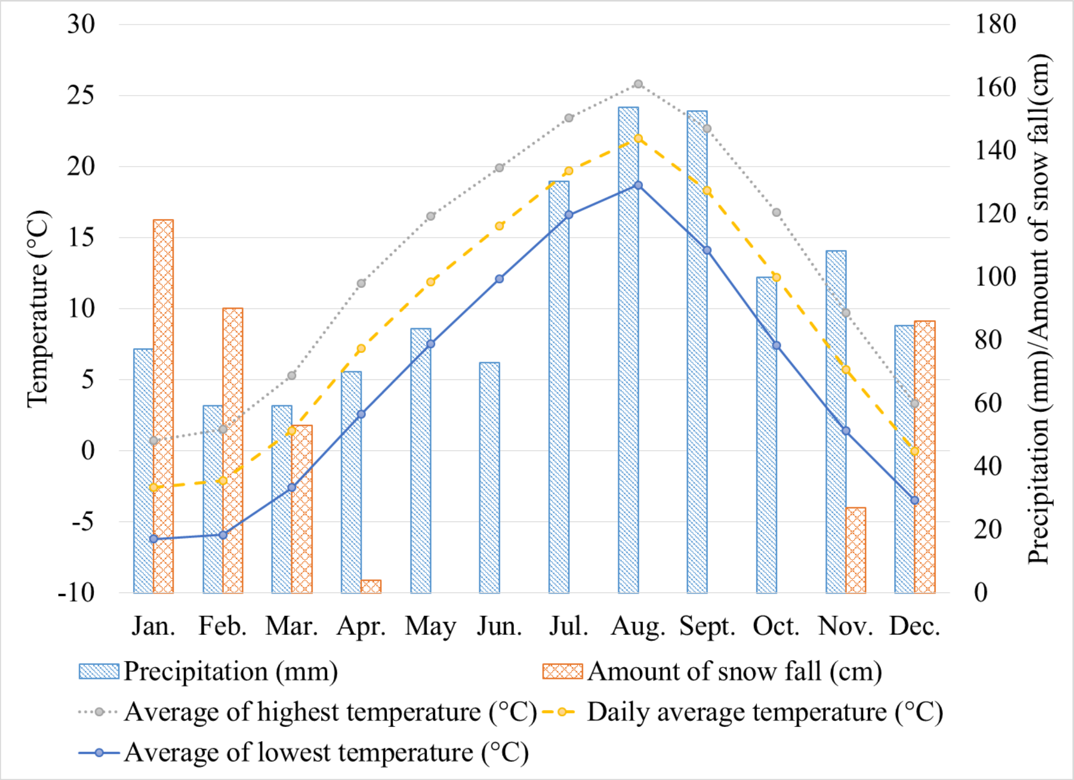
B in winter

Figure 3.5 Visiting destinations and frequency by season

3.3.3. Climate introduction in Hakodate City

Finally, in this part, a general introduction of Hakodate city is necessarily made because of its characteristics of distinct seasons.

Hakodate is a city with distinct seasons which can be concluded by the information in Figure 3.6 which shows average weather conditions over the 30 years from 1981 to 2010. It is clear that the temperature, precipitation, and amount of snowfall are distinctly different from summer to winter. The lowest daily average temperature occurs in January with -2.6°C while the highest daily average temperature occurs in August with 22°C. Snowfall is concentrated in the period from December to March, while July, August and September are the three months with the most precipitation. Spring and autumn have moderate temperatures but are relatively short. So it is appropriate to use the data collected in Hakodate with totally different weather characteristics to analyze behavioral variability according to weather’s variability across seasons.



Data source: Japan Meteorological Agency

Figure 3.6 Average weather conditions in Hakodate (30 years from 1981 to 2010)

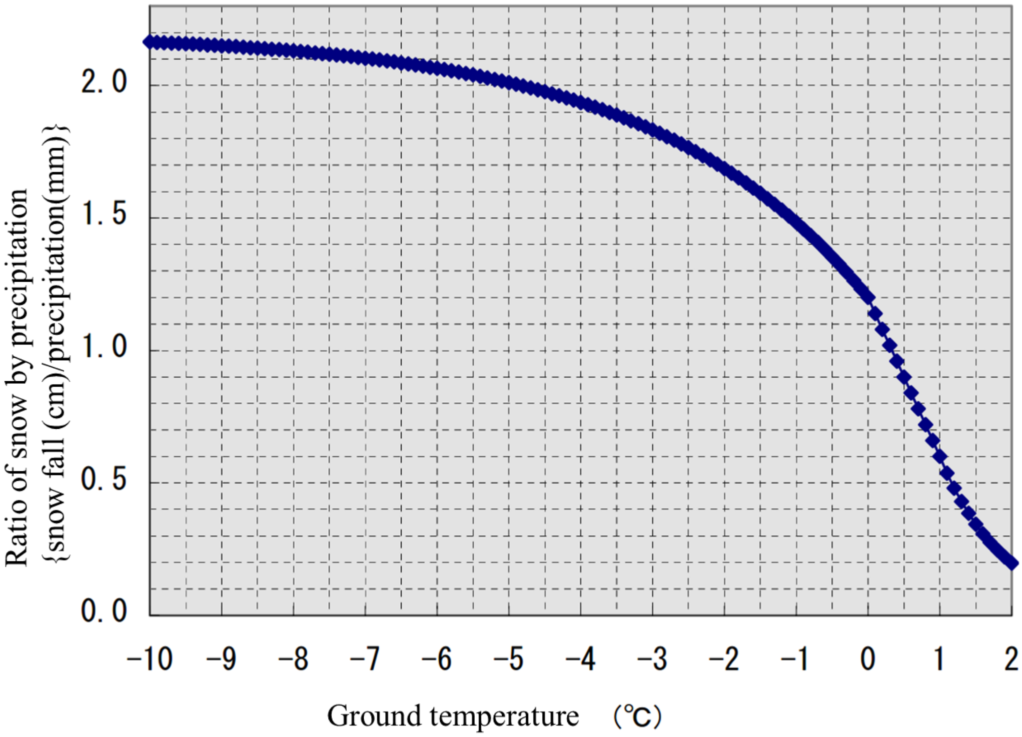


Figure 3.7 Relationship between ground temperature and ratio of snow by precipitation[1]

[1] Translated from the figure in Japanese. Original figure is from this following website: http://www.applenet.jp/dataout/19/5/5_5001.pdf (Visited on 2015/11/18)

Here, the difference between precipitation and amount of snow fall should be noted. Actually "precipitation" includes rain, snow, sleet, drizzle, etc. In case of snow, it should be converted to rain when counted as "precipitation". And the ratio (snow in cm/rain in mm) depends on the temperature. The relationship between the ratio and the temperature can be found in Figure 3.7. The ratio is not always 1 especially in winter.

3.4. SUMMARY

In this chapter, the data sets utilized in this thesis are introduced. There are totally two data sets, collected in Nagoya metropolitan area and Hakodate city respectively. Data set of Nagoya area has 80 participants' data collected during 5 weeks; data set of Hakodate city has 2 participants' data collected in 8 months. Data set of Nagoya area will be used for activity stop identification, activity type identification and activity sequence generation model; data set of Hakodate city will be used for 1) investigation of the performance of machine learning when identifying activity type or travel mode with additional weather and GIS related variables, as well as 2) daily activity-travel pattern analysis.

3.5. REFERENCE

Sano, S., Kanamori, R., Hirata, K. and Nakashima, H. (2013). *Smart City Hakodate Project: Traffic Behavior Reports for a Traffic Flow Simulator.* Workshop of Social System and Information Technology (WSSIT2013) (in Japanese).

Chapter 4. Activity Stop Identification

4.1. INTRODUCTION

Activity stop identification is to segment the trips and activities from continuous GPS trajectories. Most of the existing research used criteria of zero-speed lasting for a given period to detect the activity. However, not all GPS data include the feature of speed. What is more, the threshold of the given period is usually decided in an arbitrary way.

The records in the data sets collected in Nagoya area or Hakodate city do not include the feature like “speed”. It makes the above method unsuitable to these data. In order to segment the activities and trips in these data set, a totally different method must be used according to the data sets’ characteristics. Taking the data collected in Nagoya area as an example, it is collected in an intensive way, in which the GPS record is collected every 10 seconds. When demonstrating these GPS points on a map, scattered points along the road/railway links mean trips while gathered points around a location possibly mean activities. Due to this characteristic, an improved density-based algorithm (called C-DBSCAN, constrained density-based spatial clustering of applications with noise) with support vector machines (SVMs) is proposed to be applied to distinguish the trips and activities at the trip ends.

During continuous GPS trajectories, stops are taken to be obvious signs of some activity taking place or the trip starting/ending. However, a stop does not necessarily equate to activity. Some stops are followed by a certain activity while others are not. In this chapter, two types of stops are defined: activity stops and non-activity stops, respectively. An activity stop is a stop followed immediately by some activity such as work, shopping, recreation and so on; a non-activity stop is one that is not followed by a particular activity, such as waiting for a green light at an intersection or being stuck in traffic. C-DBSCAN algorithm is used to detect all types of stops first. Then SVMs is utilized to distinguish these two kinds of stops.

The data set used in this chapter is from the 1st group of data set I. The full GPS data set is divided almost equally by time sequence for each participant into two parts: a training dataset for estimating parameters in the C-DBSCAN algorithm and training models in SVMs; and a prediction dataset for validating the C-DBSCAN algorithm and testing the learned SVMs. A training set is used to train the machine learning method to learn the potential relationship between independent variables and dependent variables. The learnt relationship is tested on a test set to check the effect of learning from training set and the transferability of the learned relationship.

4.2. STOP LOCATION IDENTIFICATION

In section 2.1.3, the methods of trip segmentation have been summarized. In case of GPS data without feature of "speed", using GPS features to build a bit more complicated methods is necessary. Considering the particular features of the data collected in Nagoya metropolitan area, this chapter proposes a two-step method which uses a density-based clustering method to identify all types of stops at the first step and then uses a supervised machine learning method to distinguish the locations of activity stops and non-activity stops at the second step.

4.2.1. Density based algorithm

In this section, key definitions in the original DBSCAN algorithm are introduced first. Then the improved DBSCAN algorithm, named C-DBSCAN, is interpreted.

4.2.1.1. DBSCAN Algorithm

The same notation as presented by Ester et al. (1996) is used in this section. Applying the key definitions of DBSCAN in the context of GPS tracing points is to separate the stop points and the moving points from the GPS trajectories.

Definition 1: (**Eps-neighborhood of a point**) The Eps-neighborhood of a point *p*, denoted by $N_{Eps(p)}$, is defined by $N_{Eps(p)}=\{q\in D \mid dist(p,q)\leq Eps\}$. Here, $N_{Eps(p)}$ is a set of points in which each point *q* belongs to database *D* and has a distance shorter than *Eps* to point *p*, and *Eps* is a given distance threshold.

Definition 2: (directly density-reachable) A point *p* is directly density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if

1) $p\in N_{Eps(q)}$ and
2) $|N_{Eps(q)}| \geq MinPts$

where *MinPts* is the minimum number of points in the Eps-neighborhood of point *q*.

Definition 3: (density-reachable) A point *p* is density-reachable from a point *q* w.r.t. *Eps* and *MinPts* if there is a chain of points $p_1, \ldots, p_n$, (where $p_1=q$, and $p_n=p$) such that p_{i+1} is directly density-reachable from p_i.

Definition 4: (**density-connected**) A point *p* is density-connected to a point *q* w.r.t. *Eps* and *MinPts* if there is a point *o* such that both *p* and *q* are density-reachable from *o* w.r.t. *Eps* and *MinPts*.

Definition 5: (**cluster**) Let *D* be a database of points. A cluster *C* w.r.t. *Eps* and *MinPts* is a non-empty subset of *D* satisfying the following conditions:

1) $\forall p, q$: if $p \in C$ and *q* is density-reachable from *p* w.r.t. *Eps* and *MinPts*, then $q\in C$. (Maximality)
2) $\forall p, q \in C$: *p* is density-connected to *q* w.r.t. *Eps* and *MinPts.* (Connectivity)

Definition 6: (**noise**) Let $C_1, \ldots, C_k$ be the clusters of database *D* w.r.t. parameters Eps_i and $MinPts_i$, $i=1, \ldots, k$. Then we define the noise as the set of points in database *D* not belonging to any cluster C_i, i.e. noise = $\{p \in D \mid \forall i: p \notin C_i\}$

4.2.1.2. Disadvantage and improvement

When DBSCAN is applied in the situation of GPS track points, points in a cluster are the equivalent of stop points, which gather together with a higher density; on the other hand, points in the noise are the equivalent of moving points along road or rail network links with a lower density.

However, the DBSCAN algorithm was developed to solve the spatial point classification problem without consideration of their temporal sequence. Consequently, in a detoured trajectory, a distinguished stop cluster may contain other moving points or points in subsequent clusters sharing the same location. Moreover, due to the definitions and concepts of the original DBSCAN algorithm, points representing movement along a straight road at low speed when the GPS signal transmission frequency is high may be grouped into a single cluster under certain given parameter values. As a result, applying the original DBSCAN algorithm to GPS trajectories may lead to errors. Here, we advance the C-DBSCAN algorithm in which two constraints are added in order to avoid these two potential errors.

The first constraint is all points in a cluster should be temporally sequential. This means the sequential order should increase one by one and no “sudden increase” is allowed in the cluster. If such a “sudden increase” is found, the cluster will be divided into two potential clusters at the point of sudden increase and each one will be tested to see if it satisfies the condition of minimum number of points in one cluster. If not, the points in the potential cluster will be labeled as moving points. Otherwise the points in the potential cluster will be tested by the second constraint.

The second constraint is that the percentage (*PCT*) of abnormal points in a cluster should not exceed a given threshold named PCT_{AP}. To be specific,

$$PCT \le PCT_{AP} \tag{4.1}$$

where $PCT = \frac{|AP|}{|C|}$, $|AP|$ is the number of abnormal points in the cluster, $|C|$ is the total number of points in the same cluster.

Before giving a definition of an abnormal point, the direction and direction change of a point in a cluster need to be explained, as follows. The **direction** of a point is defined in an imaginary situation in Cartesian coordinates where the point is the origin and the direction is defined as the angle between the negative direction of the vertical axis and the line between the point and the previous point, like α_1 for point P_1 and α_2 for point P_2 in Figure 4.1. Suppose three points in the cluster are marked sequentially as P_0, P_1 and P_2. The **direction change** from point to point P_1 is defined as the angle from ray $\overline{P_0P_1}$ to ray $\overline{P_1P_2}$, shown as $\Delta\alpha$ in Figure 4.1. $\Delta\alpha$ is the angular change between α_1 and α_2, i.e. $\Delta\alpha = \alpha_2 - \alpha_1$. Since the cosine value of $\Delta\alpha$ is used, it does not matter whether $\Delta\alpha$ is negative or positive.

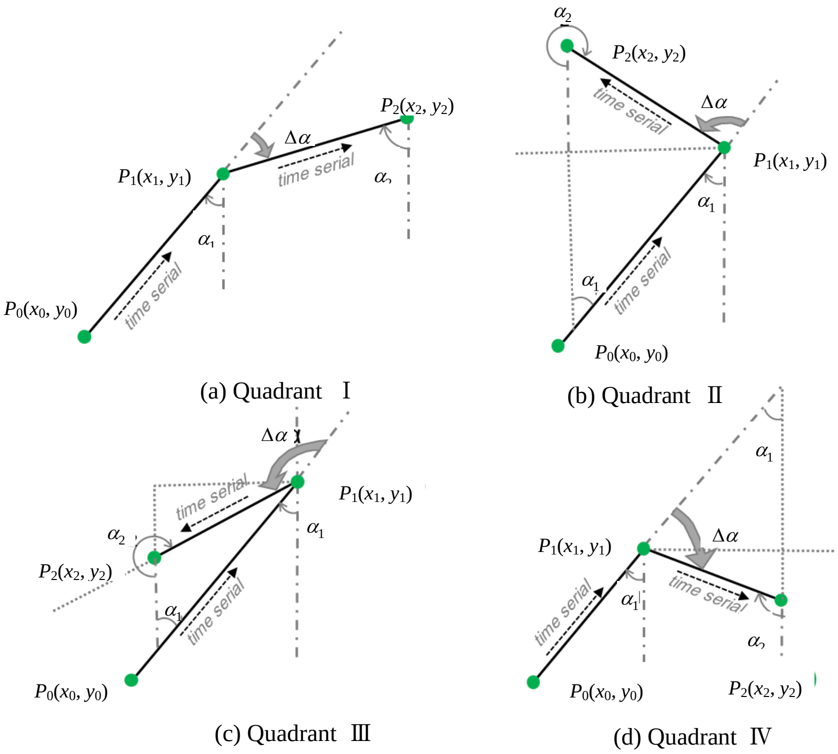


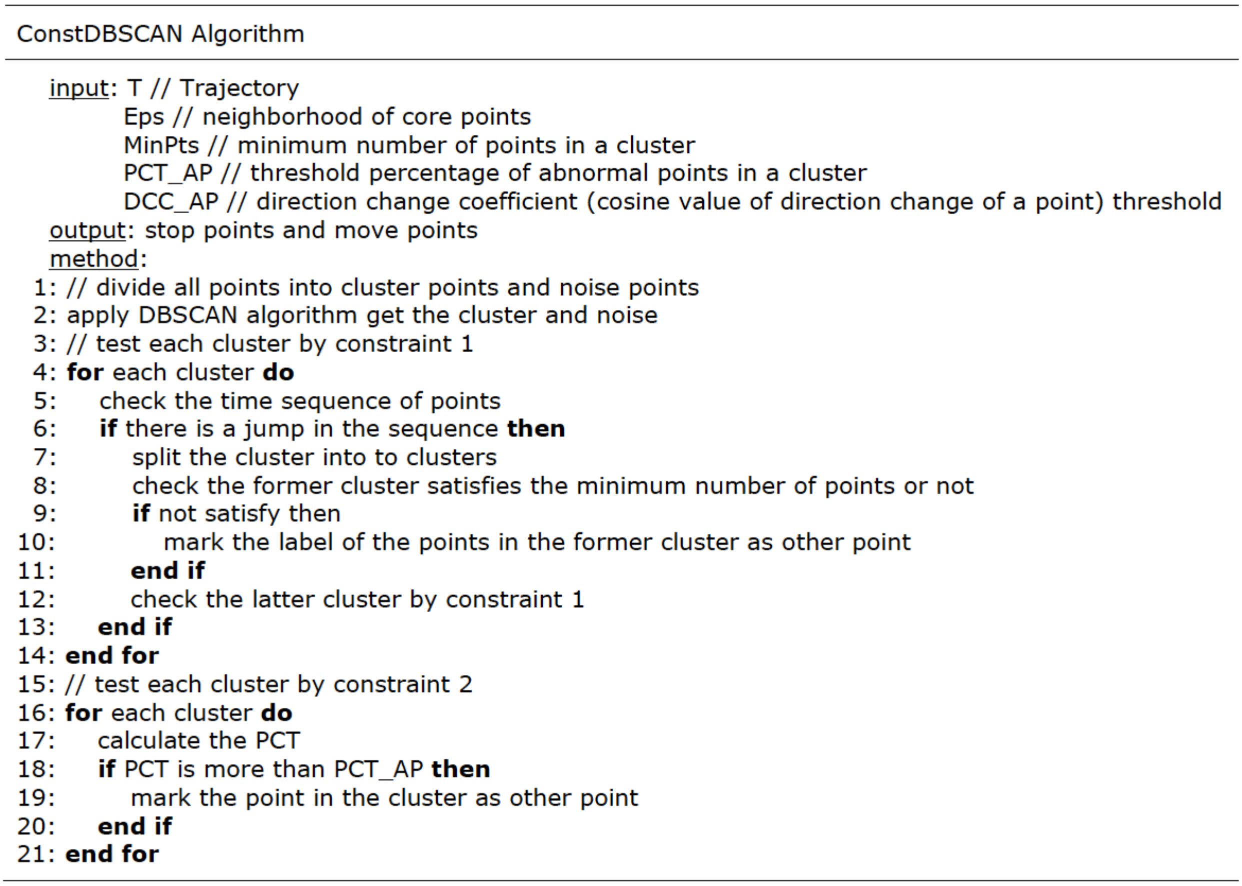
Figure 4.1 Direction of a point and direction change of two points for a second point in different quadrants

If a cluster represents a stop location, the points are scattered around the location and the points should have an even distribution of direction changes. This means that the cosine of direction change (or the **direction change coefficient**, *DCC*) should nearly always differ from 1. Points with a *DCC* value close to 1 probably represent movement of the subject along a straight link in the network. In a cluster, the points should have an even distribution of *DCC*. So **abnormal points** are those points without an even distribution of *DCC*, to be specific, a *DCC* close to 1. Here we use DCC_{AP} denote the approximation to 1.

$$Abnormal\ Point = \{DCC \geq DCC_{AP} | \ Point \in Cluster\} \tag{4.2}$$

4.2.1.3. C-DBSCAN algorithm

The improved DBSCAN algorithm used in this research, C-DBSCAN, is shown in Figure 4.2. Firstly, the DBSCAN algorithm is applied to obtain the cluster points (stop points) and noise points (moving points) in line 2. Then each cluster is tested against constraint 1. Here a new cluster may be split from the older one or the old cluster may be labeled as noise if it does not follow the cluster rule. Finally a cluster satisfying constraint 1 will be tested against constraint 2. Clusters that satisfy both constraints 1 and 2 are marked as stop points; other points are marked as moving points.



```
ConstDBSCAN Algorithm

input: T // Trajectory
       Eps // neighborhood of core points
       MinPts // minimum number of points in a cluster
       PCT_AP // threshold percentage of abnormal points in a cluster
       DCC_AP // direction change coefficient (cosine value of direction change of a point) threshold
output: stop points and move points
method:
// divide all points into cluster points and noise points
apply DBSCAN algorithm get the cluster and noise
// test each cluster by constraint 1
for each cluster do
    check the time sequence of points
    if there is a jump in the sequence then
        split the cluster into to clusters
        check the former cluster satisfies the minimum number of points or not
        if not satisfy then
            mark the label of the points in the former cluster as other point
        end if
        check the latter cluster by constraint 1
    end if
end for
// test each cluster by constraint 2
for each cluster do
    calculate the PCT
    if PCT is more than PCT_AP then
        mark the point in the cluster as other point
    end if
end for
```

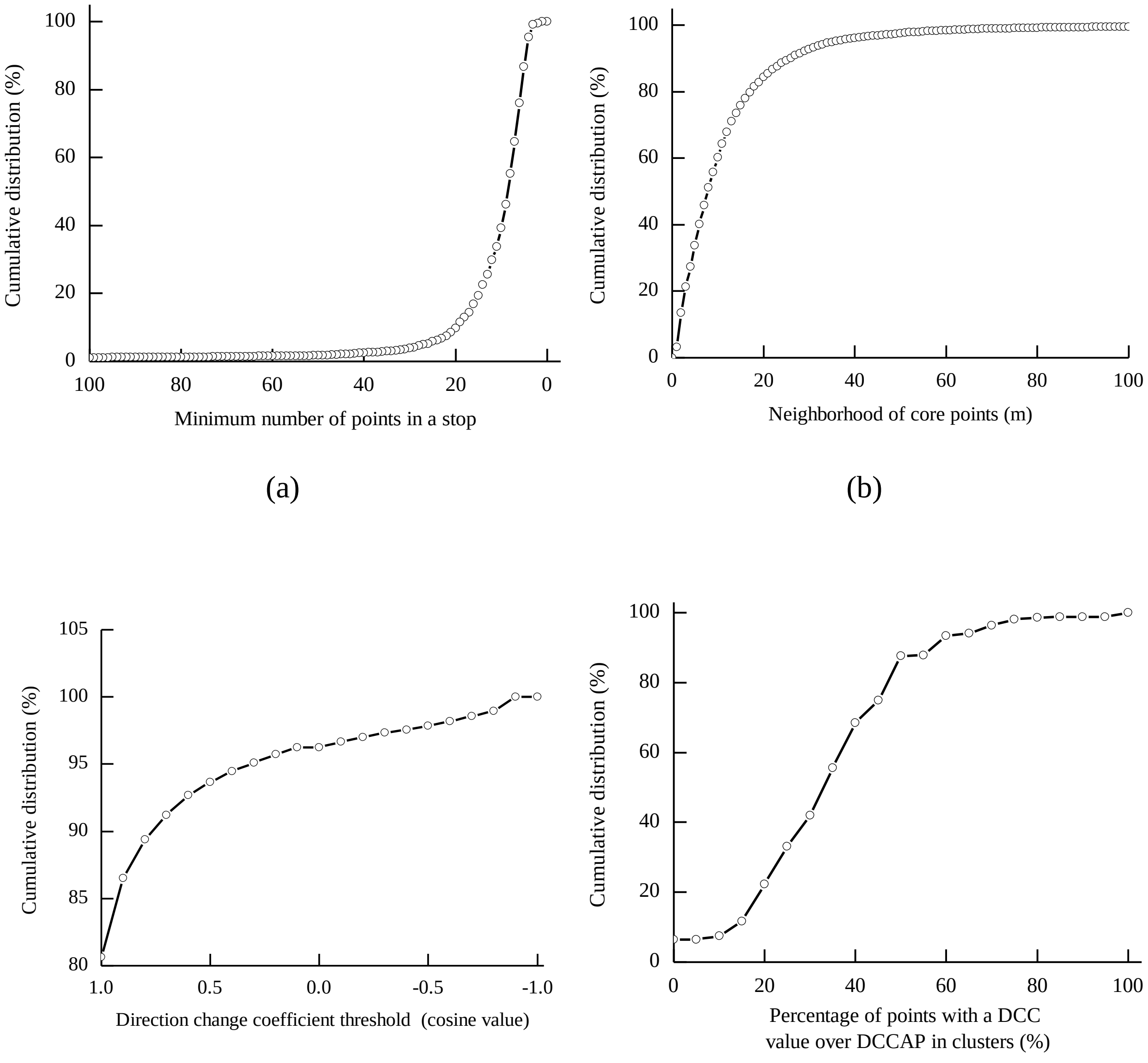
Figure 4.2 C-DBSCAN algorithm

4.2.2. Parameter estimation

In the C-DBSCAN algorithm, there are four parameters which need to be estimated. They are *Eps*, *MinPts*, DCC_{AP} and PCT_{AP}. The cumulative frequency method (at least 90%) is used to estimate these four parameters using the samples in the training data set. The estimation results of these four parameters are shown in Figure 4.3.

Figure 4.3-a demonstrates that if *MinPts* equals four points in the neighborhood, there is 95% probability a stop point is identified and included in a cluster. Figure 4.3-b shows that with the premise that *MinPts* equals four points, if an *Eps* value of less than 25 meters for a cluster candidate means there is a 90% probability that a stop point is included in a cluster. Figure 4.3-c indicates that there is 90% probability that a point with *DCC* value more than 0.8 is a moving point. Figure 4.3-d shows that with the premise that DCC_{AP} is equal to 0.8, a PCT_{AP} value of less than 60% for a cluster candidate means there is a 93% probability that the cluster candidate is a stop. Consequently, we obtain the estimated parameters as follows: *Eps* = 25 meter, *MinPts* = 4, DCC_{AP} = 0.8 and PCT_{AP} = 60%.

Note that demographic information and available modes of transport may influence the thresholds of parameters which need to be estimated. However, in this thesis, the data sets used for training and testing are obtained from the same group of volunteers using several modes of transport, so this influence is not verified in detail.



(a) (b)

(c) (d)

Figure 4.3 Estimation results

4.3. DISTINGUISHING ACTIVITY STOP LOCATION FROM NON-ACTIVITY STOP LOCATION

4.3.1. Support vector machines

SVMs is a supervised machine learning method which can be used for classification or regression analysis. It was developed by Vladimir N. Vapnik and the current standard incarnation (soft margin) was proposed by Cortes and Vapnik in 1993 and published in 1995 (Cortes and Vapnik 1995).

For application to classification, SVMs divides a training dataset with a hyperplane that maximizes the margin between two classes in the first step. In the second step, the learning from this training dataset is applied to the prediction dataset and the classification is implemented. The hyperplane can be linear or non-linear depending on whether the input data is linear or non-linear. In the non-linear case, SVMs use a kernel function to map the points in the dataset linearly separable in the higher dimensions using a non-linear mapping function ϕ. The separating hyperplane in the higher dimensions can be represented by the following formula:

$$\boldsymbol{\omega}^{\boldsymbol{T}}\phi(\boldsymbol{x_i}) + b = 0 \tag{4.3}$$

where $\boldsymbol{\omega}$ is the weight vector normal to the hyperplane, $\boldsymbol{x_i} \in R^n$, i = 1, ... , l are the n-dimensional vectors used for training or that are to be divided into two classes, and b is the intercept associated with decision boundaries.

Since the instance counts of different labels are not balanced in this dataset, assignment of the same cost value to the two classes would cause skewing of the separating hyperplane towards the minority class as a result of this imbalance. In order to avoid suboptimal SVMs models arising because of the imbalance, two different misclassification costs, C^+ and C^-, are assigned in the SVMs models (Chang and Lin 2011) as the following formulas:

$$\min_{\omega,b,\varepsilon} \quad \frac{1}{2}\boldsymbol{\omega}^{\mathbf{T}}\boldsymbol{\omega} + C^+ \sum_{i|y_i=+1}^{l} \xi_i + C^- \sum_{i|y_i=-1}^{l} \xi_i, \tag{4.4}$$

$$\begin{aligned} s.t. \quad & y_i(\boldsymbol{\omega}^{\boldsymbol{T}}\phi(\boldsymbol{x_i}) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, . . l. \end{aligned}$$

where C^+ and C^- are the misclassification cost (or penalty) for the positive class examples and negative class examples, respectively, $\boldsymbol{y} \in R^l$ is an indicator vector such that $y_i \in \{1,-1\}$, and ξ_i is the slack variable.

The dual problem of the situation represented by equations (4.4) is as follows:

$$\min_{\omega,b,\varepsilon} \quad \frac{1}{2}\boldsymbol{\alpha}^{\mathbf{T}}Q\boldsymbol{\alpha} - \boldsymbol{e}^{\mathbf{T}}\boldsymbol{\alpha} \tag{4.5}$$

$$\begin{aligned} s.t. \quad & 0 \leq \alpha_i \leq C^+, if \quad y_i = 1; \\ & 0 \leq \alpha_i \leq C^-, if \quad y_i = -1; \\ & \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{\alpha} = 0\,, \end{aligned}$$

where $\boldsymbol{e}$ = [1,....,1]$^{\mathrm{T}}$ is the vector of ones, $\boldsymbol{Q}$ is an l by l positive semi-definite matrix, $Q_{ij} \equiv y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j})$, and $K(\mathbf{x_i}, \mathbf{x_j}) \equiv \phi(\mathbf{x_i})^T \phi(\mathbf{x_j})$ is the kernel function.

After problem (4.5) is solved, using the primal-dual relationship, the optimal model can be represented as follows:

$$\boldsymbol{\omega} = \sum_{i=1}^{l} y_i \alpha_i \phi(\boldsymbol{x_i}) \tag{4.6}$$

and the decision function is expressed as

$$sgn\left(\boldsymbol{\omega}^{T}\phi(\boldsymbol{x})+b\right)=sgn\left(\sum_{i=1}^{l}y_i\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{4.7}$$

For the kernel function, $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$, a Gaussian kernel is believed to be the most suitable function given our data size and attribute size (Hsu et al 2010). The Gaussian kernel function is shown as follows.

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=e^{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2} \tag{4.8}$$

where σ is the Gaussian parameter and $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the Euclidean distance between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

To implement this SVMs model, the software LibSVM (Chang and Lin 2011) is utilized. LibSVM applies the SVMs to the training dataset and stores the values of $y_i\alpha_i\forall i,b$, label names, support vectors and other information such as kernel parameters in the trained model file for implementing forecasts on the prediction dataset.

4.3.2. Attributes selection

This subsection describes the features in GPS trajectories that are selected for implementing SVMs in order to distinguish activity stops and non-activity stops. Stop duration is the first attribute that comes to mind, because no matter what kind of activity takes place, a certain period of time is the basic requirement. The distribution of stop duration for the two kinds of stops in the training data trajectories is illustrated in Figure 4.4.

This shows that stop duration is an important stop-distinguishing feature of the trajectories. Firstly, a sharp increase in accumulative frequency of activity stop after 300 seconds, which means almost 80% of activity stops have a duration more than 300 seconds, demonstrates that much more activity stops have a much longer duration than non-activity stops. Secondly, it is found that a threshold can be used to distinguish these two kinds of stops. If 105 seconds is taken to be the threshold, almost 92.5% of stops are accurately distinguished. However, this leaves 7.5% of stops with an erroneous classification. Specifically, almost 7.5% of activity stops have a stop duration from 30 seconds to 105 seconds while 7.5% of non-activity stops have a stop duration from 105 seconds to 170 seconds. The reasons for these short activity stops may be that subjects incorrectly turn off the GPS function immediately upon arriving home or at the workplace, or particularly efficient deliveries to customers. Non-activity stops with a longer duration may be caused by long waits at major intersections with longer signal cycles. This overlap in stop duration means that other features extracted from the trajectories are also needed as attribute inputs for the SVMs.

Non-activity stops with a longer duration may be caused by long waits at major intersections. It shows in Figure 4.4 that longer duration of non-activity stop is more than 105 seconds and the percentage of more than 180 seconds is almost zero. It means that non-activity stop in our data set does not include situation of traffic congestion which usually lasts for a long time. For non-activity stop such as waiting at major intersections, unlike activity stops, it does not involve any local walking, as would be the case with an activity stop at home, a workplace or at a convenience store, post office and so on. As a result, GPS points collected during non-activity stops are scattered over a very limited area. This area can be measured by taking the average (mean) distance of each of the scattered GPS points to their common centroid. Figure 4.5 shows the average distance to the centroid for each stop; almost all non-activity stops have an average distance from the centroid of less than 30 meters.

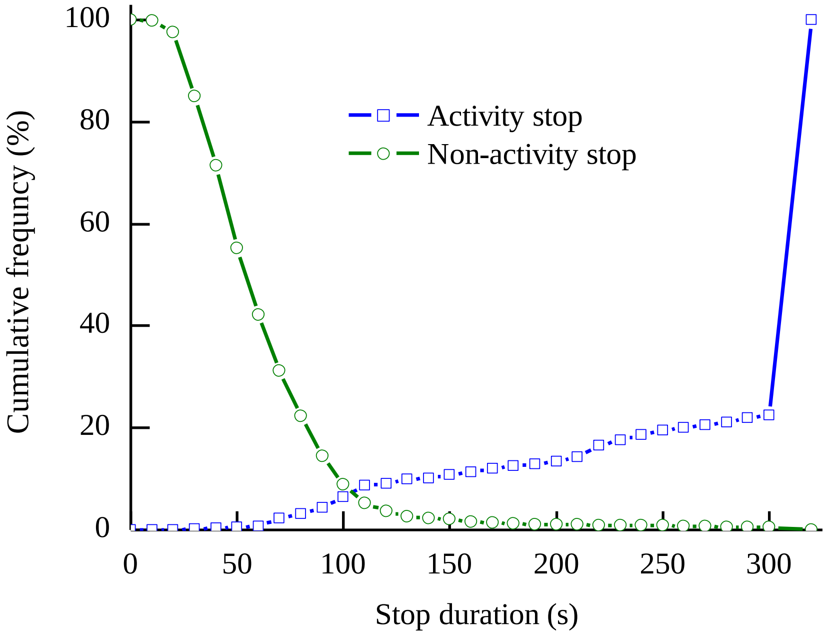


Figure 4.4 Distribution of stop duration for two kinds of stops

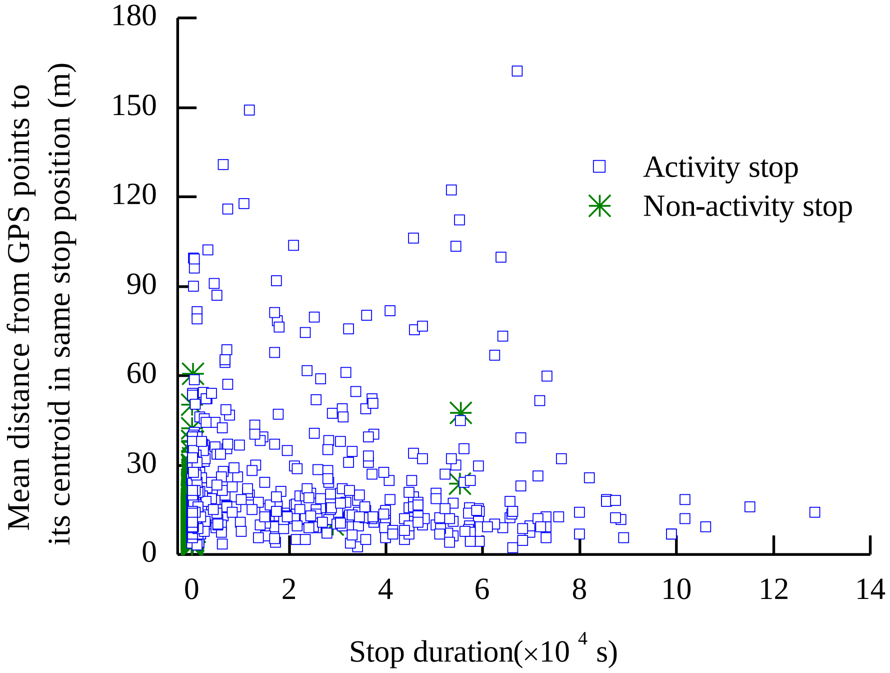


Figure 4.5 Mean distance from GPS points to the common centroid by stop duration

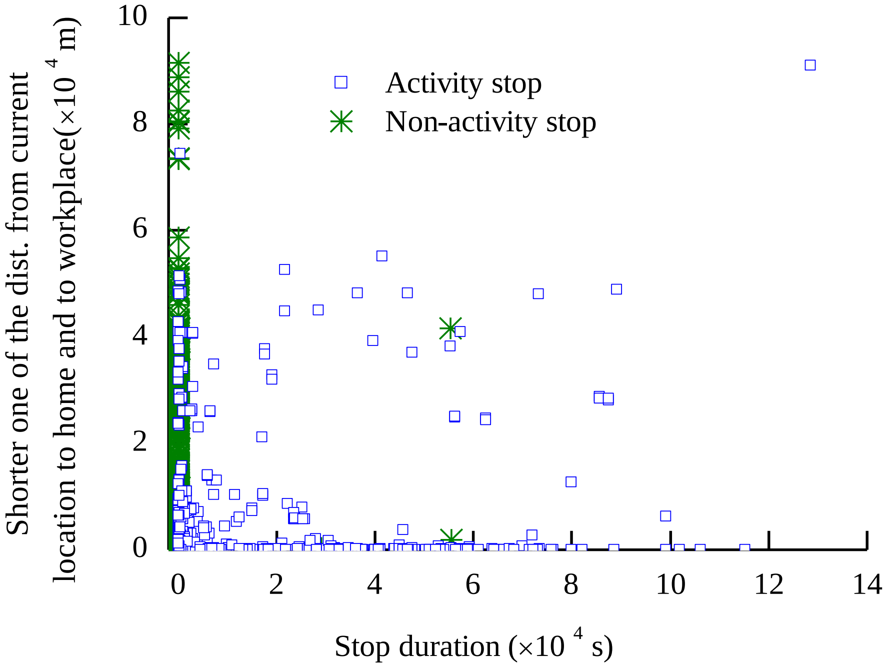


Figure 4.6 Shorter of distances from current location to home and to work place

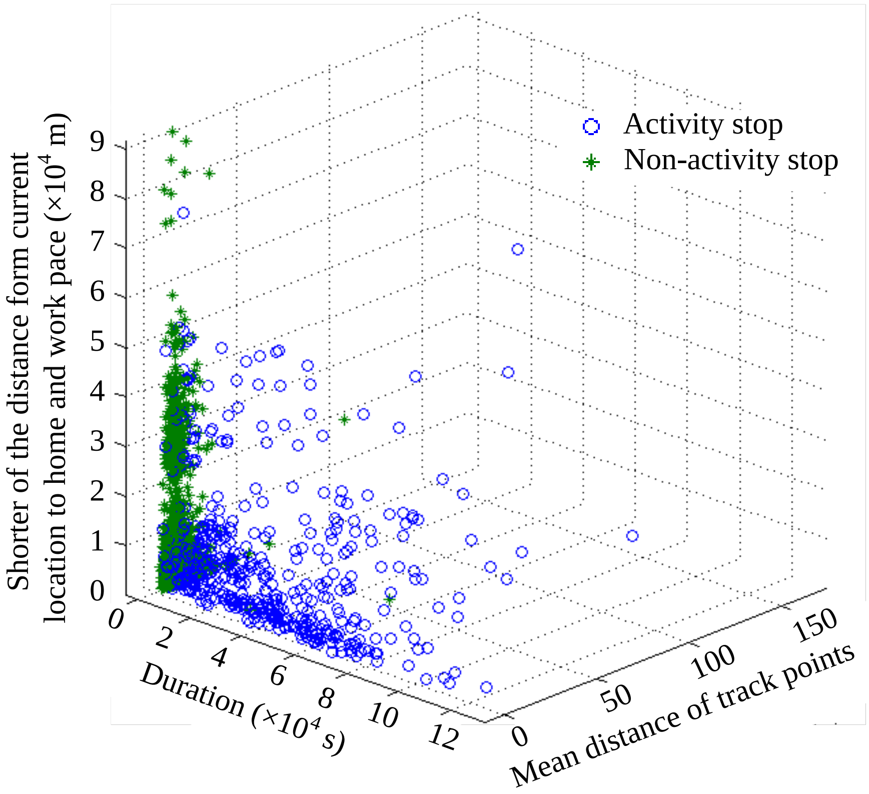


Figure 4.7 Three attributes shown together in three dimensions

Considering the immediate turning off of GPS functions by some subjects, these short-duration activity stops may be identified by the distance between stop location and the home or workplace. In this research, the shorter of these two distances is used for each stop. This distance is plotted against stop duration in Figure 4.6. This demonstrates that activity stops are concentrated within 10km whereas non-activity stops are at a greater distance.

Figure 4.7 plots stop duration, mean distance of GPS points to the cluster centroid and the shorter of the distances from the current location to home and to the workplace by activity stop and non-activity stop in three-dimensional space. It can be concluded from this that most stops can be distinguished by the use of the three features. Consequently, these three features are selected and utilized as input features in the SVMs.

4.3.3. Result analysis

After processing the data for scale, optimal values of the parameters cost, *C*, and gamma, γ are calculated by the grid module in LibSVM for different settings of C^+ and C^-. A "grid-search" using cross-validation is implemented. Various pairs of (C,γ) in increasing order are tested exponentially, a method found to be suitably practical to identify good values of parameters by Hsu et al. (2010). The test values of parameters *C* and γ were as follows: C=2^{-30}, 2^{-29},…, 2^{30}; γ =2^{-30}, 2^{-29},…, 2^{30}. These pairs of (C,γ) were tested in five situations: with C^+=*C, 2C, 3C, 4C, 5C* and C^-=*C*. And a five-fold cross-validation was implemented to avoid the overfitting problem. The results for the tested situations are shown in Table 4.1. The cross-validation accuracy reaches a peak of 95.65% when C=2^{21}, $\gamma=2^{-3}$, $C^{+}=2C$ and $C^{-}=C$.

Table 4.1 SVM results for different (C, γ) pairs for five situations of C^{+} and C^{-}

Situation	C^{+}	C^{-}	C	γ	Accuracy
I	1C	1C	2^{30}	2^{-3}	95.55%
II	2C	1C	2^{21}	2^{-3}	95.65%
III	3C	1C	2^{17}	2^{-2}	95.45%
IV	4C	1C	2^{18}	2^{-3}	95.36%
V	5C	1C	2^{18}	2^{-3}	95.16%

The learning by the SVMs in situation II is then used to test the data set, achieving an accuracy of 94.12%.

This still leaves about 4–6% of stops falsely distinguished as activity stops or non-activity stops. These are the stops with very similar vector values. In future work, these results might be improved by increasing the number of vector dimensions, such as by including the frequency of stops at the same destination, distances to locations on the important address list, GIS information and so on.

4.4. COMPARISON WITH OTHER METHODS

4.4.1. Other methods for comparison

Since the two-step method proposed in this paper is an extension of the DBSCAN algorithm, in this subsection two variants of the DBSCAN method (one in the duration-based category and the other in the density-based category, respectively) are tested using data collected in Nagoya in order to compare accuracy. Here the primary characteristics of these algorithms are listed. Details can be found in the corresponding papers.

The density-based method is called the DJ-cluster algorithm (Zhou et al. 2007), a simplified version of the DBSCAN algorithm. It uses the same concept of a core point as DBSCAN. However, concerning expansion of the cluster, density-reachability and density-connectivity are replaced by the concept of density-joinability. To be specific, instead of using core points to expand a cluster, any shared point is used to combine clusters.

The duration-based method is called the TrajDBSCAN algorithm (Tran et al. 2011). In contrast with the original DBSCAN algorithm, the key feature of TrajDBSCAN is that it defines a temporal linear neighborhood with a core point determined based on a minimum stop time (not a minimum number of points).

4.4.2. Accuracy of calculation and results

The aim here is to show differences in accuracy when applying these methods to the identification of activity stop locations in continuous GPS trajectories. Four indexes are selected and listed below with which to comprehensively describe accuracy. These indexes were calculated for each stop location and Table 4.2 shows the average value of these indexes among all activity stops. Parameter values used in the comparison are the recommended ones or as calculated following the methods in the corresponding papers.

1) Ratio of number of locations identified by method over ground truth.

This index is used to evaluate the redundancy of activity stop location candidates. When equal to one, all identified activity stop location candidates are actual activity stop locations. If greater than 1, then activity stop locations have been erroneously determined as candidates by the method. If less than 1, the method has failed to identify some actual activity stop locations.

2) Average distance between center of identified location and ground truth.

This index is used to evaluate the geographical accuracy of activity stop location candidates. A shorter distance between the center of an identified location and the ground truth location means higher accuracy has been achieved. Here the center of an activity stop location is calculated as the centroid of the stop points indicating the stop location.

3) Percentage of points in the ground truth correctly identified.

This index is used to indicate identification accuracy at the level of GPS points. It shows the percentage of points in the ground truth that are identified by each method. The higher the value of this index, the more actual stop points are correctly identified by the method.

4) Percentage of stop points identified by the method in ground truth.

This index is used to show redundancy at the level of GPS points. It shows the percentage of points identified by each method that really exist in the ground truth. The higher the value of this index, the lower the percentage of useless GPS points in the identified stop locations by the method.

The comparative results shown in Table 4.2 clearly show that the two-step method proposed in this chapter gives generally better performance than the other two DBSCAN variants in identifying activity stop locations in the continuous GPS trajectories.

Table 4.2 Four indexes used to compare the proposed method with other DBSCAN variants

Method	Index 1	Index 2	Index 3	Index 4
DJ-cluster	1.46	8.9m	90%	95%
TrajDBSCAN	6.62	3.4m	93%	94%
C-DBSCAN and SVM	1.02	3.0m	88%	99%

4.5. SUMMARY

In this chapter, a two-step methodology is proposed for identifying activity stops in continuous trajectories utilizing a variation of the DBSCAN algorithm and the SVMs method. In order to adjust DBSCAN to the context of GPS trajectories, two constraints are applied as improvements: a time sequence constraint and a direction change constraint. Three major features are extracted for utilization in the SVMs method: stop duration, mean distance to the centroid of a cluster of points at the stop location, and the shorter of the distances from the current location to home and to the workplace.

Application of this proposed methodology to GPS data collected using mobile phones in Nagoya area of Japan in 2008 demonstrates that the improved DBSCAN algorithm (C-

DBSCAN) achieves an accuracy of 90% in identifying stop locations and the SVMs method is almost 96% accurate in distinguishing activity stops from non-activity stops. Therefore, this two-step method may be suitable for identifying activity stops in continuous GPS trajectories with a higher frequency of data points, especially those that lack feature of "speed" for any reason. In comparison with other similar methods, this two-step procedure demonstrates better performance overall.

With the latest GPS-capable devices, the corresponding data collection interval can be as short as one second, and the information collected also covers speed and acceleration. However, the method advanced in this chapter can certainly be applied to GPS data with more features than were used in this work. On the other hand, this method also offers the possibility of reducing the number of features required when collecting GPS data, thereby reducing memory requirements.

Since the data set used in this study has the feature demonstrating the GPS signal quality which was used to exclude the trajectory point with low signal quality caused by the underground or tunnel condition. One future research trend can be trying to include this part of data with the help of supplement data collected by other types of sensors on the mobile phone. This chapter used three features as independent variables in SVMs and there still be improving space. So another direction of future research could be focusing on trying to increase the dimensionality of the vectors utilized in the SVMs for higher accuracy. Data collected in Nagoya do not include traffic congestion. So for the situations like this, it may be similar to an activity stop and current methods and attributes used in SVMs may not handle it well. But it may also be worthy to try by adding dimensionality of the vectors in the SVMs in the future research.

4.6. REFERENCE

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Cortes, C.and Vapnik, V. (1995). Support-vector networks. *Machine Learning 20 (3)*: 273.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd, Vol. 96,* 226-231.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2010). A Practical Guide to Support Vector Classification, <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf> (Jul. 1, 2014)

Tran, L. H., Nguyen, Q. V. H., Do, N. H., and Yan, Z. (2011). Robust and Hierarchical Stop Discovery in Sparse and Diverse Trajectories (No. EPFL-REPORT-175473).

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS), 25(3)*, 12.

Chapter 5. Activity Type Identification

5.1. INTRODUCTION

After distinguishing activity stops from non-activity stops, the next step is to identify activity types. Many researchers have contributed to this field and tried to achieve better accuracy by advancing and testing many methods. Most of these existing methodologies are rule-based methods, and machine learning methods have been rarely applied in this field (Gong et al. 2014).

However, one disadvantage of rule-based methods is that the rules may vary depending on features of data sets; rules suitable for one data set may lead to low accuracy of another data set. On the contrary, machine learning is believed suitable for classification based on the features and potential algorithms learned from a given training data set. It achieves satisfactory results when applied to transportation mode identification. However, only limited research (Deng and Ji, 2010; Griffin and Huang, 2005; McGowen and McNally, 2007) have applied machine learning methods on activity type identification. What is more, which method is suitable to identify activity type is still unknown. In this chapter, several machine learning methods are tested and compared for identifying activity types from mobile phone GPS data. To be specific, Support Vector machine (SVM), Discriminant Analysis (DA), Classification Trees, and Neural Network (NN) are applied on the mobile phone GPS data collected in the Nagoya metropolitan area in Japan.

The reasons of choosing these four machine learning methods are as follows. 1) Discriminant analysis and decision tree have already been applied for identifying the activity types by other research. 2) Support vector machines, neural network and decision trees have already been applied for travel mode identification and the accuracy varies in the existing research. Some papers (Reddy et al. 2008; Gonzalez et al. 2010; Zhang et al. 2011) achieves a good results while others (Byon et al. 2007; Zheng et al. 2008) do not. It is true that the same method achieves quite different accuracy on different data sets due to the data's quality and independent variables used in the machine learning methods. And these methods can be treated as the candidates of good performance of activity type identification. Consequently, these four methods are chosen for the comparison in this chapter.

5.2. METHODOLOGY

This section introduces the machine learning methods used for test and comparison on activity type identification. Since more than two types of activities need to be identified, machine learning methods here are used for multi-class classification.

5.2.1. Support vector machines

SVM is invented for binary-class classification originally and detailed instruction of this can be found in section 4.3.1. As far as application on multi-class classification, there are two types of methods in SVMs. One is by constructing and combining several binary classifiers while the other is by directly considering all data in one optimization formulation. However, the latter has a limitation that the number of variables should be proportional to the number of classes (Hsu and Lin, 2002). Consequently, in this chapter, only the former is implemented.

Applied for multi-class classification, the former type of SVM can construct and combine several binary classifiers in 3 ways: one-against-one SVM, one-against-rest SVM and DAG (Directed Acyclic Graph) SVM.

5.2.1.1. One-against-rest SVM

One-against-rest SVM constructs *k* SVM classifiers where *k* is the number of classes. As for the *i*-th SVM classifier, items in the *i*-th class will be marked as positive labels while items in all other class will be marked as negative labels. The *i*-th SVM classifier solves the following problem (Hsu and Lin, 2002).

$$\min_{\omega^i, b^i, \xi^i} \quad \frac{1}{2}\left(\omega^i\right)^T \omega^i + C\sum_{j=1}^{l} \xi_j^i \tag{5.1}$$

$$\left(\omega^i\right)^T \phi\left(\boldsymbol{x}_j\right) + b^i \geq 1 - \xi_j^i, if\ y_j = i,$$

$$\left(\omega^i\right)^T \phi\left(\boldsymbol{x}_j\right) + b^i \leq -1 + \xi_j^i, if\ y_j \neq i,$$

$$\xi_j^i \geq 0, j = 1, \dots, l,$$

where l is the sample size of data set; for each item $(\boldsymbol{x_i}, y_i)$, $y_i \in \{1, \dots, k\}$ is the class label of $\boldsymbol{x_i}$;. ϕ is the function mapping data $\boldsymbol{x_i}$ to a higher dimensional space; C is the misclassification cost parameter (or penalty parameter); b is the intercept associated with decision boundaries; ξ_j is slack variables.

After solving problem (5.1), there are k decision functions as follows.

$$(\omega^1)^T \phi(\boldsymbol{x}) + b^1,$$

$$\vdots$$

$$(\omega^k)^T \phi(\boldsymbol{x}) + b^k.$$

Then x is assigned in the class which has the largest value of the decision function.

$$\text{class of } x \equiv\ argmax_{i=1,\dots,k}\left(\omega^i\right)^T \phi(x) + b^i \tag{5.2}$$

In practice, the dual problem of (5.1) is solved and it has the same number of variables as the number of data in (5.1). Consequently, *k l*-variable quadratic programming problems are solved.

5.2.1.2. One-against-one SVM

The second is called one-against-one SVM which constructs *k(k-1)/2* binary-class classifiers where k is the number of classes. Each classifier is trained by arbitrarily selected two classes until each class is trained with any other class. For training data from the i-th class and j-th class, the following binary classification problem should be solved (Hsu and Lin, 2002; Knerr et al. 1990).

$$\min_{\omega^{ij},b^{ij},\xi^{ij}} \quad \frac{1}{2}\left(\omega^{ij}\right)^T \omega^{ij} + C\sum_t \xi_t^{ij} \tag{5.3}$$

$$\left(\omega^{ij}\right)^T \phi(\boldsymbol{x_t}) + b^{ij} \geq 1 - \xi_t^{ij}, if\ y_t = i,$$

$$\left(\omega^{ij}\right)^T \phi(\boldsymbol{x}_t) + b^{ij} \leq -1 + \xi_t^{ij}, if\ y_t = j,$$

$$\xi_t^{ij} \geq 0.$$

Then a voting strategy, called "Max Wins" is used to decide which class the item should be assigned in. If sign $\left(\left(\omega^{ij}\right)^T \phi(\boldsymbol{x}_t) + b^{ij}\right)$ says $\boldsymbol{x}$ is in the i-th class, then the vote for the i-th class is added by one. Otherwise, the j-th is added by one. Finally, $\boldsymbol{x}$ is assigned into the class with the largest vote result.

5.2.1.3. DAG SVM

The third one is DAG SVM. It has the same training phase as one-against-one SVM by solving *k(k-1)/2* binary SVM classifiers. However, the testing phrase is replaced by using a rooted binary directed acyclic graph which has *k(k-1)/2* internal nodes and k leaves. Each node is a binary SVM of i-th and j-th classes. It starts from the root node and moves to either one of the leave depending on the result of output value of binary decision function of upper level. An example of DAG classification for three classes is shown in Figure 5.1. $\bar{i}$ represents the rejection that x belongs to class i. At the top level of classification, any pair of classes can be selected and either class is rejected. Then it moves to the left or right branch depending on the result. Finally, it reaches to one of the leaves representing the predicted class.

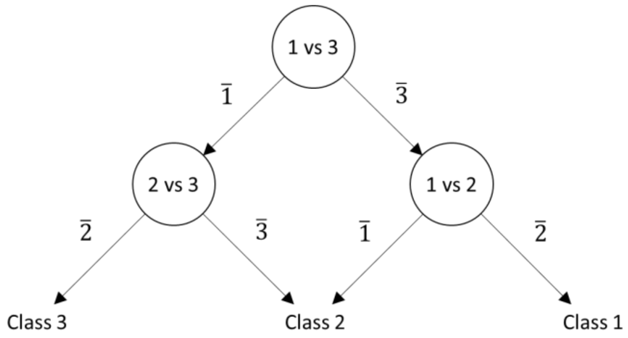


Figure 5.1 DAG classification for 3 classes

5.2.2. Classification trees

Classification tree is a method based on a series of rules in an optimal order. Rules are tested in the form of a chain of simple questions. After replying to the question at each split node, the answer decides the direction of which child node will be chosen as the next split node. To be more specific, a classification tree predicts a category of an item by following a series of decisions in the tree from the root (beginning) node down to a leaf node (the final child node). Leaves represent class labels and branches represent conjunctions of features that lead to those class labels. These questions-and-answers are actually a set of logical if-then conditions for classifying items into corresponding categories.

A basic way to build a classification tree from labeled examples proceeds in a greedy manner: the most informative questions are asked as nearer the root as possible in the hierarchy. In a greedy algorithm, the first question is designed to get the two children subsets consisting of sample cases of the same class in each subset as purely as possible. After the first subdivision is done, one proceeds in a recursive manner, by using the same method for the left and right children sets, designing the appropriate questions, and so on and so forth until the remaining sets are sufficiently pure to stop the recursion. It is the time when the leaves are reached after the chain of questions descending from the root, and then the final remaining set in the leaf should be almost pure, i.e., consisting of sample cases of the same class. (Battiti and Brunato, 2014)

Algorithms for constructing trees usually work top-down, by choosing a variable at each step that best splits the set of items. Different algorithms use different metrics for measuring “best” or purity. In this research, Gini Impurity (GI) is used to measure “best”. GI (Battiti and Brunato, 2014) is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. It is computed as the expected value of the mistake probability; as usual, the expectation is given by adding, for each class, the probability of mistaking the classification of an item in that class (i.e., the probability of assigning it to any class but the correct one: $1 - p_i$) times the probability for an item to be in that class (p_i). Suppose that there are m classes, and let f_i be the fraction of items labeled with value i in the set. Then, by estimating probabilities with frequencies ($p_i \approx f_i$):

$$\mathrm{GI(f)} = \sum_{i=1}^{m} f_i\,(1 - f_i) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2 \tag{5.4}$$

GI reaches its minimum (zero) when all cases in the node fall into a single target category; otherwise GI is positive.

5.2.3. Neural networks

Neural network is believed to be capable of handling nonlinear pattern recognition but it is difficult to interpret. A neural network consists of three layers: input layer, hidden layer and output layer. With the known correct output values, the network can learn the unrevealed patterns between input and output values by cycles of training in the hidden layers. Then this trained network can be used to predict the output with new input values.

Backpropagation (BP) neural network is a popular kind of neural network used in practice and it is usually considered to be a supervised machine learning methodology. Backpropagation means "backward propagation errors". It is usually used in conjunction with an optimization method such as gradient descent. In an attempt to minimize the loss function; on each of training iteration, the current gradient of a loss function with respects to all the calculated weights in the network is evaluated and then the gradient is fed to the optimization method which employs it to update the weights (Sayed and Baker, 2015).

Before training the network, the topological structure of the network needed to be designed. The topological structure includes number of hidden layers and number of neurons in each hidden layer. There is a consensus that the performance difference from adding additional hidden layers: the situations in which performance improves with a second (or third, etc.) hidden layer are very small (Hassen, 2013). One hidden layer is sufficient for the large majority of problems. Regarding the number of points in hidden layer, there are several formulas (Sheela and Deepa, 2013; Stathakis, 2009) or rules (Karsoliya, 2012) for calculation but they generate results in a big difference. In this paper, the designed structure of neural network contains one hidden layer and the number of neuron points in the hidden layer is tested from 1 to 100 (this range can satisfy almost all the formulas or rules) in order to reach a satisfactory accuracy. The number of points in the hidden layer achieving the highest accuracy will be used for training the network and prediction. Finally we obtained that 31 neuron points in the hidden layer can achieve the best accuracy of 89.2%. The neural network structure utilized in this research is shown in Figure 5.2.

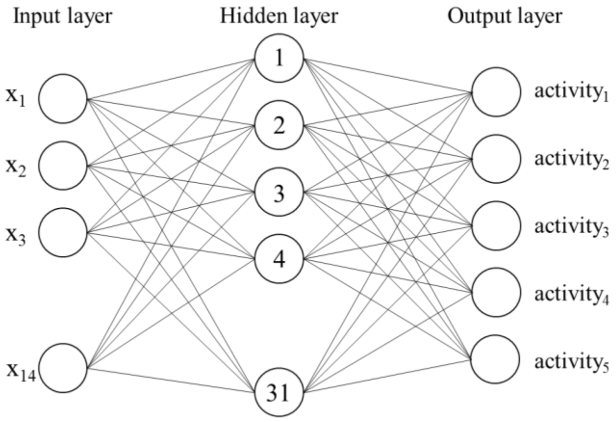


Figure 5.2 Neural network for identifying activity types

5.2.4. Discriminant analysis

DA is usually used in statistics, pattern recognition and machine learning to characterize or separate two or more classes of items or examples by a combination of attributions. DA is closely related to multivariate analysis of variance (MANOVA) and shares the assumptions of MANOVA (Poulsen and French 2008). One of the vital assumptions is related to normal distribution: it is assumed that the data represent a multivariate normal distribution; different classes generate data based on different normal distributions. However, violations of the normality assumption are not "fatal" and the resultant significance test is still reliable as long as non-normality is caused by skewness and not outliers (Tabachnick and Fidell, 1996).

DA can be linear or quadratic depending on variance-covariance matrices are homogeneous or heterogeneous. That is, linear DA requires different classes share the same covariance structure (homogeneous) while quadratic DA does not have this constraint. Based on the features of data set in our research, quadratic DA is used in this paper.

DA creates a classifier which will minimize the possibility of misclassifying cases into their respective groups or categories. The classifier is trained by estimating parameters of a normal distribution for each class. Then for predicting the classes of new data, the trained classifier finds the class with the smallest misclassification cost.

$$\hat{y} = arg \min_{y=1,\dots,K} \sum_{k=1}^{K} \hat{P}(k|\boldsymbol{x})C(y|k) \tag{5.5}$$

where $\hat{y}$ is the predicted classification; K is the number of classes; $\hat{P}(k|\boldsymbol{x})$ is the posterior probability of class k for observation $\boldsymbol{x}$; $C(y|k)$ is the cost of classifying an observation as y when its true class is k, and $\hat{P}(k|\boldsymbol{x})$ is calculated as follows.

$$\hat{P}(k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|k)P(k)}{P(\boldsymbol{x})} \tag{5.6}$$

$$P(\boldsymbol{x}|k) = \frac{1}{2\pi|\Sigma_k|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_K^{-1}(x-\mu_k)\right) \tag{5.7}$$

where $|\Sigma_K|$ is the determinant of Σ_K, and Σ_K^{-1} is the inverse matrix.

5.3. ATTRIBUTE SELECTION

Variables related to activity and trip which first come to mind are used as input variables. Besides, demographic information of volunteers and land use of trip end may also have influence on activity types. However, this data set does not own detailed land use information by now, so only demographic information is added as input variables with activity and trip related variables. Variables from these three dimensions are interpreted in detail as follows.

The first dimension is related to activity features, including activity duration, periods of activity start and end, distance from activity location to home and work place, day characteristics of activity start and end; the second dimension is related to trip features and time cost of trip before the current activity is used; the third dimension is related to demographic information of volunteers, including gender, age, annual household income, frequency of using auto and public transit. Actually driving license is also included in the questionnaire; however, since every volunteer owns it, this variable was finally eliminated from the independent variables for machine learning application. Specification of the selected attributes can be found in Table 5.1.

Table 5.1 Attribute selected for analysis

Attribute name	Description
Activity dimension	
Duration (C)	Time of the activity lasts
Period of activity start	One of the seven period when activity starts
Period of activity end	One of the seven period when activity ends
Distance to home (C)	Distance from trip end to participant’s home
Distance to work place (C)	Distance from trip end to participant’s work place
The day when activity starts	The day type when activity starts, dummy variables: 1 for weekday, 0 for weekend
The day when activity ends	The day type when activity ends, dummy variables: 1 for weekday, 0 for weekend
[Previous activity type]	Type of previous activity
[Next activity type]	Type of next activity
Trip dimension	
Trip duration (C)	Time cost of the trip spent to before current activity
[Travel mode before activity]	Travel mode of trip before current activity
[Travel mode after activity]	Travel mode of trip after current activity
Demographic dimension	
Gender	Dummy variable, male and female
Age	One of the eight age intervals
Annual household income	One of the ten income intervals
Occupation	One of the eight occupation categories
Auto usage frequency	One of the five frequency types
Public transit usage frequency	One of the five frequency types

Note:
1) Attribute name with a “(C)” means continuous variable; others without it means categorical variables.
2) 7 categories of period are periods: NP-b-MP: non-peak before morning peak; MP: morning peak; NP-b-MP-NOON: non-peak between morning peak and noon; Noon; NP-b-NOON-EP: non-peak between noon and evening peak; EP: evening peak; NP-a-EP: non-peak after evening peak. Peak periods and noon period are defined as follows: morning peak starts from 7am to 9am on weekdays and weekends; evening peak starts from 4pm to 7pm on weekdays and 3pm to 6pm on weekends. Noon period starts from 12pm to 1pm on weekdays and weekends. Morning peak and evening peak are equivalent to those in the report of current traffic characteristics in Chukyo Area in Japan.
3) 8 categories of age are as follows: 0~10 years old; 11~20 years old; 21~30 years old; 31~40 years old; 41~50 years old; 51~60 years old;

61~70 years old; 71~80 years old; more than 80 years old.
4) 10 categories of annual household income are following intervals: [0,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8), [8,10), [10,15), [15,+∞); unit is million Japanese Yen.
5) 8 categories of occupation include: employee; self-employer or management; part time or freelance; government/school related; student; housewife; other; no occupation.
6) 5 categories of auto / public transit usage frequency cover: more than 5 days per week; 3~4 days per week; 1~2 days per week; 2~3 times per month; less than once per month.
7) Attributes in [ ] are only used for performance comparison of 3 types of SVMs. Involvement of previous activity type and next activity type is trying to make the accuracy as high as possible; for practical calculation, these two should not be involved because of the endogeneity.

5.4. RESULTS AND DISCUSSION

5.4.1. Comparisons between different types of SVMs

5.4.1.1. Scenario settings

In the performance comparison, 7 scenarios are designed for checking the variance of results when using different combinations of variables from three dimensions. Scenario 1 uses independent variables from all three dimensions. Scenario 2, scenario 3 and scenario 4 use any two dimensions of variables. Other scenarios use only one dimension of variables. Detailed description of each scenario can be found in Table 5.2.

Table 5.2 Scenarios settings

Scenarios	Explanatory Variables Used
Scenario 1	All three dimensions
Scenario 2	Activity dimension and trip dimension
Scenario 3	Trip dimension and demographic dimension
Scenario 4	Activity dimension and demographic dimension
Scenario 5	Activity dimension only
Scenario 6	Trip dimension only
Scenario 7	Demographic dimension only

Each scenario is experimented with different pairs of parameters of cost in optimization function and gamma in kernel function in the following way: from -1 to 30 with a step of 1 for cost and from -30 to 1 with a step of -1 for gamma. Consequently, in each scenario, there are 1024 times of calculation for different pairs of cost and gamma tested.

As far as tools calculating SVM algorithms are concerned, we use LibSVM 3.20 (Chang et al. 2011) to implement one-against-one SVM and DAG SVM (for DAG SVM, some revise is needed, Hsu et al. 2002); for one-against-other SVM, Matlab 2012b is used as the platform for the implement.

5.4.1.2. Results

The highest value of cross validation accuracy with corresponding values of cost and gamma for each scenario are summarized in Table 5.3.

Table 5.3 Highest cross validation accuracy in each scenario

SVM type	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7
one-against-one SVM	*85.47*	*85.99*	55.88	*85.99*	86.30	49.09	43.75
(cost, gamma)	$(2^{22},2^{-15})$	$(2^{23},2^{-15})$	$(2^{7},2^{-4})$	$(2^{19},2^{-15})$	$(2^{21},2^{-14})$	$(2^{18},2^{-2})$	$(2^{0},2^{-5})$
one-against-rest SVM	*87.51*	*87.03*	56.56	87.82	*87.40*	47.05	39.73
(cost, gamma)	$(2^{23},2^{-16})$	$(2^{21},2^{-14})$	$(2^{19},2^{-11})$	$(2^{24},2^{-15})$	$(2^{24},2^{-15})$	$(2^{21},2^{-10})$	$(2^{21},2^{-30})$
DAG SVM	*84.84*	*84.47*	55.99	*85.10*	85.52	48.98	43.75
(cost, gamma)	$(2^{11},2^{-5})$	$(2^{19},2^{-13})$	$(2^{21},2^{-5})$	$(2^{19},2^{-15})$	$(2^{22},2^{-15})$	$(2^{19},2^{-3})$	$(2^{-1},2^{-4})$

Note: figure in red bold underline is the most accurate among scenarios; figure in green bold italic is that close to the highest accuracy.

As far as comparison among different kinds of SVM, one-against-rest shows a higher accuracy compared to one-against-one SVM and DAG SVM in most of the scenarios. One-against-other SVM is defeated by the other two SVMs only in scenario 6 and scenario 7; from scenario 1 to scenario 5, one-against-other SVM demonstrates up to 3% higher accuracy to the other two. However, one-against-other SVM is very time-consuming and is about ten times of running time of the other two. Since the three kinds of SVM were tested on two platforms, platform difference may contribute to the time difference to some extent. But it is believed that the main reason for variance in time consumption is the difference among the combination structures of these SVM methods. This result is consistent with the conclusion when using these SVM on other data sets for classification (Hsu and Lin, 2002).

Overall, there seems no big difference in accuracy of these three kinds of SVM. However, considering the time consumed for calculation, one-against-one SVM and DAG SVM are more suitable for data set with samples in large quantity. Since in our data set, the sample size is not so many, one-against-rest SVM is used in the comparison with classification tree, neural network and discriminant analysis.

5.4.2. General accuracy and time cost of machine learning methods comparison

Matlab 2014a is used as the platform of implementing empirical analysis of the four machine learning methods above and got the accuracy results and time elapsed for comparison. Regarding the time cost of calculation, in order to avoid the effect from hardware, these four methods are calculated one by one on the same computer and there is no other applications kept running at the same time.

Table 5.4 shows the general accuracy and time cost of each machine learning method utilized in this research. Figure in bold italic means the highest accuracy or largest accuracy difference among methods or least time cost; while figure in bold underlined, on the contrary, means the lowest accuracy or smallest accuracy difference among methods or most time cost. The accuracy difference, defined as the accuracy difference between test set and training set, is a reflection of transferability of the method. A method with larger accuracy difference means it can achieve a more stable and transferrable parameters from the training set to other sets.

Table 5.4 General accuracy and time cost of each method

Method	Accuracy			Time (seconds)		
	Training data	**Predicting data**	**Difference**	**Optimizing Parameter**	**Training**	**Predicting**
One-against-rest SVM	91.8%	84.3%	**-7.5%**	**21249.717**	**36.302**	**0.044**
Neural Network	87.8%	**80.5%**	-7.3%	50.243	0.597	0.021
Discriminant Analysis	**86.0%**	82.8%	***-3.2%***	---	0.319	0.024
Classification Tree	***96.7%***	***89.2%***	**-7.5%**	---	***0.301***	***0.003***

Classification tree obtains the highest general accuracy of 96.7% for training on training set and 89.2% for predicting on test set. One-against-rest SVM achieves the second best general accuracy followed by neural network and discriminant analysis. Discriminant analysis obtains the largest accuracy difference which shows the accuracy of predicting is only 3.2% lower than the training; the other three methods make similar accuracy difference of 7.3% ~ 7.5%.

As far as time cost is concerned, among these four methods, classification tree uses the least time 0.301 seconds for training on training set and 0.003 seconds for predicting on test set. The most time consuming method is one-against-rest SVM, which needs 36.302 seconds for training and 0.044 seconds for predicting. Time cost of neural network and discriminant analysis are between the time costs of these two. One disadvantage of one-against-rest SVM and neural network is the process of obtaining the optimal parameter value which needs much more time. Especially the one-against-rest SVM takes nearly 6 hours for optimal parameter calculation.

From the point of view of general accuracy and time cost, it seems that classification tree is the most appropriate method in these four machine learning methods for identifying activity type from our data set.

5.4.3. Accuracy for each activity type among methods

Table 5.5 shows the accuracy of identifying each kind of activity type in training set and test set as well as the accuracy difference when applying each method between these two sets. Figures in bold italic and underlined have the same meanings as in Table 5.4.

Table 5.5 Accuracy of activity type identification by each method

unit: %

Methods	Activity type	Business	Recreation	Back home	Others	Work / school
One-against-rest SVM	**Training set**	97.1	77.7	99.4	48.4	93.1
	Test set	***95.2***	57.7	97.8	10.7	85.9
	Difference	-1.9	-20.0	**-1.5**	-37.7	-7.1
Neural Network	**Training set**	96.0	74.3	**94.0**	**37.1**	**87.3**
	Test set	92.0	62.8	**93.5**	**7.1**	**76.8**
	Difference	-4.0	-11.5	-0.6	-30.0	**-10.6**
Discriminant Analysis	**Training set**	**91.1**	**37.4**	***100.0***	51.6	98.2
	Test set	89.9	**34.6**	99.3	25.0	95.1
	Difference	***-1.2***	***-2.8***	-0.7	***-26.6***	-3.1

Classification Tree	**Training set**	***97.3***	***92.2***	***100.0***	***71.0***	***100.0***
	Test set	**89.4**	***70.5***	***100.0***	***32.1***	***100.0***
	Difference	**-8.0**	**-21.7**	***0.0***	**-38.8**	***0.0***

For “business” activity, all these methods can achieve a high accuracy around 90%. Although classification tree achieves the highest accuracy on training set, also gets the smallest accuracy difference between test set and training set. However, one-against-rest SVM achieves the highest accuracy on test set and the accuracy difference is only -1.9%. So it seems that one-against-rest SVM is the most capable of identifying “business” activity.

Regarding “recreation” activity, none of these four methods can achieve a satisfactory accuracy on both training set and test set. Classification tree achieves the highest accuracy on both training set and test set, but the accuracy difference is also the smallest one. On the contrary, discriminant analysis achieves the largest accuracy difference but with the lowest accuracy on training set and test set. So it seems classification tree is the most appropriate for identifying “recreation” activity among the four methods.

For “back home” activity, all these methods can achieve a high accuracy over 93% no matter on training set or test set. Classification tree can get the highest accuracy and biggest accuracy difference which means it is the most appropriate method for identifying “home” activity.

As far as “others” activity is concerned, the result is similar to “recreation” activity; none of these four methods can achieve a satisfactory accuracy on training set and test set. Compared to the other three, classification tree can achieve a higher accuracy on both training set and test set. Although the accuracy difference is the smallest, classification tree is still the most appropriate for identifying “others” activity.

Concerning “work-or-school” activity, classification tree and discriminant analysis perform much better than the other two. Especially classification tree reaches 100% accuracy on both training set and test set.

For more detailed information, ratios of each activity type identified correctly or as other types by each method, which can be found in the confusion matrix, are shown in Table 5.6.

Table 5.6 Confusion matrix of each method applied on training set and test set

unit: %

Methods	**Training set**						**Test set**					
	Truth/predicted	***business***	***recreation***	***home***	***other***	***work or school***	***Truth/predicted***	***business***	***recreation***	***home***	***other***	***work or school***
One-against-rest SVM	***business***	97.1	2.0	0.4	0.0	0.4	***business***	95.2	3.2	0.0	0.5	1.1
	recreation	16.2	77.7	2.2	1.7	2.2	***recreation***	30.8	57.7	1.3	2.6	7.7
	home	0.0	0.6	99.4	0.0	0.0	***home***	0.7	1.4	97.8	0.0	0.0
	other	14.5	30.6	1.6	48.4	4.8	***other***	21.4	57.1	0.0	10.7	10.7
	work or school	6.0	0.9	0.0	0.0	93.1	***work or school***	12.0	2.1	0.0	0.0	85.9
Neural Network	***business***	96.0	1.3	0.4	0.0	2.2	***business***	92.0	4.3	0.0	0.0	3.7
	recreation	20.1	74.3	1.7	1.7	2.2	***recreation***	21.8	62.8	3.8	2.6	9.0
	home	3.4	2.2	94.0	0.0	0.3	***home***	4.3	0.7	93.5	0.7	0.7
	other	38.7	17.7	0.0	37.1	6.5	***other***	25.0	46.4	14.3	7.1	7.1
	work or school	10.6	2.1	0.0	0.0	87.3	***work or school***	14.1	3.5	4.9	0.7	76.8
Discriminant Analysis	***business***	91.1	3.6	0.0	4.5	0.9	***business***	89.9	4.3	0.0	5.9	0.0
	recreation	49.7	37.4	0.0	12.8	0.0	***recreation***	48.7	34.6	0.0	16.7	0.0
	home	0.0	0.0	100.0	0.0	0.0	***home***	0.7	0.0	99.3	0.0	0.0
	other	22.6	25.8	0.0	51.6	0.0	***other***	28.6	46.4	0.0	25.0	0.0
	work or school	0.0	0.3	0.0	1.5	98.2	***work or school***	0.7	3.5	0.0	0.7	95.1
Classification Tree	***business***	97.3	1.6	0.4	0.4	0.2	***business***	89.4	7.4	0.0	3.2	0.0
	recreation	5.0	92.2	0.0	2.8	0.0	***recreation***	19.2	70.5	0.0	10.3	0.0
	home	0.0	0.0	100.0	0.0	0.0	***home***	0.0	0.0	100.0	0.0	0.0
	other	11.3	17.7	0.0	71.0	0.0	***other***	25.0	42.9	0.0	32.1	0.0
	work or school	0.0	0.0	0.0	0.0	100.0	***work or school***	0.0	0.0	0.0	0.0	100.0

In a summary, these four methods are all good at identifying activity of “business”, “home” and “work-or-school” with a satisfactory; but none of them is capable of identifying “recreation” and “others” with a satisfactory accuracy. From the view point of the exact accuracy value and transferability of the method, it seems that classification tree is the most appropriate method for identifying activity type among these four when being applied on Nagoya data set.

The optimally estimated structure of classification tree is shown in Figure 5.3 (for sake of simplification, just show and explain the first three levels). A brief explanation of the first three levels is given as follows. It starts a question of the current distance to the work place at the root (first level of parent node, node 1), if it is less than 55 meters, it directs to the left child node, node 2; else it directs to the right child node, node 3. At the second level, node 2 and node 3 become parent nodes and split again based on the question of the distance to home. At node 2, if the distance is less than 1535 meters, then the activity type is business, else is work/school. At node 3, if the distance is less than 54 meters, then current activity type might be home directing to the new left child node, node 6; else it directs to a new right child node, node 7 and needs further questions in the next level. In a recursive manner, the parent node splits into child nodes, and the child node is treated as parent node for the next split in the lower level based on questions until all the activity type can be identified.

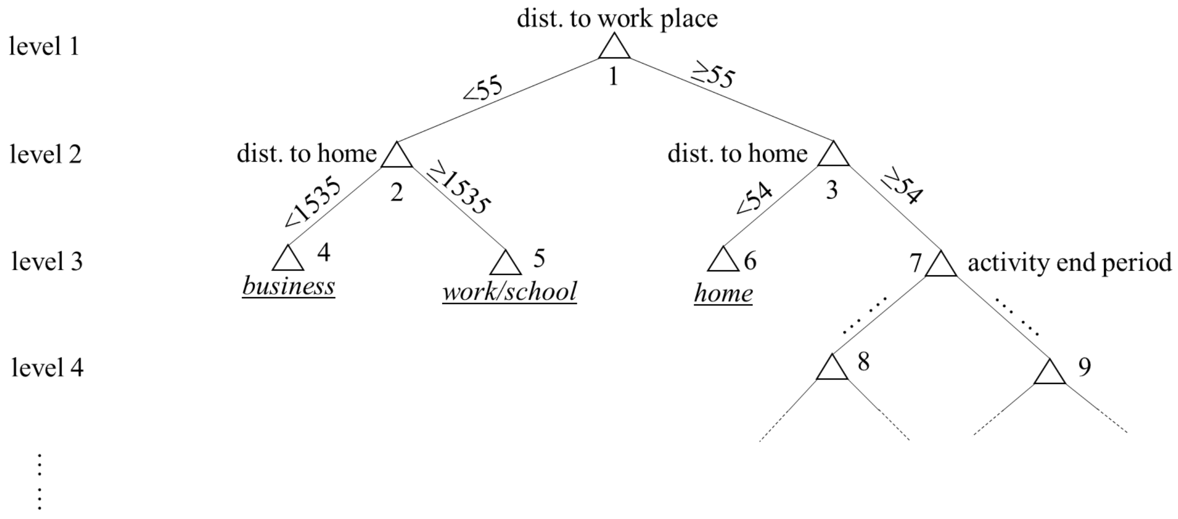


Figure 5.3 Estimated structure of classification tree with data in Nagoya

5.4.4. Discussion on low accuracy of certain activity types

From the results above, it is clear that none of these four methods are capable of identifying activity of “recreation” and “others” with a high accuracy. Low accuracy may be explained from the following two perspectives.

1) Homogeneity among activities to some extent.

Table 5.6 demonstrates more detailed results in the form of confusion matrix where the ratio of one kind of activity is identified as each kind of activity. It shows that “recreation” activity is easily misidentified as “business” activity while “others” activity is easily misidentified as “recreation” or “business” activity. However, “business” activity is not easily misidentified as “recreation” and “others” activity; and “recreation” activity is not easily misidentified as “others” activity. It means this kind of homogeneity is unidirectional. The first possible reason is the homogeneity among activities. It is a limitation of activity category in the survey. Some activities in different categories may happen at the similar period of a day, costing similar duration and locating near home or work place etc. In the current category system, “recreation” also includes “eating out” and “shopping”. While, “others” include escort, medical dental care and visit friend/relative etc. It can be possible that “meals” and “medical care” happens near one’s home with almost the same duration. These similar features will make the identification fail. However, some features of business activity and recreation activity may prevents the homogeneity turn to bidirectional.

Also, participants’ misunderstanding the category will also make contribution to the failure. In the dataset, some volunteers treat “visiting parks” as “recreation” while treat “visiting temples” as “other”. Consequently, based on the current attributes used, “recreation” activity and “others” activity are not as feature-distinctive as the other three. To be more specific, "other" now includes 1) "visit friend/relatives", 2) "visit temple", 3) "exercise", 4) "medical services", 5) "buy services (ward offices, bank etc.)", 6) "buy gas", 7) "escort" etc. In fact, 1) ~ 3) should be categorized as "social and recreational"; 4) ~ 7) should be shopping or personal business. This is may be one of the possible reasons of low accuracy of identifying "recreation" and "others".

2) No effective attributes are available in the existing data sets.

As explained in the first respective, independent variables in current data sets are not distinctive. Consequently, additional features may be necessary to make these two kinds of activity less homogeneous. Actually, any two kinds of activities can be distinguished by some distinctive features. Otherwise, they should be allocated into the same category. The only problem is whether the distinctive features can be obtained or not. Additional features that can be obtained without too much difficulty may include inter-persons related information, land use information, POI (point of interest, such as supermarkets, restaurants, banks, post offices, kindergarten, elementary school, railway/subway stations etc.) information, visit frequency, social network information etc. Inter-persons related information can be obtained from several persons’ GPS trajectories and their social relationship. Data set of Nagoya Metropolitan area only includes information of participants’ home and work place; other GIS information may also be helpful to improve the accuracy. About social network information, since some person likes updating their status on twitter real time; so adding this kind of information as independent variables also has the potential to improve the accuracy.

Based on the discussion above, possible solutions for more accurate results include: 1) refining activity type into more specific ones, as well as fixing the erroneously marked activity types and 2) adding possible effective information as independent variables.

5.5. SUMMARY

In this chapter, four machine learning methods are applied on identifying activity types with mobile phone GPS data collected in Nagoya Metropolitan area. Based on metrics of accuracy and time cost, it seems that classification tree shows superiority over one-against-rest SVM, neural networks, and discriminant analysis. Except "business" activity, classification tree identifies other four types of activity with higher accuracy and uses less time. Although its accuracy of identifying "business" is lower than the other three methods, it is still a satisfactory one. However, none of the four kinds of methods can handle well identifying "recreation" activity and "others" activity at a satisfactory accuracy. This may be caused by the homogeneity of features in these two types of activities and deficiency of effective features that can identify these two activities.

Future research can be furthered by including inter-persons related information, land use information, POI information, visit frequency, social network information etc. as input variables to check the performance of improving the accuracy. If each type of activity can be properly identified with a satisfactory accuracy by machine learning methods, GPS module on the mobile phone can be used as a reliable alternative to collect personal trip data with a lower cost during a longer period.

5.6. REFERENCE

Battiti, R., Brunato, M. (2014). *The LION Way: Machine Learning plus Intelligent Optimization*. LIONlab, University of Trento, Italy.

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM *Transactions on Intelligent Systems and Technology*. Vol, 2(3), 27.

Deng, Z., and Ji, M. (2010) Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. *Traffic and Transportation Studies* 2010. 768-777.

Griffin, T., Huang, Y. (2005). A decision tree classification model to automate trip purpose derivation. In *The Proceedings of the ISCA 18th International Conference on Computer Applications in Industry and Engineering*. 44-49.

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.

Hsu, C. W., and Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2), 415-425.

Hassen, A. M. A. (2013). The proposed System for Indoor Location Tracking. *International Journal of Computer Applications Technology and Research*, 2(6), 690-698.

Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer

BPNN architecture. *International Journal of Engineering Trends and Technology*, 3(6), 713-717.

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing* (pp. 41-50). Springer Berlin Heidelberg.

McGowen, P., and McNally, M. (2007). Evaluating the Potential to Predict Activity Types from GPS and GIS Data. *Western Regional Science Association 46th Annual Meeting*, , Newport Beach, CA

Poulsen, J., French, A. (2008). Discriminant function analysis. San Francisco State University: San Francisco, CA. [online]. Available at : http://userwww.sfsu.edu/efc/classes/biol710/discrim/discrim.pdf.

Sayed, M., Baker, F. (2015). E-Learning Optimization Using Supervised Artificial Neural-Network. *Journal of Software Engineering and Applications*, 8(01), 26.

Sheela, K. G., Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*.

Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of Remote Sensing*, 30(8), 2133-2147.

Tabachnick, B.G., L.S. Fidell. (1996). *Using Multivariate Statistics*. Harper Collins College Publishers: New York.

Chapter 6. Techniques of Improving Accuracy of Identifying Activity Type

6.1. INTRODUCTION

In the previous chapter, several machine learning methods have been applied on activity type identification. Some activity types, like "commute" and "home", can be identified with a higher accuracy due to the effectively distinctive attributes, distance from trip end to participant's home or work place. On the contrary, other activity types, such as "recreation" and "other" cannot be identified with a satisfactory accuracy due to currently used independent variables. In this chapter, several techniques are discussed trying to improve the average accuracy of activity type identification or certain type of activity types.

GIS information such as POI information, more specific activity type category, and social media information like twitter status are mentioned as possible ways to improve the accuracy in the previous chapter. Besides, data collected during longer time, especially several months, offers the possibility to analyze the weather's influence on the accuracy of identifying activity types. Consequently, in this chapter, POI information, and weather are used to analyze their influence on the accuracy of activity type identification.

Data collected in Hakodate city are used to analyze the influence from both GIS information and weather information. This data set is suitable to do the analysis considering weather's influence, since the data collection for each participant lasted for 8 months including winter and summer. Moreover, Hakodate city is not a mega city with many complex buildings like Nagoya. Complex building usually contains the shops and facilities in one building and it can provide the multiple service like "eating", "recreation", "shopping", and "personal affairs" etc. Too much complex buildings in a city will make the GIS information invalid to some extent and some other additional variables are needed very badly. Consequently, Hakodate data is suitable to analyze the influence from GIS information, too.

6.2. DATA PREPROCESSING

In the data set of Hakodate city, 8 trips with mode as "taxi" in the data base are treated as done by "auto", since the sample size of "taxi" is small and it usually has the same characteristics as "auto". 84 trips with activity type of "exercise" are treated as with a purpose of "recreation" since "exercise" is one kind of "recreation".

Categorical variables are encoded using one-of-K or one-hot encoding, in which the explanatory variable is encoded using one binary feature for each of the variable's possible values (Hackeling, 2014).

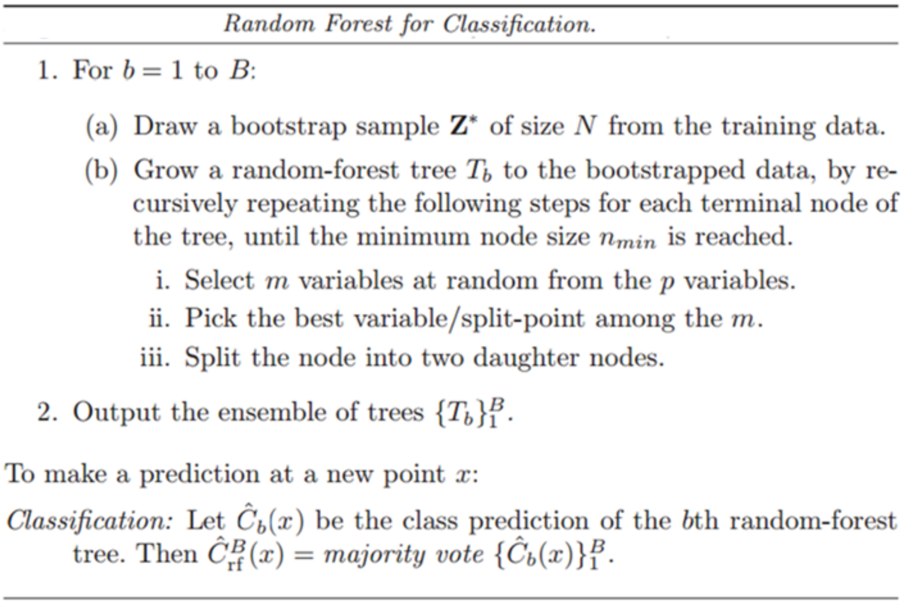
6.3. METHODOLOGY--RANDOM FORESTS

In this chapter, a new machine learning methods, random forests is used for analyzing the effect from features of weather and GIS dimension. In Chapter 5, after comparison, classification trees show a good performance in time cost and accuracy over support vector machines, discriminant analysis, and neural networks. Random forests can be regarded as an upgraded version of classification tree.

RF is a classification algorithm that uses an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training data using a randomly selected subset of variables or features (Breiman, 2001; Statnikov et al., 2008).

RF is derived from the decision tree method. The traditional decision tree method suffers from the overfitting problem when the trees grow very deep. The result is low bias but very high variance. However, RF is a substantial modification of the technique of bagging (or bootstrap aggregating), in which a large collection of de-correlated trees is built and then averaged. Bagging reduces the variance of an estimated prediction function (Hastie et al., 2009).

Bagging takes place in selecting the samples of the training set and the features used for training. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples: $b = 1, \ldots, B$. The reason for "feature bagging" is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable, these features will be selected in many trees, causing them to become correlated (Chauhan and Chauhan, 2014). The specific procedure for implementing RF is shown in Figure 6.1 (Friedman, J. et al., 2001).



Random Forest for Classification.

```
1. For $b = 1$ to $B$:
    (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
    (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
```

To make a prediction at a new point x:

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\mathrm{rf}}^B(x) = \textit{majority vote } \{\hat{C}_b(x)\}_1^B$.

Figure 6.1 Algorithm for random forest classification

The advantage of RF is that it not only provides accuracy in the training set and test set, but also indicates the importance of each attribute in the process. Thus, RF can divulge which weather related variable will have an effect on structuring machine learning for identifying activity type and travel mode.

In this chapter, RF is implemented using Python and the scikit-learn tool box (Pedregosa et al. 2011).

6.4. EFFECT FROM VARIABLES IN ADDITIONAL DIMENSIONS

Usually, more useful features in the data set contribute higher accuracy of activity type identification. In Chapter 5, features from activity dimension, trip dimension and demographic dimension are used in the comparison of activity type identification using four types of machine learning methods. And it concludes that the features from activity dimension contribute the most and features from demographic dimension have the least contribution. In this chapter, features from more dimensions are tested in the improvement of activity type identification. To be specific, features from weather dimension and GIS dimension are tested.

Weather features are obtained from Japan meteorological Agency. GIS features are obtained by calculating some variables based on the POI (Point of Interest) information. POI information includes the locations where activity happens. To be specific, these locations include offices, factories, ward office, police stations, banks, post offices, hospitals, universities, schools, kindergartens, temples, parks, sight-seeing spots and so on. Besides, bus stops, railway stations and streetcar stations are also included. Bus stop information was provided by Hakodate City Bus Corporation. And streetcar stations and other GIS information, are from AlpsMap data set. Because there is no shape of the location but only a point as the representative of each GIS location, the calculation is based on the coordinates of this representative point of location.

6.4.1. Scenario settings

In order to test the impact of weather features and GIS features in the activity type identification, four scenarios are set as in Table 6.1.

Table 6.1 Scenario settings for test additional features in activity type identification

Scenario	Set for usage	Summer data	Winter data	GIS related features	Weather related features
Scenario 1	Training set	70%	70%	○	○
	Test set	30%	30%	○	○
Scenario 2	Training set	70%	70%	○	●
	Test set	30%	30%	○	●
Scenario 3	Training set	70%	70%	●	○
	Test set	30%	30%	●	○
Scenario 4	Training set	70%	70%	●	●
	Test set	30%	30%	●	●

Note: ● means to be included, ○ means to be excluded
GIS related features are those obtained using GIS points of interests.

In each scenario, training set include 70% of summer data and 70% of winter data while test set include 30% summer data and 30% winter data. First, data set with only trip & activity features is calculated. Then weather features and GIS features are added separately to check how the accuracy changes. Finally, both the features from weather dimension and GIS dimension are added and check the corresponding accuracy.

6.4.2. Features as explanatory variables

Explanatory variables come from three dimensions: trip & activity dimension, weather dimension and GIS dimension. Explanatory variables in each dimension are explained in Table 6.2 and detailed explanation of GIS related variables is below the table.

Table 6.2 Explanatory variables description

Explanatory variables	Description
Trip and Activity Dimension	
Trip duration	Time cost of the trip
Trip length	The length of trip, calculated by summing consecutive GPS points with map-matched coordinates
Average speed of trip	Trip length divided by the trip duration
The day when trip starts	The day type when trip starts, dummy variables: 1 for weekday, 0 for weekend
The day when trip ends	The day type when trip ends, dummy variables: 1 for weekday, 0 for weekend
Activity duration	Time cost of the activity happening at trip end
Trips starts before morning peak?	Dummy variable, 1 for yes, 0 for no
Trip starts at morning peak?	Dummy variable, 1 for yes, 0 for no
Trip starts after morning peak and before noon?	Dummy variable, 1 for yes, 0 for no
Trip starts during the noon?	Dummy variable, 1 for yes, 0 for no
Trip starts after the noon and before evening peak?	Dummy variable, 1 for yes, 0 for no
Trip starts at evening peak?	Dummy variable, 1 for yes, 0 for no
Trip starts after evening peak?	Dummy variable, 1 for yes, 0 for no
Trip ends before morning peak?	Dummy variable, 1 for yes, 0 for no
Trip ends at morning peak?	Dummy variable, 1 for yes, 0 for no
Trip ends after morning peak and before noon?	Dummy variable, 1 for yes, 0 for no
Trip ends during the noon?	Dummy variable, 1 for yes, 0 for no
Trip ends after the noon and before evening peak?	Dummy variable, 1 for yes, 0 for no
Trip ends at evening peak?	Dummy variable, 1 for yes, 0 for no
Trip ends after evening peak?	Dummy variable, 1 for yes, 0 for no
Distance to home	Distance from trip end to participant’s home
Distance to work place	Distance from trip end to participant’s work place
Weather dimension	
Temperature when trip starts	Temperature value of the first GPS point in a trip
Temperature when trip ends	Temperature value of the last GPS point in a trip
Average temperature during the trip	Averaged value of temperature of all GPS points in a trip
Precipitation when trip starts	Precipitation value of the first GPS point in a trip
Precipitation when trip ends	Precipitation value of the last GPS point in a trip
Average precipitation during the trip	Averaged value of precipitation of all GPS points in a trip
Snow accumulation when trip starts	Snow accumulation value of the first GPS point in a trip
Snow accumulation when trip ends	Snow accumulation value of the last GPS point in a trip

Clear or cloudy when trip starts	Dummy variable of weather status, 1 for yes, 0 for no
Rainy when trip starts?	Dummy variable of weather status, 1 for yes, 0 for no
Snowy when trip starts?	Dummy variable of weather status, 1 for yes, 0 for no
Clear or cloudy when trip ends?	Dummy variable of weather status, 1 for yes, 0 for no
Rainy when trip ends?	Dummy variable of weather status, 1 for yes, 0 for no
Snowy when trip ends?	Dummy variable of weather status, 1 for yes, 0 for no
GIS dimension	
Density of non-activity stop	Number of non-activity stops divided by trip length
Maximum distance to bus stop	Maximum distance of non-activity stops in a trip to their nearest bus stops
Minimum distance to bus stop	Minimum distance of non-activity stops in a trip to their nearest bus stops
Average distance to bus stop	Average distance of non-activity stops in a trip to their nearest bus stops
Maximum distance to streetcar station	Maximum distance of non-activity stops in a trip to their nearest streetcar stations
Minimum distance to streetcar station	Minimum distance of non-activity stops in a trip to their nearest streetcar stations
Average distance to streetcar station	Average distance of non-activity stops in a trip to their nearest streetcar stations
Average distance to bus route	Average distance of the distances from GPS points in a trip to the bus route
Average distance streetcar line	Average distance of the distances from GPS points in a trip to the streetcar line
Number of shops	Number of shops in a given length of radius of the trip end
Distance to the nearest shop	Distance from the trip end to the nearest shop
Number of restaurants	Number of restaurants in a given length of radius of the trip end
Distance to the nearest restaurant	Distance from the trip end to its nearest restaurant
Number of leisure facilities	Number of leisure facilities in a given length of radius of the trip end
Distance to the nearest leisure facility	Distance from the trip end to the nearest leisure facility
Number of offices	Number of offices in a given length of radius of the trip end
Distance to the nearest office	Distance from the trip end to its nearest office

Note: distance from trip end to respondent’s home/work place is not put in the GIS dimension and it is because the address of their homes and workplaces are not obtained from GIS data but a prior questionnaire.

Among the GIS features in the above table, a key definition used in the calculation of GIS features is “non-activity stop” which first used in Chapter 4. Introducing this definition in this section is because of the fact that non-activity stop for a bus or a streetcar may be a temporary stop of these public transport system. Using GIS features related to non-activity stop may improve the accuracy of travel mode identification.

In the data set of Hakodate, for each trip trajectory, since the activity at the trip end has already been segmented from the trip, all identified stops using C-DBSCAN algorithm will be non-activity stops. Furthermore, as the GPS interval in Hakodate data set is a bit sparse (30 sec on average and 92% of the interval are less than 60 sec) and there are at most two points in a non-activity stop in most temporary stop location, the original DBSCAN algorithm without additional constraints is enough to identify the non-activity stop during the non-detoured trips in Hakodate. The distance from GPS point to the tram line or the bus route is also used as the GIS related feature.

In order to improve the accuracy of identifying the activity type such as “shopping”, “meal”, “recreation”, the shortest distance to these facilities and the number of these facilities in a radius of the trip end are also used as GIS features. The radius used in this research is 250 meter. A sensitivity analysis of this radius is also done in this research. According to the activity type of “shopping”, “meal”, “recreation” and “other” which show a lower accuracy of “commute” and “back-home”, the POI locations are categorized into corresponding groups. These groups include supermarkets, restaurants, recreational facilities and offices. Recreational facilities at the trip ends include hot spring shop, park, baseball ground, racecourse etc. Offices at the trip ends include bank, hospital or clinic, school, post office, police station, ward office, etc.

Frequency of visiting certain places is not used to identify the participants’ home and work places, since the address of their homes and work places has already been known by the questionnaire.

A map-matching method developed by Li et al. (2013) was used to match the original GPS trajectory points to the road network and decide which route the trip passed. The specific steps are as follows. Step 1, extract trip data from GPS trajectory point data. Step 2, extract links for a sub road network by selecting links locating within a radius (200m used in this research) of each GPS point of current trip; length of road links in this sub networks are shrunk by a given a factor valued from 0 to 1 to keep them in the shortest path with a higher probability. Step 3, search the shortest path from origin to the destination. Several shrunk factors have been tested and 0.1 can achieve the highest accuracy. Map-matched GPS points are used to calculate the trip length and the average speed of that trip.

6.4.3. Results and discussion

Identification accuracy and confusion matrix are used as performance metrics of classification used in this research. 90% is set as a satisfactory accuracy threshold.

6.4.3.1. Results of activity type identification

The accuracy results of activity type identification in the above scenarios are shown in Table 6.3. It can be concluded that including weather related features will decrease the general accuracy; while including GIS related features will increase the accuracy. It may because weather features have no influence on people’s intention of activity type. Including weather related features in the identification will lead to over-fitting relationship in the training process while it turns out bad performance in the prediction in the test set.

Table 6.3 Activity type identification accuracy with/without additional features

Activity type identification accuracy (training set) [test set]		weather related features involved			
		No		Yes	
GIS related features involved	No	(99%)	[88%]	(99%)	[85%]
	Yes	(100%)	[91%]	(100%)	[90%]

Table 6.4 shows the confusion matrix of activity type identification of test sets in the above scenarios. The specific results and explanation are as follows.

The purpose of “back home” and “commute” can be 100% identified in the scenarios no matter weather or GIS related variables involved or not. It is because the features of distance from trip end to the participant’s home/work place that guarantee the high accuracy. These two explanatory variables are ranked as the most important variables in the importance factor lists in each scenario.

Table 6.4 Confusion matrix of activity type identification of test sets in scenarios 13-16

Predicted	Scenario1: weather × GIS ×						Scenario2: weather ○ GIS ×						Scenario3: weather × GIS ○						Scenario4: weather ○ GIS ○					
Truth	**B**	**C**	**M**	**O**	**R**	**S**	**B**	**C**	**M**	**O**	**R**	**S**	**B**	**C**	**M**	**O**	**R**	**S**	**B**	**C**	**M**	**O**	**R**	**S**
Back home	**128**	0	0	0	0	0	**128**	0	0	0	0	0	**128**	0	0	0	0	0	**128**	0	0	0	0	0
Commute	0	**79**	0	0	0	0	0	**79**	0	0	0	0	0	**79**	0	0	0	0	0	**79**	0	0	0	0
Meal	0	0	**9**	6	0	3	0	0	**7**	3	0	8	0	1	**10**	1	1	5	0	1	**8**	2	0	7
Other	0	1	1	**9**	1	18	0	1	0	**5**	2	22	0	1	1	**14**	2	12	0	1	1	**13**	2	13
Recreation	0	1	0	1	**33**	1	0	0	1	0	**30**	5	0	0	2	1	**32**	1	0	1	1	1	**32**	1
Shopping	2	0	0	10	0	**69**	2	0	1	8	1	**69**	2	1	0	3	0	**75**	2	1	0	3	0	**75**

The purpose of “meal” can be identified with limited improvement when GIS related features involved, but its accuracy gets worse to some extent when inputting weather related features. The improvement of identifying “meal” purpose in scenario 3 and 4 is because of the input features related to restaurant in GIS point data.

The accuracy of identifying purpose of “other” can be improved a lot when involving GIS related variables but decreases when inputting weather related variables. With the improvement by GIS related variables, still almost half of the “other” trip will be identified as “shopping” erroneously. It is because the destinations of “other” purpose, like bank, post office, etc. are usually very close to the shopping facilities, and the current added GIS related variables still cannot distinguish them well.

Weather related variables’ involvement have negative influence on the identification of “recreation” trip while the GIS related variables have limited negative influence. Possible reasons include: 1) some types of “recreation” like “exercise” are not sensitive to rainy/snowy weather status. 2) “recreation” trip happened on a bad weather day may be identified as a trip non-sensitive to weather, like “shopping” which has similar features to “recreation”.

Adding weather related variables has no influence on identifying “shopping” purpose but GIS related variables does. It increases the accuracy due to the features like distance to the nearest shops from trip end and the density of shops near the trip end.

6.4.3.2. Results of travel mode identification

Table 6.5 shows the accuracy of travel mode identification in the above scenarios. It can be concluded that including either/both of weather related features and/or GIS related features will increase the accuracy. Weather’s improvement on the result can be explained by weather’s influence on people’s mode choice and it happens only to “bicycle”. If “bicycle” samples are excluded, weather related features will only have little negative effect in the identification process, and the conclusion will be the same as activity type identification. The negative effect from weather information can be explained as the overfitting in the training process.

Table 6.5 Travel mode identification accuracy with/without additional features

Travel mode identification accuracy (training set) [test set]		weather related features involved			
		No		Yes	
GIS related features involved	No	(100%)	[87%]	(100%)	[90%]
	Yes	(100%)	[88%]	(100%)	[92%]

Table 6.6 shows the confusion matrix of travel mode identification of test sets in the above scenarios. The specific results and explanation are as follows.

Involving weather related variables in identifying "auto" mode decreases the accuracy a little. On the contrary, inputting GIS related variables can improve the accuracy, due to GIS related information of bus and streetcar. With these information, trips by "auto" can be better distinguished from those by public transport system sharing a similar trip distance on the same route.

Table 6.6 Confusion matrix of travel mode identification of test sets in scenarios 13-16

Predicted / Truth	Weather × GIS ×					Weather ○ GIS ×					Weather × GIS ○					Weather ○ GIS ○				
	A	**BI**	**BU**	**S**	**W**	**A**	**BI**	**BU**	**S**	**W**	**A**	**BI**	**BU**	**S**	**W**	**A**	**BI**	**BU**	**S**	**W**
Auto	**206**	1	2	2	2	**205**	1	2	2	3	**210**	1	0	1	1	**209**	1	1	2	0
BIcycle	2	**28**	2	3	12	1	**43**	0	2	1	2	**27**	0	3	15	1	**45**	0	1	0
BUs	3	1	**7**	6	0	3	1	**6**	7	0	3	0	**8**	6	0	4	0	**7**	6	0
Streetcar	3	1	1	**33**	0	3	0	3	**32**	0	2	0	0	**36**	0	2	0	0	**36**	0
Walk	3	4	0	1	**49**	1	8	1	0	**47**	3	6	0	1	**47**	3	8	1	1	**44**

Weather related variables input in identifying "bicycle" can make higher accuracy while inputting GIS related variables decrease the accuracy a lit. It is because compared to other modes, "bicycle" mode more likely happens on a no-rain day (85%). Except the sensitivity to weather and distinct speed to motorized mode, additional GIS related features has no contribution to identifying "bicycle" mode. That is why when GIS related features are added, the general accuracy of identifying "bicycle" decreases a bit.

Involving weather related features decreases the accuracy of "bus" identification a bit while involving GIS related variables can improve the accuracy to some extent. However, still almost half of "bus" trips are erroneously identified as "streetcar". It is because some bus routes share the same road as streetcar line especially in the downtown area and it makes some trips cannot be distinguished correctly.

Involving weather related variables in identifying "streetcar" has a bit negative effect on the accuracy due to streetcar's insensitivity to weather while involving GIS related variables can evidently improve the accuracy due to the GIS information related to the "streetcar". Besides, no trip will erroneously be identified as "bus". It can be explained by the fact that trip by streetcar usually has a longer trip length than that by bus and it makes the misidentification more difficult.

Involving weather related or GIS related variables cannot improve the accuracy of identifying “walk”. It can be explained by the fact that trip by walk usually has a much short trip length which is insensitive to weather’s changing. And there is no features related to walk in the GIS related variables. Including them will make the learned algorithm in the training set become overfitting and make the prediction in the test set less accurate.

General conclusion of the impact from GIS features and weather features in activity type identification and travel mode identification is summarized as follows. 1) Including weather related and GIS related features have different influence on each specific type of activity types or travel modes. 2) For the average accuracy of identification, including weather related features has negative influence identifying activity type but positive influence on identifying travel mode (when “bicycle” sample is included in the data set); including GIS related features has positive influence on both activity type identification and travel mode identification.

6.4.3.3. Sensitivity analysis of accuracy due to radius variation at trip end

In order to test how the value of radius at the trip end will affect the general accuracy result, a series of radius value starting from 25 meter to 1000 meter with a step of 25 meters are tested in the above scenarios. The accuracy results of these tests are shown in Figure 6.2. It can be concluded that the accuracy is almost stable when the radius increases. And the minimum, maximum, mean and standard deviation of accuracies in each scenario is shown in the following table. It can be concluded that the accuracy is not dependent on the radius value. This conclusion is consistent with the ranking result of importance factor of GIS variables related to radius in RF result. RF results of scenarios with GIS variables, none of the variables of the number of shops/restaurants/leisure facilities/offices in a given radius are at a high ranking among explanatory variables.

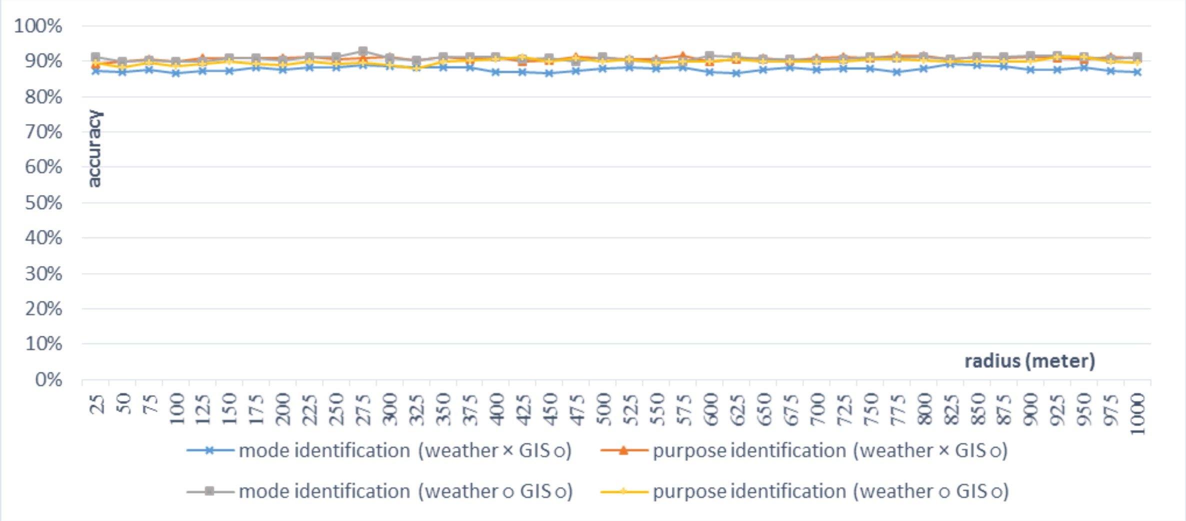


Figure 6.2 Accuracy changes due to different radius value in scenario 13~16

Table 6.7 Usual statistics of accuracies in each scenario

	mode identification (weather × GIS ○)	purpose identification (weather × GIS ○)	mode identification (weather ○ GIS ○)	purpose identification (weather ○ GIS ○)
Minimum	86.56%	89.25%	89.78%	87.90%
Maximum	89.25%	91.67%	92.74%	91.13%
Mean value	87.77%	90.77%	90.85%	89.87%
Std Dev.	0.71%	0.52%	0.62%	0.73%

6.5. DATA SELECTION FOR TRAINING SET AND TEST SET

This section shows the results of data selection for training set and test set when data are collected in distinct seasons. There are reasons why this investigation is important to this field. Mobile phone as a tool of PT data collection, can obtain the data not just for multiple days. The collection may last for several month or several years. When using machine learning methods to identify the activity type or travel mode, the ground truth of the identification is necessary. Whether a data set of one season with ground truth can be used as training set to predict a data set of opposite season is the first question to do the investigation.

Since Hakodate city is a city with distinct seasons, using winter data and summer data of Hakodate is suitable to achieve this goal. Before the calculation, some obvious uneven prerequisite should be cleared. In Hakodate data set, trips by bicycle only occurs in summer. So using winter data without bicycle trips surely cannot predict a good accuracy on the summer data set with bicycle trips. So 154 trips by bicycle are not included in the calculation in this section. And also for weather features, since there is no snow and snow accumulation in summer, weather features have also been processed as follows. 1) Weather status when trip starts/ends has been aggregated from 3 category (rainy, snowy, clear/cloudy) to 2 category (rainy/snowy, clear/cloudy) in this section for eliminating obvious feature difference among seasons. 2) Snow accumulation when trip starts/ends are not included in this section for the same reason. Other features are used the same as shown in Table 6.2.

6.5.1. Scenario settings

In order to test how to select data for training set and test set, 12 scenarios are set in this section. Detailed information can be found in Table 6.8

Table 6.8 Scenario settings for test data selection

Scenario	Set for usage	Summer data	Winter data	GIS related features	Weather related features
Scenario 1	Training set	70%	---	○	○
	Test set 1	---	100%	○	○
	Test set 2	30%	---	○	○
	Test set 3	30%	100%	○	○
Scenario 2	Training set	---	70%	○	○
	Test set 1	100%	---	○	○
	Test set 2	---	30%	○	○
	Test set 3	100%	30%	○	○

Scenario 3	Training set	50%	50%	○	○
	Test set 1	50%	---	○	○
	Test set 2	---	50%	○	○
	Test set 3	50%	50%	○	○
Scenario 4	Training set	70%	---	○	●
	Test set 1	---	100%	○	●
	Test set 2	30%	---	○	●
	Test set 3	30%	100%	○	●
Scenario 5	Training set	---	70%	○	●
	Test set 1	100%	---	○	●
	Test set 2	---	30%	○	●
	Test set 3	100%	30%	○	●
Scenario 6	Training set	50%	50%	○	●
	Test set 1	50%	---	○	●
	Test set 2	---	50%	○	●
	Test set 3	50%	50%	○	●
Scenario 7	Training set	70%	---	●	○
	Test set 1	---	100%	●	○
	Test set 2	30%	---	●	○
	Test set 3	30%	100%	●	○
Scenario 8	Training set	---	70%	●	○
	Test set 1	100%	---	●	○
	Test set 2	---	30%	●	○
	Test set 3	100%	30%	●	○
Scenario 9	Training set	50%	50%	●	○
	Test set 1	50%	---	●	○
	Test set 2	---	50%	●	○
	Test set 3	50%	50%	●	○
Scenario 10	Training set	70%	---	●	●
	Test set 1	---	100%	●	●
	Test set 2	30%	---	●	●
	Test set 3	30%	100%	●	●
Scenario 11	Training set	---	70%	●	●
	Test set 1	100%	---	●	●
	Test set 2	---	30%	●	●
	Test set 3	100%	30%	●	●
Scenario 12	Training set	50%	50%	●	●
	Test set 1	50%	---	●	●
	Test set 2	---	50%	●	●
	Test set 3	50%	50%	●	●

6.5.2. Results and discussions

Identification accuracy and confusion matrix are used as performance metrics of classification used in this research. 90% is set as a satisfactory accuracy threshold.

6.5.2.1. Results of activity type identification

The results of activity type identification in scenario 1 to 12 are shown in Figure 6.3 and Table 6.9. Detailed explanation is as follows.

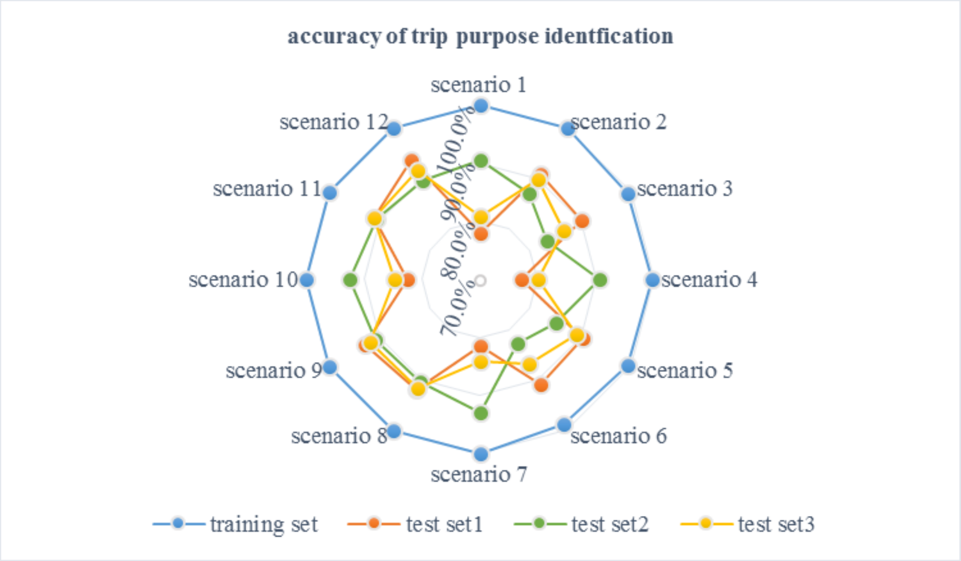


Figure 6.3　Accuracy of activity type identification in training set and test sets

Note: specific percentage of accuracy can be found in the following table.

Table 6.9　Accuracy of activity type identification in training set and test sets

purpose	**scenario 1**	**scenario 2**	**scenario 3**	**scenario 4**	**scenario 5**	**scenario 6**
Training set	100.0%	100.0%	99.3%	99.7%	99.5%	98.7%
Test set 1	77.9%	90.7%	90.2%	77.1%	90.5%	90.9%
Test set 2	90.6%	86.9%	83.2%	90.6%	85.1%	82.9%
Test set 3	80.7%	89.8%	86.6%	80.1%	89.2%	86.8%
purpose	**scenario 7**	**scenario 8**	**scenario 9**	**scenario 10**	**scenario 11**	**scenario 12**
Training set	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Test set 1	81.6%	92.0%	92.8%	82.5%	91.1%	93.6%
Test set 2	93.1%	90.5%	90.7%	92.5%	90.5%	89.6%
Test set 3	84.1%	91.7%	91.7%	84.7%	90.9%	91.5%

1) Summer data as training set

The test set composed only by summer data can reach a satisfactory accuracy. But the test set composed only by winter data or the mixed season data cannot reach a satisfactory accuracy.

2) Winter data as training set

The test set composed only by summer data can reach a satisfactory accuracy. But the test composed only by winter data or mixed season data cannot reach a satisfactory accuracy without GIS features as explanatory variables.

3) Two-season data as training set

If GIS features are not available in the test set, only the test set composed of summer data can achieve a satisfactory accuracy. If GIS features are available in the test set, test sets either composed only by summer data or winter data or mixed season data can reach a satisfactory.

Besides, there are several issues worth being noted. 1) Whether GIS related or weather related features are used as explanatory variables or not, training set from summer data can achieve a good accuracy on a test set from summer data, but cannot achieve a good result on a test set from winter data. And the accuracy gap between these two test sets are big, around 10%

to 13%. 2) When using winter data as training set, the accuracy on test set from summer data is higher than that from winter set unexpectedly. 3) When training set is made up of two-season data, the test set of only winter data show a worse performance than the test set of only summer data. From issues 1) to 3), it seems that no matter what kind of data is used in the training set, the test set of only winter data cannot achieve a better accuracy than a test set of only summer data. The reason of accuracy gap between test set of winter data and test set of summer data is the composition of activity type types in winter and summer. Winter season has a higher percentage of “meal”, “other”, “recreation”, and “shopping”, any of which is difficult to distinguish with current explanatory variables. However, 4) involving GIS related features can improve the identification accuracy of these activity type types and decrease the accuracy gap between these two kinds of test set to some extent.

Consequently, in a situation of low identification accuracy of some types of activity type, two season data should be used as training set and additional GIS related features should be included as explanatory variables. Then composition of activity type types can be balanced and effective GIS features can improve the performance of identifying activity types of “meal”, “other”, “recreation” and “shopping”.

6.5.2.2. Results of travel mode identification

The results of travel mode identification in scenario 1 to 12 are shown in Figure 6.4 and Table 6.10. The specific explanation are as follows.

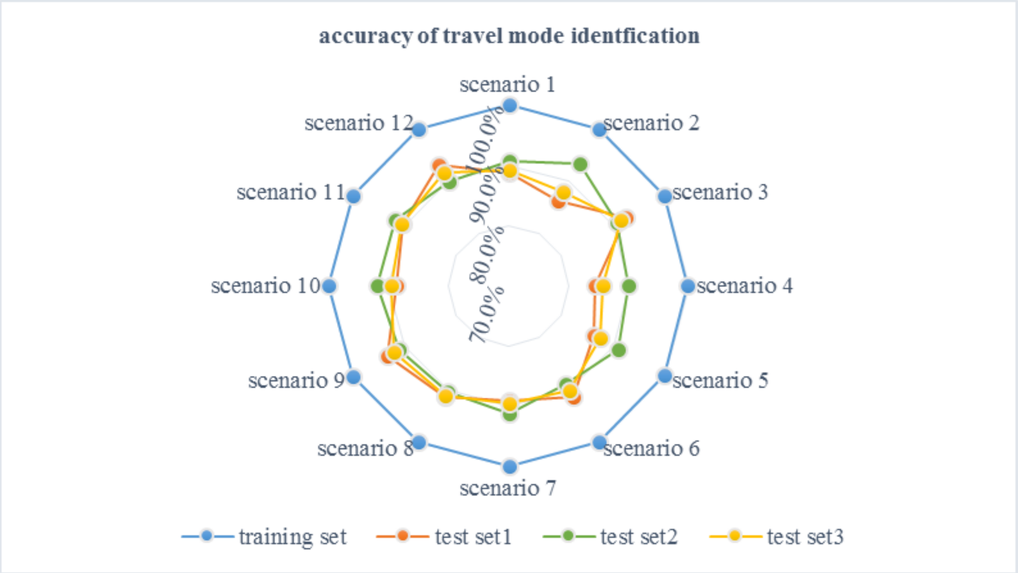


Figure 6.4 Accuracy of travel mode identification in training set and test sets

Note: specific percentage of accuracy can be found in the following table.

Table 6.10 Accuracy of travel mode identification in training set and test sets

Mode	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5	scenario 6
Training set	100.0%	100.0%	99.8%	99.7%	99.7%	100.0%
Test set 1	88.7%	86.2%	92.4%	84.3%	86.4%	91.3%
Test set 2	90.6%	93.5%	90.7%	89.9%	91.1%	88.9%

Test set 3	89.2%	87.9%	91.5%	85.5%	87.5%	90.1%
Mode	scenario 7	scenario 8	scenario 9	scenario 10	scenario 11	scenario 12
Training set	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Test set 1	88.9%	91.3%	93.2%	88.7%	90.3%	93.2%
Test set 2	91.2%	90.5%	91.1%	91.8%	91.7%	90.0%
Test set 3	89.6%	91.1%	92.1%	89.4%	90.7%	91.5%

1) Summer data as training set

The test set composed only by summer data can reach a satisfactory accuracy. But the test set composed only by winter data or mixed season data cannot reach a satisfactory accuracy.

2) Winter data as training set

The test set composed only by summer data or winter data or mixed season data can reach a satisfactory accuracy only when GIS features included as explanatory variables.

3) Two-season data as training set

The test sets either composed only by summer data or winter data or mixed season data can reach a satisfactory accuracy when weather features are not used as explanatory variables or GIS features are used as explanatory variables.

Besides, there are several issues worth being noted.1) Test set made up of samples from the same single season as training set can always get a higher accuracy than the test set of opposite season (except scenario 8 which show the opposite result. However, the two test sets get very close results in this scenario). It can be concluded that training set cannot be from one single season while the test set from the opposite season. This seasonal difference is caused by the different distribution of average speed of "auto" and "bus" among seasons. However, when GIS related features included as explanatory variables, the seasonal effect comes weaker since the accuracy gap between test set 1 and test set 2 are very small. 2) When using summer data as training set, the accuracy of test set from winter set are stable except scenario 4. It is because summer has less sample of trips by auto with speed less than 10 km/h than winter and the lower speed of auto in winter makes the auto erroneously identified as streetcar and walk. The overfitting caused by involving weather features in the training set of summer data will make this erroneous identification more easily happen. However, if GIS related features are included, the negative effect caused by the speed can be neglectable.

General conclusion for data selection for activity type identification and travel mode identification with data from several seasons: 1) For activity type identification, in the case of unbalanced composition of purpose type in different seasons and low accuracy of identifying some difficult activity types, test set composed of single season data same as those in training set cannot always achieve a satisfactory accuracy. Instead, it is better to use the mix-season data as the training set and GIS features to improve the accuracy of these difficult activity types. And the test sets can be composed of summer data, or winter data or mixed season data. 2) For travel mode identification, test set composed of single season data same as those in training set indeed can achieve a satisfactory accuracy. But mixed season data as training set can achieve a satisfactory accuracy on the test set of summer data or winter data or mixed season data. And including GIS features can improve the accuracy result.

The reason of the difference of accuracy between winter and summer for activity type identification and travel mode identification are different. 1) For activity type identification: winter-data test set always have a lower accuracy than summer data test set. The reasons are unbalanced composition of activity types among seasons and the low accuracy of some type of activity. 2) For travel mode identification: one season data as test set always has a lower accuracy than other test set whose data same as training set. The reason is that for each mode, the speed distribution of summer or winter is not from the same distributions.

The distributions of each mode's average speed during the trip in summer and winter are tested by univariate Kolmogorov-Smirnov test (K-S test) to check if their distributions are identical across seasons. For each travel mode, null hypothesis is that average speed during the trip has the same distribution in winter as in summer. It is found that for "auto" or "bus" (account for 71% of the whole sample size), the distribution in each season is not from the same distribution at both the significant levels of 1% and 5%. And for "streetcar" and "walk", the distribution in each season is from the same distribution at both the significant levels of 1% and 5%. This result can be explained by the fact that auto and bus are more easily to be effected by weather in winter, like the snow and ice on the road.

6.6. SUMMARY

In this chapter, several techniques are tested whether can improve the accuracy of activity type identification and travel mode identification.

Adding GIS features indeed can improve the accuracy of both the activity type identification and travel mode identification. Adding weather features only can improve the accuracy of travel mode identification.

About the data selection for training set and test sets: 1) For activity type identification, in the case of unbalanced composition of purpose type in different seasons and low accuracy of identifying some difficult types of activity, test set composed of single season data same as those in training set cannot always achieve a satisfactory accuracy. Instead, it is better to use the mix-season data as the training set and GIS features to improve the accuracy of these difficult activity types. And the test sets can be composed of summer data, or winter data or mixed season data. 2) For travel mode identification, test set composed of single season data same as those in training set indeed can achieve a satisfactory accuracy. But mixed season data as training set can achieve a satisfactory accuracy on the test set of summer data or winter data or mixed season data. And including GIS features can improve the accuracy result.

6.7. REFERENCE

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

Chauhan, M. & Chauhan, N. (2014). Application of random forest in various fields as the most reliable and effective data mining technique. *International Journal of Advance Research in Science and Engineering (IJARSE)*, Vol. No.3, Issue No.12, 281-286.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9, 1871-1874.

Friedman, J., Hastie, T. and Tibshirani, R., (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Gong, L., Sato, H., Yamamoto, T., Miwa, T., & Morikawa, T. (2015). Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation*, 23(3), 202-213.

Hackeling, G. (2014). Mastering Machine Learning with scikit-learn. Packt Publishing Ltd.

Li, Q., Cao, P. & Miao, L. (2013). Offline Map-matching for Archived Probe Vehicle Data. *Geomatics and Information Science of Wuhan University*, Vol. 38 No.2, 244-247 (in Chinese)

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1), 319.

Yuan, G. X., Ho, C. H., & Lin, C. J. (2012). An improved GLMNET for L1-regularized logistic regression. *The Journal of Machine Learning Research*, 13(1), 1999-2030.

Chapter 7. Daily Activity-Travel Pattern Analysis Focusing on Weather's Influence

7.1. INTRODUCTION

7.1.1. Background

GPS censors attached to mobile phones provide a possible way to collect PT data lasting for multiday with much less burden on the participants. And these multi-day data provide the possibility of analyzing the activity-travel pattern variations during days, under influence of some factors that only can be found changing during a longer time, like weather. Analyzing weather's influence on activity-travel behavior can contribute to improve the travel demand modeling, especially for the regions and cities with distinct seasons. In this chapter, the activity-travel pattern is examined in terms of how participants organize their trips during the day focusing weather's influence with data collected in Hakodate city.

Generally, people organize trips under the influence of various factors and try to obtain the maximum utility during the process of organizing trips. The number of trips in a trip chain, the number of trip chains in a day, and the total number of trips are three variables related to how people organize their trips (or travel demand) during the day. And it becomes a hot topic of identifying and analyzing key factors influencing variability of trip organizing. Existing research has already revealed the influencing factors from the dimensions of gender (McGuckin and Murakami, 1999), occupation (Bayarma et al. 2007; Chu, 2004), land use together with density (Dharmowijoyo et al. 2015; Noland and Thomas, 2007; Schmöcker el al. 2010; Susilo and Maat, 2007), weather (Liu et al. 2014; Liu et al. 2015), etc.

Among these factors above, weather is a key factor influencing not only the traffic volume (Cools et al. 2010b; Keay and Simmonds, 2005) in the macroscopic way, but also the travel behavior (Cools et al. 2010a) in the microscopic way. Generally speaking, weather variables change temporally through the year and spatially here and there. People in a specific geographical location may adapt their travel behavior to the local climate (Liu et al. 2015). So it would also be necessary to analyze weather's influence on people's travel behavior in the temporal prospective, to be specific, through a longer period of time, like several months covering different seasons. However, existing research of analyzing trip organizing, especially under the influence from weather factors, is almost based on one-day trip data. It would be anticipatory to derive the results of analyzing people how to organizing trips from the trip data collected during several months. In this chapter, three dependent variables are used to demonstrate the process of how people organize trips: the number of trips in a trip chain, the number of trip chains in a day, and the total number of trips in a day, and analyze the relationship between these dependent variables and independent variables from the several dimensions including weather dimension.

An important term in this paper is trip chain. Primerano et al. (2008) summarized the definition of trip chaining, and two most commonly accepted definitions of trip chains are a sequence of trips with anchored points of 1) both homes, and 2) home and work/school. In this paper, the first definition is used to integrate the sequence of trips in this chapter.

7.1.2. Descriptive analysis

The distribution of number of trips in a trip chain according to each type of weather condition in the day time is shown in Table 7.1. During a day of snow, Trip chains with more than three trips have a higher percentage (18%) of those during the day of rain (14%) or other (14%).

Table 7.1 Distribution of number of trips in a trip chain according to each type of weather condition

Number of trips	Other		Rain		Snow	
in a chain	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
2	65	38%	75	49%	46	41%
3	80	47%	57	37%	45	40%
4	17	10%	13	8%	12	11%
5	4	2%	6	4%	2	2%
6	2	1%	2	1%	3	3%
7	1	1%	1	1%	2	2%
8	0	0%	0	0%	1	1%
9	0	0%	0	0%	1	1%
Total	169	100%	154	100%	112	100%

The distribution of number of trip chains in a day according to each type of weather condition in the day time can be found in Table 7.2. Day of snow has a lower percentage of more trip chains.

Table 7.2 Distribution of number of trip chains in a day according to each type of weather condition

number of trip	Other		Rain		Snow	
chains in a day	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
1	71	60%	67	61%	59	70%
2	43	36%	39	36%	22	26%
3	4	3%	3	3%	3	4%
Total	118	100%	109	100%	84	100%

The distribution of the total number of trips in a day according to each type of weather condition is shown in Table 7.3. Day of other has the highest percentage of trips more than two, while day of snow is has the lowest percentage of trips more than two. It means that people is likely to have more trips on a day with good weather.

Table 7.3 Distribution of total number of trips in a day according to each type of weather condition

Total number of trips in a day	Other		Rain		Snow	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
2	10	8%	19	17%	21	25%
3	49	42%	34	31%	23	27%
4	23	19%	21	19%	12	14%
5	16	14%	21	19%	10	12%
6	11	9%	7	6%	8	10%
7	6	5%	5	5%	7	8%
8	0	0%	2	2%	1	1%
9	1	1%	0	0%	1	1%
10	1	1%	0	0%		0%
11	1	1%	0	0%	1	1%
Total	118	100%	109	100%	84	100%

The average number of trips per day, average daily trip sum of duration and average daily sum of trip distance are shown in Table 7.4. It is clear that when it is no rain or no snow, the volunteers tend to have more trips, and travel longer and further.

Table 7.4 Average value of number of trips per day and daily trip sum of duration & trip distance according types of weather condition

Weather conditions	Number trips in a day	Daily trip sum of duration (minutes)	Daily sum of trip distance (km)
other	4.04	55.12	12.71
rain	3.98	51.88	12.18
snow	3.98	54.71	10.84

From a qualitative point of view, the general weather condition indeed has an influence on the number of trips in a trip chain, the number of trip chains in a day and the number of trips in a day, which indicates specific weather variable may impact how people organize trips during the day and it is worth to analyze these kind of impact in a detailed and quantitative way.

7.2. METHODOLOGY

7.2.1. Model specification

Ordered logit models are used in this research to analyze how the variables from demographic dimension, time dimension, trip (chain) dimension, and weather dimension have impact on people's decision of organizing trips in trip chains. As mentioned in the introduction section, the number of trips in one day, the number of trip chains in one day, and the number of trips in a trip chain are chosen to be analyzed as the dependent variables respectively.

The ordered logit model is a regression model for ordinal dependent variables. Ordered logit model for a single latent variable y* can be expressed as:

$$y^* = \boldsymbol{X\beta} + \varepsilon \tag{7.1}$$

where y* is the unobserved propensity to increase the number of trips in a trip chain, or the number of trip chains in a day, or the number of trips in a day in our paper. $\boldsymbol{X}$ is the vector of independent variables, and $\boldsymbol{\beta}$ is the vector of coefficients representing the effect of the covariates. ε is the random error term that is assumed to be uncorrelated with $\boldsymbol{X}$. y^* cannot be observed, instead the categories of response can only be observed as follows.

$$\mathrm{y} = \begin{cases} 1, if\ y^* \leq \ \mu_1, \\ 2, if\ \mu_1 \leq y^* \leq \mu_2, \\ \quad\quad \vdots \\ j, if\ u_{j-1} \leq y^* \leq \mu_j, \\ \quad\quad \vdots \\ m, if\ y^* \geq \ \mu_{m-1}. \end{cases} \tag{7.2}$$

where $\mu_1, \mu_2, \ldots \mu_{m-1}$ are unknown cut points need to be estimated.

$\boldsymbol{\beta}$ is the partial change in y^* with respect to $\boldsymbol{X}$. To be specific, when $\boldsymbol{X}$ changes in a unit and all other variables keep constant, y^* is expected to change by $\boldsymbol{\beta}$ unit.

The probability that observation *i* chooses alternative *j* is:

$$p_{ij} = p(y = j) = \mathrm{p}\big(u_{j-1} \leq y^* \leq \mu_j\big) = F\big(\hat{\mu}_j - \boldsymbol{X_i}\widehat{\boldsymbol{\beta}}\big) - F\big(\hat{\mu}_{j-1} - \boldsymbol{X_i}\widehat{\boldsymbol{\beta}}\big) \tag{7.3}$$

where *F* is the logistic cumulative distribution function as follows:

$$F(z) = e^z/(1 + e^z) \tag{7.4}$$

Maximum likelihood is used to estimate $\boldsymbol{\beta}$ and $\mu_1, \mu_2, \ldots \mu_{m-1}$.

7.2.2. Explanatory variables

Explanatory variables considered in this paper are from three dimensions: time dimension, trip or trip chain dimension, and weather dimension. All available explanatory variables are shown in Table 7.5. Some variables from weather dimension are correlated from qualitative analysis. So excluding the variables with high correlation is necessary before estimating the ordered logit models. In order to test the correlation among variables, Pearson's correlation coefficient is used in this paper, and we use 0.75 as the threshold to decide the correlation is high or not. For the pairs of high correlated variables, only one variable in each pair is used in the model estimation. Variables in underline in Table 7.5 are those excluded variables based on the Pearson’s correlation coefficients.

Then the explanatory variables are tested for significance in a step-wise procedure and the variables with a significance more than 95% will be used in the final estimation.

Table 7.5 Specifications for explanatory variables

Explanatory variables	Description
Time Dimension	
Day of travel	What day is it when travel happens: Weekday (reference); weekend
Season	What season is it when travel happens: Summer (reference); winter
Trip (chain) dimension	
Period when trip chain starts	Morning peak (reference); during morning peak and noon; noon; during noon and evening peak; evening peak; after evening peak
Whether commute trip is in the trip chain	Yes (reference); no
Main mode in the trip chain	walk / bicycle; public transport; private car
Average speed (C)	Average speed of trip chain or all trips during one day; it is equal to total distance in a trip chain or one day divided by the corresponding trip time cost
Weather dimension	
Average atmospheric pressure in city (C)	Daily average atmospheric pressure in Hakodate city
Average atmospheric pressure at sea level (C)	Daily average atmospheric pressure at the sea near Hakodate city
Total precipitation (C)	Total precipitation during one day
Max precipitation in one hour (C)	Maximum of precipitation in continuous 60 minutes during one day
Max precipitation in ten minutes (C)	Maximum of precipitation in continuous 10 minutes during one day
Average temperature (C)	Average temperature on the survey day
Highest temperature (C)	Highest temperature on the survey day
Lowest temperature (C)	Lowest temperature on the survey day
Average humidity (C)	Average humidity on the survey day
Minimum humidity (C)	Minimum humidity on the survey day
Average wind speed (C)	Average wind speed on the survey day
Max wind speed (C)	Maximum of average wind speeds in every 10 minute
Instant max wind speed (C)	Maximum of all instant wind speeds on the survey day
Sunshine duration (C)	The time of sunshine lasting on the survey day
Total snow fall (C)	Total snow fall on the survey day
Maximum snow accumulation (C)	Maximum snow accumulation on the survey day
Weather condition during day time	No rain/snow (reference), rain, snow
Weather condition during night	No rain/snow (reference), rain, snow

Note: variables with a C in parentheses are continuous variables; others are dummy variables. Variables in underline are not included in the following analysis due to correlation to other weather related variables.

Besides the explanatory variables above, demographic information was also collected by a questionnaire together with the GPS-based trip survey. The information such as gender, age, whether holding a driving license, and the driving frequency etc. is included. However, since this chapter is focusing on using person-specific model to analyze the trip/travel behavior of two volunteers, these demographic variables are not included in the explanatory variable list.

Furthermore, the correlation among the three dependent variables is also tested. The highest one is 0.71, which happens between the number of trip chains in a day and the total number of trips in a day. So the estimating these three dependent variables individually is reasonable for behavioral analysis purpose.

7.3. ESTIMATION RESULTS

The estimation results of number of trips in a trip chain, number of trip chains in a day, and number of trips in a day for the two volunteer A and B are shown in Table 7.6~Table 7.8. The estimation results using two volunteers' data are also in the tables just for an informal comparison and an attempt of future calculation with data of all twenty volunteers.

Table 7.6 Estimation results-the number of trips in a trip chain

Explanatory variables	Volunteer A		Volunteer B		Volunteer A and B	
	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Time Dimension						
weekday	---	---	Ref.	---	---	---
weekend	---	---	-0.558	0.359	---	---
summer	Ref.	---	---	---	Ref.	---
winter	**-1.226**	**0.534***	---	---	**-0.880**	**0.399***
Trip chain start at morning peak	Ref.	---	---	---	Ref.	---
Trip chain start at non-peak between morning peak and noon	**2.253**	**1.132***	---	---	0.349	0.367
Trip chain start at noon	**2.615**	**1.137***	---	---	0.494	0.404
Trip chain start at non-peak between noon and evening peak	0.716	1.162	---	---	**-0.810**	**0.373***
Trip chain start at evening peak	0.649	1.297	---	---	-0.845	0.465
Trip chain start after evening peak	1.483	1.637	---	---	**-3.071**	**0.747****
Trip (chain) dimension						
Trip chain with commute	---	---	Ref.	---	Ref.	---
Trip chain without commute	---	---	0.358	0.350	**0.984**	**0.335****
Main mode is walk / bicycle	Ref.	---	---	---	Ref.	---
Main mode is public transport	**1.290**	**0.396****	---	---	**1.726**	**0.450****
Main mode is private car	**2.031**	**0.564****	---	---	**2.289**	**0.358****
average speed	**-0.110**	**0.047***	**0.099**	**0.032****	---	---
Weather dimension						
Average humidity	**-0.050**	**0.018****	---	---	**-0.031**	**0.015***
Sunshine duration	---	---	0.036	0.047	---	---
Snow accumulation	**0.032**	**0.014***	**0.057**	**0.015****	**0.041**	**0.011****
No rain/snow during day time	---	---	Ref.	---	Ref.	---
Rain during day time	---	---	-0.177	0.395	0.021	0.268
Snow during day time	---	---	**-1.585**	**0.627***	-0.744	0.401
Cut points						
μ1	**-2.166**	**1.828**	**0.946**	**0.741**	**-0.921**	**1.167**
μ2	**-0.441**	**1.819**	**3.951**	**0.799**	**1.412**	**1.168**
Model fit						
Number of observations	**231**		**204**		**435**	
Log-likelihood at zero	**-226.017**		**-195.938**		**-441.943**	
Log-likelihood at final	**-199.342**		**-180.669**		**-392.855**	

Note: --- Variables not included in the analysis. ** Significant at 1% level; * significant at 5% level

In terms of variables from time dimension, volunteer B is apt to have more trips and trip chains on weekends; volunteer A tends to make less trips in winter. But as for the number of trips in each trip chain, no significant difference is found between weekend and weekday, and either between seasons.

Concerning the variables from the trip and trip chain dimension, volunteer A tends to have more trips in a trip chain when the mode is public transport or private car, which is in line with previous research (Liu et al. 2015; Schmöcker et al. 2010; Ye et al. 2007); since volunteer B is vehicle-dependent type (there are no trips by other modes but only private car in the data set), this kind of tendency cannot be observed. However, volunteer A tends to have less trips in the trip chain when the average speed of the trip chain increases, while volunteer B tends in the opposite way. Volunteer B behaves contradictorily to the findings of previous research (Liu et al. 2015; Schmöcker et al. 2010), which indicates there is a negative correlation between average speed of trip chain and the number of trips in a trip chain due to a short dwelling time within 30 minutes between two trip chains at home. It may be because volunteer B prefers to achieve a series of trips in a trip chain without have a dwell at home. Volunteer A tends to have more trip chains and trips in a day when the average speed increases.

Table 7.7 Estimation results- the number of trip chains in a day

Explanatory variables	Volunteer A		Volunteer B		Volunteer A and B	
	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Time Dimension						
Weekday	---	---	Ref.	---	Ref.	---
Weekend	---	---	**2.843**	**0.461****	**1.039**	**0.260****
Summer	Ref.	---	---	---	---	---
Winter	**-1.667**	**0.624****	---	---	---	---
Trip (chain) dimension						
Average speed	**0.123**	**0.041****	-0.116	0.064	---	---
Weather dimension						
Total precipitation	-0.076	0.040	---	---	---	---
Average wind speed	-0.246	0.140	---	---	---	---
Snow accumulation	---	---	---	---	**-0.015**	**0.007***
No rain/snow during night	Ref.	---	Ref.	---	---	---
Rain during night	-0.028	0.488	0.343	0.439	---	---
Snow during night	**1.534**	**0.681***	**-1.521**	**0.631***	---	---
Cut points						
μ1	**0.331**	**0.608**	**-0.610**	**1.169**	**0.683**	**0.162**
μ2	**5.343**	**1.172**	**2.135**	**1.189**	**3.656**	**0.348**
Model fit						
Number of observations	**164**		**147**		**311**	
Log-likelihood at zero	**-115.716**		**-116.023**		**-238.242**	
Log-likelihood at final	**-100.781**		**-90.097**		**-227.805**	

Note: --- Variables not included in the analysis. ** Significant at 1% level; * significant at 5% level

As for the variables from the weather dimension, both of the volunteers tend to have more trips in a trip chain when there is more snow accumulation, which corresponds to the previous research (Liu et al. 2015), who shows that travelers tend to do more trips when the ground is covered with snow. One thing should be noted that volunteer B tend to have less trips in a trip chain when the weather condition is snow during the day time, which seems to be contradictory with the estimated result of variable of snow accumulation. It may be explained as that she tends to have less trips in a trip chain when there is snow during the day time compared to the day when it is no rain/snow. But if she decides to have trips on a snowy day, she prefers to achieve more trips in a trip chain. It can be also proved by the estimated results of the number of trips in a day. She tends to have less trips when the day is snowy compared to day without rain/snow; but if she decides to travel outside, she tends to have more trips when the snow fall increases or the snow accumulation increases. Volunteer A tends to have less trip chains in a day when the total precipitation increases; he tends to have more trips and trip chains when the weather condition (night) is snowy compared to the night without rain/snow. It seems to be a bit strange that volunteer A tends to have more trips when it is snowy and after checking the detailed information of trips on each type of weather condition, it may be explained by his trips with "other" purpose, like picking up family members at the station or somewhere else on a snowy night, which is very common in daily life. The same thing does not happen to Volunteer B who tends to have less trip chains during the night with snow. It could be explained as that because this kind of task is done by other family members.

Table 7.8 Estimation results- the number of trips in a day

Explanatory variables	Volunteer A		Volunteer B		Volunteer A and B	
	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Time Dimension						
weekday	---	---	Ref.	---	Ref.	---
weekend	---	---	**2.934**	**0.448****	**0.851**	**0.241****
summer	Ref.	---	---	---	---	---
winter	**-1.041**	**0.425***	---	---	---	---
Trip (chain) dimension						
average speed	**0.080**	**0.034***	---	---	**0.052**	**0.018****
Weather dimension						
Total precipitation	**-0.054**	**0.021***	---	---	**-0.021**	**0.011***
Average wind speed	**-0.259**	**0.105***	---	---	**-0.156**	**0.071***
Sunshine duration	---	---	0.079	0.056	---	---
Total snow fall	---	---	**0.271**	**0.087****	---	---
Snow accumulation	---	---	**0.061**	**0.016****	---	---
No rain/snow during day time	---	---	Ref.	---	---	---
Rain during day time	---	---	-0.010	0.445	---	---
Snow during day time	---	---	**-3.466**	**0.790****	---	---
No rain/snow during night	Ref.	---	---	---	---	---
Rain during night	-0.056	0.381	---	---	---	---
Snow during night	**1.193**	**0.477***	---	---	---	---
Cut points						
µ1	**-1.742**	**0.522**	**-1.941**	**0.563**	**-1.431**	**0.392**
µ2	**-0.681**	**0.505**	**1.538**	**0.542**	**0.319**	**0.376**
µ3	**0.714**	**0.511**	**1.901**	**0.554**	**1.117**	**0.379**
µ4	**1.864**	**0.541**	**2.913**	**0.588**	**2.006**	**0.395**
Model fit						
Number of observations	**164**		**147**		**311**	
Log-likelihood at zero	**-255.560**		**-193.546**		**-483.309**	
Log-likelihood at final	**-240.848**		**-162.957**		**-470.297**	

Note: --- Variables not included in the analysis. ** Significant at 1% level; * significant at 5% level

7.4. SUMMARY

In order to explore how people organize trips in a day, ordered logit models are used in this chapter to estimate the number of trips in a trip chain, the number of trip chains in a day and the total number of trips in a day. Explanatory variables from time dimension, trip (chain) dimension, and weather dimension are used in the estimation. The result shows that the significant variables from weather dimension are related to snow and rainfall. To be specific, it reveals that snow accumulation will make people tend to make the trip chain more complex. Total precipitation in a day will make people tend to decrease the number of trip chains in a day. Compared to fine weather, weather condition of snow make people have less trips; but if people decides to make trips, s/he tends to make more trips when the total snow fall or snow accumulation increases. Since the conclusion are significant variables related to rainfall and snow, which corresponds to previous research with one-day trip data with spatial variability, it proves to some extent that it is applicable and acceptable to use data with spatial variability to replace the time-series data which abound with weather variability.

The limitation of this chapter is the sample size is too small. Only two person's data are used. The reason is the ground truth of each trip has to be checked with related GIS information, due to some mistakes by the participants. The future work is to include the other 18 participants' data in the analysis as well as the demographic information.

7.5. REFERENCE

Bayarma, A., Kitamura, R., and Susilo, Y. (2007). Recurrence of daily travel patterns: stochastic process approach to multiday travel behavior. *Transportation Research Record: Journal of the Transportation Research Board,* (2021), 55-63.

Chu, Y. L. (2004). Daily stop-making model for workers. *Transportation Research Record: Journal of the Transportation Research Board,* (1894), 37-45.

Cools, M., Moons, E., Creemers, L. and Wets, G. (2010a). Changes in travel behavior in response to weather conditions. *Transportation Research Record: Journal of the Transportation Research Board*, 2157(1), 22-28.

Cools, M., Moons, E. and Wets, G. (2010b). Assessing the impact of weather on traffic intensity. *Weather, Climate, and Society*, 2(1), 60-68.

Dharmowijoyo, D. B., Susilo, Y. O., and Karlström, A. (2015). Day-to-day variability in travellers' activity-travel patterns in the Jakarta metropolitan area. *Transportation*, 1-21.

Keay, K. and Simmonds, I. (2005). The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. *Accident Analysis & Prevention,* 37(1), 109-124.

Liu, C., Susilo, Y. O., and Karlström, A. (2014). Examining the impact of weather variability on non-commuters' daily activity–travel patterns in different regions of Sweden. *Journal* *of Transport Geography*, 39, 36-48.

Liu, C., Susilo, Y. O., and Karlström, A. (2015). Measuring the impacts of weather variability on home-based trip chaining behaviour: a focus on spatial heterogeneity. *Transportation,* 1-25.

McGuckin, N., and Murakami, E. (1999). Examining trip-chaining behavior: Comparison of travel by men and women. *Transportation Research Record: Journal of the Transportation Research Board*, (1693), 79-85.

Noland, R. B., and Thomas, J. V. (2007). Multivariate analysis of trip-chaining behavior. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 34(6), 953.

Schmöcker, J. D., Su, F., and Noland, R. B. (2010). An analysis of trip chaining among older London residents. *Transportation,* 37(1), 105-123.

Susilo, Y. O., and Maat, K. (2007). The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation,* 34(5), 589-609.

Primerano, F., Taylor, M. A., Pitaksringkarn, L., and Tisato, P. (2008). Defining and understanding trip chaining behaviour. *Transportation*, 35(1), 55-72.

Ye, X., Pendyala, R. M., and Gottardi, G. (2007). An exploration of the relationship between mode choice and complexity of trip chaining patterns. *Transportation Research Part B: Methodological*, 41(1), 96-113

Chapter 8. Activity Sequence Generation Model

8.1. INTRODUCTION

Activity-travel pattern can be decomposed into following components: activity type, activity location, activity duration, trip duration, and travel mode. Using PT data can regenerate/ simulate the activity travel pattern. Especially with multiday or multi-month or even continuous collected PT data by mobile phone, the activity-travel pattern of the mobile phone user can be predicted. If there is a tremendous difference between the predicted activity-travel pattern with the real one, it may mean an anomaly pattern is detected. And some further countermeasures are needed, in the case of the mobile phone users are the elderly, living alone. It is important to a society with a huge percentage of old people, especially in Japan.

However, with current features used as explanatory variables, activity type and travel mode identified with machine learning methods are still with errors. It means that this error should also be included in the regeneration/ simulation model. Although using traditional PT data can also do the estimation of this regeneration/ simulation model, the unknown error will make the model not precise as it is thought. With the data already known its error, the model can be built in a way to including it in the model itself.

In this chapter, just for demonstrating the procedure of including the identification error from GPS data, a simplified version of activity-travel pattern, “activity sequence” is analyzed as the object.

Data collected in Nagoya Metropolitan area are used to do the estimation. And the whole data set of activity sequence are randomly stratified into two sub sets with almost equal sample size. The number of activities in each sub set varies due to the difference of number of activities in each activity sequence. The first sub set is used as training set for activity type identification in machine learning. The second sub set is used as testing set for activity type identification in the machine learning and the identified result will also be used as input for parameter estimation in activity sequence model. Activity types of ground truth in the second sub set will also be used for parameter estimation in activity sequence model. Two groups of estimated parameters will be used for prediction in the first sub set to compare the performance.

8.2. METHODOLOGY

8.2.1. Model specification

A similar idea of activity generation to that advanced by Kitamura et al. (1997) and Kitamura et al. (2000) is used in this chapter which assumes that next activity depends on the previous activities on the same day. In this chapter, it is assumed that the next activity type depends on the previous activities on the same day. Another assumption is that the activity sequence on each day is independent from another on another day, no matter these activity sequences coming from the same participant or not.

In a sample data set consisted of *I* persons, for person i, a series of activities ($X_{i,j}$) in an activity sequence ($\boldsymbol{X_i}$) is of him/her:

$$\boldsymbol{X_i} = X_{i,J}, X_{i,J-1}, \cdots X_{i,j}, \cdots, X_{i,1} \tag{8.1}$$

Based on the dependence of activities in one activity sequence, the probability of an activity sequence of person *i* is:

$$\begin{aligned}\Pr(\boldsymbol{X_i}) = \Pr\left(X_{i,J}\middle|X_{i,J-1}, X_{i,J-2,} \cdots, X_{i,1}\right) \times \Pr\left(X_{i,J-1}\middle|X_{i,J-2}, X_{i,J-3}, \cdots, X_{i,1}\right) \\ \times \cdots \times \Pr\left(X_{i,j}|X_{i,j-1}, \cdots, X_{i,1}\right) \times \cdots \times \Pr(X_{i,1})\end{aligned} \tag{8.2}$$

where $\Pr\left(X_{i,j}|X_{i,j-1}, \cdots, X_{i,1}\right)$ is the probability of choosing *j th* activity. And the probability that a certain type of activity *k* is chosen as *j th* activity $X_{i,j}$ is denoted as for simplification:

$$p_{i,j} = \Pr\left(X_{i,j} = k\middle|\boldsymbol{X_{i,j-1}}\right) = \boldsymbol{F}\left(k\colon t_{i,j-1}, \boldsymbol{D_{i,j-1}}, \boldsymbol{Z_i}, \boldsymbol{K_{i,j}}\right), \forall k \in \boldsymbol{K_{i,j}} \tag{8.3}$$

where $\boldsymbol{X_{i,j-1}} = X_{i,j-1}, \cdots, X_{i,1}$, is the activity sequence history until *(j-1) th* activity; $t_{i,j-1}$ is the time period of *(j-1) th* trip finishes; $\boldsymbol{D_{i,j-1}}$ is a dummy matrix which demonstrates whether a type of activity is in the list of previous activities; $\boldsymbol{K_{i,j}}$ is the activity types available for the next choice of activity type.

For a certain activity *k*, the result of choice by person *i* is:

$$y_{i,k} = \begin{cases}1, when\ activity\ of\ type\ k\ is\ selected \\ 0, when\ activity\ of\ type\ k\ is\ not\ selected\end{cases} \tag{8.4}$$

And $\sum_{k \in K} y_{i,k} = 1$.

Then the probability of an activity sequence $\boldsymbol{X_i}$ on a certain day can be expressed as:

$$\prod_{j=1}^{J} \prod_{k \in K} {p_{i,j}}^{y_{i,k}} \tag{8.5}$$

Then the joint probability of *I* persons each of who chooses an activity sequence $\boldsymbol{X_i}$ on a certain day can be expressed as:

$$L^* = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k \in K} {p_{i,j}}^{y_{i,k}} \tag{8.6}$$

Maximum likelihood methods is used to do the estimation on the joint probability equation. Let $L = \ln(L^*)$, then:

$$L = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k \in K} y_{i,k} \ln(p_{i,j}) \tag{8.7}$$

If *K th* activity type is used as baseline, then the logits of the first *K-1* activity type with the baseline is:

$$\ln\left(\frac{p_{i,j}}{p_{i,J}}\right) = \ln\left(\frac{p_{i,j}}{1-\sum_{j=1}^{J-1} p_{ij}}\right) = \sum_{q=0}^{Q} x_{iq}\,\beta_{qk} \tag{8.8}$$

where x_{iq} is the *q th* explanatory variables and β_{qk} is the coefficient for the *q th* explanatory variables in the linear equation.

When identified activity type from mobile phone GPS data are used for the above estimation, $p_{i,j}$ needs an additional index, the identification accuracy, to demonstrate the real probability. So the above equation should be expressed as:

$$\ln\left(\frac{p_{i,j}}{p_{i,J}}\right) = \ln(\frac{p_{i,j} r_j}{p_{i,J} r_J}) \tag{8.9}$$

where r_j is the accuracy of identifying activity type *j*. However, the additional r_j and r_J can be treated as constant value which can be moved on the right as follows.

$$\ln\left(\frac{p_{i,j} r_j}{p_{i,J} r_J}\right) = \ln\left(\frac{p_{i,j}}{p_{i,J}}\right) + \ln\left(\frac{r_j}{r_J}\right) \tag{8.10}$$

Then

$$\ln\left(\frac{p_{i,j}}{p_{i,J}}\right) = \ln\left(\frac{p_{i,j}}{1-\sum_{j=1}^{J-1} p_{ij}}\right) = -\ln\left(\frac{r_j}{r_J}\right) + \sum_{q=0}^{Q} x_{iq}\,\beta_{qk} \tag{8.11}$$

$-\ln\left(\frac{r_j}{r_J}\right)$ can be combined with the constant value $x_{i0}\beta_{0k}$ as a new constant value $x_{i0}\beta_{0k}$. Then the above equation is converted back to the logits of the first *K-1* activity type with the baseline. It means that identified activity type with error can be estimated using the same method as there is no error in the identification.

$$p_{i,j} = \frac{e^{\sum_{q=0}^{Q} x_{iq}\beta_{qk}}}{1+\sum_{k=1}^{K-1} e^{\sum_{q=0}^{Q} x_{iq}\beta_{qk}}}, for\ k < K$$

$$p_{i,j} = \frac{1}{1+\sum_{k=1}^{K-1} e^{\sum_{q=0}^{Q} x_{iq}\beta_{qk}}}, for\ k = K \tag{8.12}$$

8.2.2. Model components and constraints

In this chapter, activity type choice models for workers and non-workers are not estimated respectively as those in Kitamura et al. (1997) and Kitamura et al. (2000). It is because that in the data set of Nagoya metropolitan area, even the non-workers have a regular part-time job.

Originally, three models, home-based" model, "work-place-based" model and "other-place-based" model are planned to develop separately with different choice sets of activity types. Same definition of "home-based" model as in Kitamura et al. (2000) is used in this chapter. "Home-based" model is dealing with the trips generating from home or activities whose previous activity happens at home. Definitions of the other two models are derived from the above definition. “Work-place-based” model is dealing with the trips generating from work place or activities whose previous activity happens at work place. “other-place-based” model, on the contrary to the former two, dealing with the trips generating from other places rather than home or work place, or activities whose previous activity happens at other places rather than home or work place. The choice sets of these models are shown in Table 8.1.

Table 8.1 Choice set composition of the four models

Item in choice set	Home-based model	Work-place-based model	Other-place-based model
Back home		●	●
Back office/school	●		●
Business	●	●	●
Commute	●		●
Eating out	●	●	●
Other	●	●	●
Personal affairs	●	●	●
Shopping	●	●	●
Social and recreational	●	●	●

Note: ● means included in the model’s choice set. “Business” means work related business.

However, since the sample size in the data set of Nagoya metropolitan area is limited. Finally, a whole general model is developed due to limited sample size.

Another issue is about the constraints in the model development. Constraints in the choice set of activity type mentioned in Kitamura et al. (2000) are not applicable in the data set of Nagoya metropolitan area. 1) Some business trip starts from participant’s home and return to the work place directly after the business trip. So it means “back work” can happen without “work” as a prerequisite. 2) Some participants go home during the noon and go to work again in the afternoon. So it means there can be twice “work” in a whole day’s activity sequence. 3) Some participants have “business” activities both in the morning and afternoon with a “back work” activity to the work place. It means “back work” can be more than once in a whole day’s activity sequence. As a result, some other constraints are applied in the data analysis in this chapter. These constraints include 1) the first activity in a day’s activity sequence cannot be “back home”; and 2) the last activity in a day’s activity sequence should be “back home”.

8.2.3. Explanatory variables

Description of explanatory variables used in the estimation is shown in Table 8.2. Explanatory variables from three dimensions are used, including time dimension, activity history dimension, and demographic dimension. Demographic features are not included in the model finally, since they cause the non-convergence to the model estimation.

Table 8.2 Description of explanatory variables used in activity sequence generation model

Explanatory variables	Description
Time Dimension	
Activity starts period	The temporal period when activity starts. Select from one of the seven categories
Activity ends period	The temporal period when activity ends. Select from one of the seven categories
Activity start day	The day type when trip starts, dummy variables: 1 for weekday, 0 for weekend
Activity end day	The day type when trip ends, dummy variables: 1 for weekday, 0 for weekend

Demographic Dimension	
Gender	Dummy variable, 0 for male and 1 for female
Age	One of the eight age intervals
Income	One of the ten income intervals
Occupation	One of the eight occupation categories
Auto frequency	One of the five frequency types
Public transit frequency	One of the five frequency types
Activity History Dimension	
Back home	Dummy variable, 1 for happened in the activity history on the same day.
Back office/school	Dummy variable, 1 for happened in the activity history on the same day.
Business	Dummy variable, 1 for happened in the activity history on the same day.
Commute	Dummy variable, 1 for happened in the activity history on the same day.
Eating out	Dummy variable, 1 for happened in the activity history on the same day.
Other	Dummy variable, 1 for happened in the activity history on the same day.
Personal affairs	Dummy variable, 1 for happened in the activity history on the same day.
Shopping	Dummy variable, 1 for happened in the activity history on the same day.
Social and recreational	Dummy variable, 1 for happened in the activity history on the same day.

Note:

1) 7 categories of period are periods: NP-b-MP: non-peak before morning peak; MP: morning peak; NP-b-MP-NOON: non-peak between morning peak and noon; Noon; NP-b-NOON-EP: non-peak between noon and evening peak; EP: evening peak; NP-a-EP: non-peak after evening peak. Peak periods and noon period are defined as follows: morning peak starts from 7am to 9am on weekdays and weekends; evening peak starts from 4pm to 7pm on weekdays and 3pm to 6pm on weekends. Noon period starts from 12pm to 1pm on weekdays and weekends. Morning peak and evening peak are equivalent to those in the report of current traffic characteristics in Chukyo Area in Japan.

2) 8 categories of age are as follows: 0~10 years old; 11~20 years old; 21~30 years old; 31~40 years old; 41~50 years old; 51~60 years old; 61~70 years old; 71~80 years old; more than 80 years old.

3) 10 categories of annual household income are following intervals: [0,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8), [8,10), [10,15), [15,+∞); unit is million Japanese Yen.

4) 8 categories of occupation include: employee; self-employer or management; part time or freelance; government/school related; student; housewife; other; no occupation.

5) 5 categories of auto / public transit usage frequency cover: more than 5 days per week; 3~4 days per week; 1~2 days per week; 2~3 times per month; less than once per month.

Before the estimation, the correlation between each of the explanatory variables in the above table are tested by phi coefficient test. As a result, only one variable in the pair of highly correlated variables is included in the estimation.

8.3. ACTIVITY TYPE IDENTIFICATION RESULT

Activity features mentioned in Table 5.1 used for comparing machine learning methods are selected as explanatory variables. Classification trees are used as the machine learning method to implement the activity type identification. Data collected in Nagoya metropolitan area is used but in a more specific category. The type “recreation” is specifically split into “eating out”, “shopping” and “social and recreational”; the type “other” is split into “personal affairs” and “other”. The composition of the activity types in training set and test set is in the following figure.

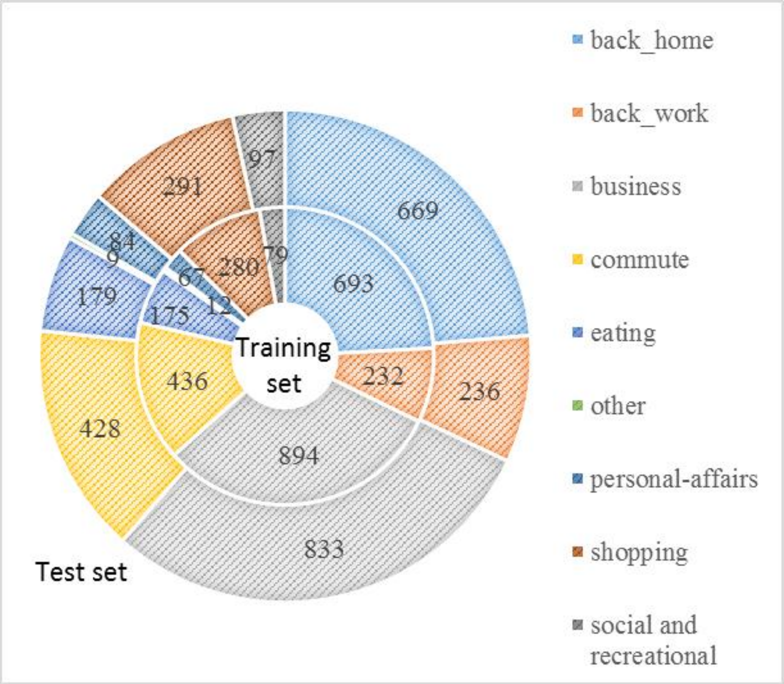


Figure 8.1 Composition

The general accuracy is 92.4% in the training set and 79.8% in the test set. The specific accuracy of each activity type can be found in Table 8.3. “Back home”, “back work”, “business” and “commute” are identified both in the training set and test set with a high accuracy. “Personal affairs”, “shopping”, “eating out” and “social and recreational” cannot be identified well. “Other” cannot be identified both in training set and test set and it may be because of the limited sample size of it.

Table 8.3 Confusion matrix of activity type identification accuracy

Training set									
Predicted Reality	**BH**	**BW**	**BU**	**C**	**E**	**O**	**P**	**SH**	**SR**
Back **H**ome	100%	0%	0%	0%	0%	0%	0%	0%	0%
Back **W**ork	0%	95%	0%	5%	0%	0%	0%	0%	0%
BUsiness	0%	0%	97%	0%	1%	0%	0%	1%	0%
Commute	0%	1%	0%	99%	0%	0%	0%	0%	0%
Eating out	0%	1%	15%	0%	78%	0%	1%	4%	1%
Other	0%	0%	33%	0%	33%	0%	8%	17%	8%
Personal-affairs	0%	0%	19%	0%	15%	0%	43%	18%	4%
SHopping	0%	0%	13%	0%	3%	0%	2%	79%	3%
Social and **R**ecreational	0%	0%	13%	0%	14%	0%	0%	11%	62%
Test set									
Predicted Reality	**BH**	**BW**	**BU**	**C**	**E**	**O**	**P**	**SH**	**SR**
Back **H**ome	100%	0%	0%	0%	0%	0%	0%	0%	0%
Back **W**ork	0%	87%	0%	13%	0%	0%	0%	0%	0%
BUsiness	0%	0%	84%	0%	6%	0%	1%	8%	2%
Commute	0%	3%	0%	96%	0%	0%	0%	0%	0%
Eating out	0%	0%	33%	0%	42%	0%	1%	17%	7%
Other	0%	0%	56%	0%	33%	0%	0%	0%	11%

Personal-affairs	0%	0%	36%	0%	10%	0%	24%	25%	6%
SHopping	0%	0%	26%	0%	15%	0%	2%	51%	5%
Social and **R**ecreational	0%	0%	24%	0%	21%	0%	2%	28%	26%

8.4. ESTIMATION RESULTS AND DISCUSSION

The test set with identified activity type in machine learning and the test set with real activity type are used for estimation respectively. There is no estimation result for the activity type of “other”, since there is no “other” activity types obtained in the results of machine learning. Type of “commute” is selected as the referenced outcome.

The two estimation results agree with each other in the significance level of most the explanatory variables.

“Back home” less possibly happens before noon and more possibly happens from evening peak which is in correspondence with most of people’s daily behavior. “Back home” most possibly ends at the morning peak which is the time when people starts to leave their home for some activities outside. “Back home” is not likely to happen when it already happened on the same day; it is likely (or certainly) to happen when other types of activity has already happened.

“Back office/school” is likely to happen from the time period between morning peak and noon, and usually to happen when “commute” has already happened. It means people intend to go to the work place first and then leave the work place for work related business. And it is likely to end from the time period between noon and evening peak. Besides “commute”, other types of activities, such as “business”, “eating out”, and “shopping” are also likely to happen before “back office/school”. “Eating out” and “shopping” listed in the activity history of “back office/school” are usually for lunch. However, “back office/school” is more unlikely to happen when “back home” has already happened and it is understandable since people have already gone home and finished one-day’s trips.

“Business” is not likely to happen in the morning but after the morning peak hour and is likely to happen after the activity “commute” and other “business” activities have already happened on the same day. Since the duration of business activity usually done in 30 minutes, it is not surprising to find that “business” is likely to end from the time period between morning peak and noon. But it is unlikely to happen when people have already reached home.

“Eating out” is not likely to happen on weekdays. It is likely to happen around noon and evening peak which is the time for lunch or dinner. And it is likely to happen when other types of activities have already happened, such as “commute”, “business”, etc. But it is not likely to happen when people have already arrived home.

“Personal affairs” is likely to happen and end around evening peak. What is more, it is likely to happen when “shopping”, “social and recreational” and other “personal affairs” have already happened on the same day.

“Shopping” is not likely to happen on weekday. It is more likely to happen and end from the period between morning peak and noon. It is also likely to happen when “personal affairs”, other “shopping”, and “social and recreational” have already happened on the same day.

“Social and recreational” is likely to happen from the time period between morning peak and noon. It is more likely to happen when “personal affairs”, other “shopping”, and “social and recreational” have already happened on the same day.

Table 8.4 Estimation results of two data sets

Back home	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	**-2.545**	**0.501**	**	**-3.038**	**0.516**	**
activity starts on weekday	-1.166	0.927		-1.728	0.939	
activity ends on weekday	-1.171	0.922		-0.714	0.934	
activity starts before morning peak	reference level			reference level		
activity starts in morning peak	---	---		---	---	
activity starts during morning peak and noon	**-0.982**	**0.465**	*	---	---	
activity starts at noon	**1.377**	**0.472**	**	**2.164**	**0.471**	**
activity starts between noon and evening peak	**3.412**	**0.460**	**	**3.798**	**0.484**	**
activity starts at evening peak	**5.993**	**0.500**	**	**5.463**	**0.491**	**
activity starts after evening peak	**8.949**	**1.073**	**	**8.893**	**1.114**	**
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	**3.971**	**0.525**	**	**3.408**	**0.486**	**
activity ends during morning peak and noon	**2.759**	**0.514**	**	**2.808**	**0.543**	**
activity ends at noon	**1.383**	**0.540**	*	**1.762**	**0.565**	**
activity ends between noon and evening peak	0.372	0.556		0.866	0.580	
activity ends at evening peak	-0.797	0.519		-0.825	0.536	
activity ends after evening peak	**-4.082**	**0.509**	**	**-3.419**	**0.534**	**
back home already happened	**-1.138**	**0.565**	*	**-1.784**	**0.584**	**
back work already happened	-0.186	0.514		-0.620	1.205	
business already happened	**1.293**	**0.317**	**	**2.678**	**0.481**	**
commute already happened	**2.620**	**0.289**	**	**3.902**	**0.415**	**
eating out already happened	0.977	0.528		1.011	0.761	
other already happened	---	---		2.854	4.080	
personal affairs already happened	**2.167**	**0.733**	**	**2.051**	**0.576**	**
shopping already happened	**1.955**	**0.375**	**	**1.961**	**0.408**	**
social and recreational already happened	**3.262**	**1.128**	**	**3.190**	**0.835**	**
Back office/school	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	**-11.737**	**3.811**	**	**-12.567**	**1.398**	**
activity starts on weekday	-0.999	2.414		-3.032	3.773	
activity ends on weekday	1.901	2.410		2.839	3.758	

activity starts before morning peak	reference level			reference level		
activity starts in morning peak	---	---		---	---	
activity starts during morning peak and noon	0.122	0.801		**4.395**	**1.047**	**
activity starts at noon	**2.445**	**0.677**	**	**4.503**	**1.016**	**
activity starts between noon and evening peak	**2.253**	**0.660**	**	**3.588**	**1.005**	**
activity starts at evening peak	**3.650**	**0.689**	**	**2.972**	**1.013**	**
activity starts after evening peak	2.417	1.280		1.701	1.518	
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	6.047	3.787		---	---	
activity ends during morning peak and noon	4.832	3.821		0.146	1.429	
activity ends at noon	3.743	3.782		0.549	1.344	
activity ends between noon and evening peak	5.392	3.771		**2.725**	**1.323**	*
activity ends at evening peak	5.812	3.762		**3.381**	**1.300**	**
activity ends after evening peak	4.822	3.757		**3.948**	**1.292**	**
back home already happened	**-1.471**	**0.630**	*	**-2.876**	**0.689**	**
back work already happened	-0.488	0.519		-1.206	1.214	
business already happened	**2.835**	**0.352**	**	**5.557**	**0.538**	**
commute already happened	**4.021**	**0.365**	**	**7.817**	**0.627**	**
eating out already happened	**1.500**	**0.546**	**	**1.546**	**0.782**	*
other already happened	---	---		5.382	4.180	
personal affairs already happened	1.273	1.070		0.930	0.701	
shopping already happened	**0.977**	**0.397**	*	**1.432**	**0.450**	**
social and recreational already happened	2.259	1.564		**2.594**	**1.019**	*
Business	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	**-5.185**	**0.689**	**	**-5.878**	**0.649**	**
activity starts on weekday	-1.364	1.006		-1.954	1.073	
activity ends on weekday	0.965	0.999		1.591	1.065	
activity starts before morning peak	reference level			reference level		
activity starts in morning peak	**-1.055**	**0.504**	*	0.691	0.513	
activity starts during morning peak and noon	**1.165**	**0.458**	*	**2.336**	**0.488**	**
activity starts at noon	**1.664**	**0.500**	**	**2.707**	**0.556**	**
activity starts between noon and evening peak	**3.004**	**0.502**	**	**3.423**	**0.558**	**
activity starts at evening peak	**4.202**	**0.562**	**	**2.723**	**0.579**	**
activity starts after evening peak	**4.238**	**1.142**	**	**3.560**	**1.187**	**
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	---	---		---	---	
activity ends during morning peak and noon	**4.276**	**0.641**	**	**3.435**	**0.622**	**

activity ends at noon	**4.236**	**0.645**	**	**3.166**	**0.629**	**
activity ends between noon and evening peak	**4.212**	**0.653**	**	**3.585**	**0.631**	**
activity ends at evening peak	**3.677**	**0.632**	**	**3.303**	**0.600**	**
activity ends after evening peak	0.319	0.629		**1.338**	**0.600**	*
back home already happened	**-1.253**	**0.548**	*	**-2.353**	**0.588**	**
back work already happened	0.537	0.488		-0.304	1.202	
business already happened	**1.736**	**0.284**	**	**3.893**	**0.466**	**
commute already happened	**2.330**	**0.256**	**	**4.310**	**0.405**	**
eating out already happened	0.784	0.525		1.154	0.763	
other already happened	---	---		1.894	4.185	
personal affairs already happened	1.121	0.709		0.837	0.593	
shopping already happened	**0.948**	**0.357**	**	0.660	0.405	
social and recreational already happened	2.079	1.162		1.006	0.947	
Commute	(base outcome)					
Eating out	*Estimation with results in ML*			*Estimation using real activity*		
Explanatory variables	*Coef.*	*Std. Err.*		*Coef.*	*Std. Err.*	
constant	**-4.143**	**1.070**	**	**-4.787**	**0.663**	**
activity starts on weekday	-1.712	0.995		**-2.129**	**1.011**	*
activity ends on weekday	-0.277	0.988		0.043	1.001	
activity starts before morning peak	reference level			reference level		
activity starts in morning peak	0.901	1.139		---	---	
activity starts during morning peak and noon	1.112	1.169		---	---	
activity starts at noon	**4.502**	**1.073**	**	**3.931**	**0.577**	**
activity starts between noon and evening peak	**3.769**	**1.083**	**	**3.530**	**0.626**	**
activity starts at evening peak	**5.614**	**1.092**	**	**4.807**	**0.616**	**
activity starts after evening peak	**7.757**	**1.445**	**	**7.218**	**1.174**	**
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	---	---		---	---	
activity ends during morning peak and noon	0.949	0.786		0.390	1.158	
activity ends at noon	**1.382**	**0.616**	*	**2.201**	**0.622**	**
activity ends between noon and evening peak	**1.473**	**0.627**	*	**2.448**	**0.630**	**
activity ends at evening peak	0.596	0.614		0.067	0.629	
activity ends after evening peak	0.046	0.553		0.100	0.557	
back home already happened	-0.869	0.572		**-2.239**	**0.614**	**
back work already happened	0.829	0.525		0.010	1.212	
business already happened	**0.693**	**0.317**	*	**3.083**	**0.489**	**
commute already happened	**1.969**	**0.300**	**	**3.926**	**0.439**	**
eating out already happened	0.885	0.543		0.319	0.785	

other already happened	---	---		4.114	4.105	
personal affairs already happened	0.318	0.855		**1.750**	**0.609**	**
shopping already happened	0.700	0.385		**1.284**	**0.423**	**
social and recreational already happened	2.141	1.187		**3.392**	**0.853**	**

Other	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	---	---		-11.299	6.207	
activity starts on weekday	---	---		-2.606	2.717	
activity ends on weekday	---	---		0.458	2.696	
activity starts before morning peak	---	---		reference level		
activity starts in morning peak	---	---		---	---	
activity starts during morning peak and noon	---	---		4.760	3.941	
activity starts at noon	---	---		---	---	
activity starts between noon and evening peak	---	---		5.242	3.956	
activity starts at evening peak	---	---		5.179	3.959	
activity starts after evening peak	---	---		**8.484**	**4.103**	*
activity ends before morning peak	---	---		reference level		
activity ends in morning peak	---	---		---	---	
activity ends during morning peak and noon	---	---		4.439	5.098	
activity ends at noon	---	---		---	---	
activity ends between noon and evening peak	---	---		---	---	
activity ends at evening peak	---	---		4.770	4.975	
activity ends after evening peak	---	---		3.548	4.969	
back home already happened	---	---		-1.876	1.253	
back work already happened	---	---		0.209	1.662	
business already happened	---	---		**2.492**	**0.986**	*
commute already happened	---	---		**3.107**	**1.009**	**
eating out already happened	---	---		1.004	1.164	
other already happened	---	---		-1.137	18.759	
personal affairs already happened	---	---		1.682	1.236	
shopping already happened	---	---		1.322	0.866	
social and recreational already happened	---	---		**2.812**	**1.394**	*

Personal affairs	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	**-3.497**	**1.103**	**	**-5.094**	**1.129**	**
activity starts on weekday	-0.650	1.175		-1.495	1.118	
activity ends on weekday	-0.992	1.162		-0.209	1.110	
activity starts before morning peak	reference level			reference level		

activity starts in morning peak	0.028	0.935		0.663	0.908	
activity starts during morning peak and noon	0.608	1.080		**2.024**	**0.893**	*
activity starts at noon	1.745	1.022		**2.889**	**0.913**	**
activity starts between noon and evening peak	1.533	1.140		**3.248**	**0.922**	**
activity starts at evening peak	**3.719**	**0.996**	**	**4.461**	**0.910**	**
activity starts after evening peak	**6.736**	**1.348**	**	**6.572**	**1.358**	**
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	**3.457**	**0.993**	**	**2.688**	**0.979**	**
activity ends during morning peak and noon	0.600	1.176		1.793	1.002	
activity ends at noon	1.503	1.024		**2.968**	**0.926**	**
activity ends between noon and evening peak	-0.270	1.393		**2.367**	**0.943**	*
activity ends at evening peak	0.216	1.017		1.503	0.889	
activity ends after evening peak	-0.521	0.897		0.264	0.867	
back home already happened	0.494	0.705		-0.691	0.607	
back work already happened	0.221	0.830		-1.129	1.363	
business already happened	-0.520	0.573		**1.336**	**0.585**	*
commute already happened	0.263	0.571		**2.699**	**0.491**	**
eating out already happened	1.189	0.697		**1.777**	**0.795**	*
other already happened	---	---		3.178	4.150	
personal affairs already happened	**1.799**	**0.835**	*	**2.380**	**0.593**	**
shopping already happened	0.877	0.543		**1.198**	**0.459**	**
social and recreational already happened	**3.216**	**1.196**	**	**2.095**	**0.898**	*

Shopping	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	**-2.847**	**0.638**	**	**-5.516**	**0.801**	**
activity starts on weekday	-1.187	0.960		**-2.391**	**0.986**	*
activity ends on weekday	-0.629	0.955		-0.035	0.981	
activity starts before morning peak	reference level			reference level		
activity starts in morning peak	-0.830	0.610		---	---	
activity starts during morning peak and noon	**1.111**	**0.549**	*	**2.949**	**0.514**	**
activity starts at noon	**1.731**	**0.581**	**	**3.736**	**0.579**	**
activity starts between noon and evening peak	**2.878**	**0.574**	**	**4.946**	**0.599**	**
activity starts at evening peak	**4.077**	**0.621**	**	**5.383**	**0.617**	**
activity starts after evening peak	**6.734**	**1.132**	**	**8.162**	**1.174**	**
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	---	---		---	---	
activity ends during morning peak and noon	**2.100**	**0.598**	**	**3.420**	**0.741**	**
activity ends at noon	**2.581**	**0.580**	**	**3.964**	**0.723**	**

activity ends between noon and evening peak	**2.110**	**0.594**	**	**3.639**	**0.729**	**
activity ends at evening peak	**1.862**	**0.562**	**	**2.413**	**0.700**	**
activity ends after evening peak	-0.044	0.543		1.234	0.685	
back home already happened	-0.154	0.526		**-1.201**	**0.551**	*
back work already happened	0.220	0.522		0.090	1.217	
business already happened	0.483	0.305		**1.884**	**0.485**	**
commute already happened	**1.517**	**0.279**	**	**3.049**	**0.421**	**
eating out already happened	0.760	0.535		1.018	0.763	
other already happened	---	---		2.262	4.140	
personal affairs already happened	1.207	0.706		**1.767**	**0.558**	**
shopping already happened	**1.380**	**0.359**	**	**1.395**	**0.395**	**
social and recreational already happened	**3.000**	**1.119**	**	**2.463**	**0.833**	**

Social and recreational	***Estimation with results in ML***			***Estimation using real activity***		
Explanatory variables	***Coef.***	***Std. Err.***		***Coef.***	***Std. Err.***	
constant	**-4.371**	**1.010**	**	**-4.047**	**0.917**	**
activity starts on weekday	-1.569	1.214		-2.109	1.103	
activity ends on weekday	-0.705	1.204		-0.288	1.095	
activity starts before morning peak	reference level			reference level		
activity starts in morning peak	---	---		0.274	0.956	
activity starts during morning peak and noon	**3.089**	**0.778**	**	**2.286**	**0.875**	**
activity starts at noon	**2.873**	**0.856**	**	**2.614**	**0.905**	**
activity starts between noon and evening peak	**3.646**	**0.831**	**	**3.443**	**0.877**	**
activity starts at evening peak	**4.473**	**0.874**	**	**3.645**	**0.884**	**
activity starts after evening peak	**7.159**	**1.289**	**	**6.437**	**1.339**	**
activity ends before morning peak	reference level			reference level		
activity ends in morning peak	---	---		---	---	
activity ends during morning peak and noon	1.025	0.964		0.636	0.943	
activity ends at noon	1.160	0.934		**1.665**	**0.804**	*
activity ends between noon and evening peak	1.638	0.907		**1.967**	**0.779**	*
activity ends at evening peak	0.577	0.915		**1.639**	**0.718**	*
activity ends after evening peak	0.452	0.834		0.396	0.701	
back home already happened	-0.264	0.632		-0.726	0.590	
back work already happened	0.955	0.642		0.339	1.275	
business already happened	0.025	0.431		**1.444**	**0.570**	*
commute already happened	**1.135**	**0.414**	**	**3.014**	**0.486**	**
eating out already happened	0.949	0.614		0.868	0.800	
other already happened	---	---		-3.291	21.139	
personal affairs already happened	1.526	0.788		**1.895**	**0.602**	**

shopping already happened	**0.915**	**0.450**	*	**1.319**	**0.436**	**
social and recreational already happened	**3.624**	**1.148**	**	**2.773**	**0.854**	**
Model fit						
Number of observations	**2826**			**2772**		
Log-likelihood at zero	**-5876.50**			**-6209.36**		
Log-likelihood at final	**-2645.17**			**-2637.53**		

Note: ** is the sign of significant at 1% level and * is for 5% level.

Then these two estimated results are used to predict the activity types in the first sub set. And the chi square test is used to compare the frequency of activity types in each time period in the two predicted results. It can be concluded that except the activity types of “eating” and “shopping”, these two predicted frequency distribution are not significantly different from each other.

Table 8.5 Frequency comparison on forecasting performance between using predicted activity in ML in test set and using real activity in test set

Time period	Back-home			Back-work		
	using prediction	using actual	χ^2	using prediction	using actual	χ^2
Before morning peak	65	64	0.02	0	0	0.00
Morning peak	5	4	0.00	0	0	0.00
Bet. morning peak and noon	2	2	0.00	1	5	3.20
Noon	15	13	0.31	29	29	0.00
Bet. noon and evening peak	25	25	0.00	36	39	0.23
Evening peak	142	149	0.33	136	123	1.37
After evening peak	403	397	0.09	0	1	1.00
p-value	0.986			0.569		
Time period	Business			Commute		
	using prediction	using actual	χ^2	using prediction	using actual	χ^2
Before morning peak	0	0	0.00	33	34	0.03
Morning peak	5	6	0.17	331	331	0.00
Bet. morning peak and noon	207	199	0.32	76	77	0.01
Noon	274	257	1.12	30	42	3.43
Bet. noon and evening peak	437	434	0.02	17	14	0.64
Evening peak	170	160	0.63	12	18	2.00
After evening peak	0	1	1.00	0	0	0.00
p-value	0.894			0.411		
Time period	Eating			Personal affairs		
	using prediction	using actual	χ^2	using prediction	using actual	χ^2
Before morning peak	0	0	0.00	0	0	0.00
Morning peak	0	0	0.00	0	0	0.00
Bet. morning peak and noon	0	0	0.00	0	0	0.00
Noon	36	10	67.60	0	2	2.00
Bet. noon and evening peak	2	0	0.00	0	1	1.00
Evening peak	15	13	0.31	0	7	7.00
After evening peak	63	70	0.70	3	1	4.00
p-value	**0.000****			0.677		
Time period	Shopping			Social and recreational		
	using prediction	using actual	χ^2	using prediction	using actual	χ^2

Before morning peak	0	0	0.00	0	0	0.00
Morning peak	0	0	0.00	0	0	0.00
Bet. morning peak and noon	36	39	0.23	1	1	0.00
Noon	18	51	21.35	2	0	0.00
Bet. noon and evening peak	75	82	0.60	3	0	0.00
Evening peak	66	72	0.50	1	0	0.00
After evening peak	94	95	0.01	2	0	0.00
p-value	**0.001****			1.000		

Note: sign ** means to reject the null hypothesis that two frequency distributions are the same at 1% level.

8.5. SUMMARY

In this chapter, an activity sequence generation model is developed and estimated with the activity identified from GPS trajectory with error. The model is decomposed as a series of multinomial logit model with the activity that have already happened on the same day as a conditional probability to the next activity choice behavior. The data set has been split into two sub sets, one for training algorithm in machine learning and the other for predicting activity types as a test set. Then the predicted and actual activity types in the test set are used to estimate the activity sequence model respectively. The estimated results are used for prediction on the first sub set to compare the performance. The chi square test shows that the predictions with the two estimated results can generate the same results except the activity type of “eating out” and “shopping”.

There are also some limitations in this chapter and they need to be solved in the future work. 1) More specific models should be estimated with bigger sample size data set. 2) Other tests, like Fisher's exact test should be used to compare the final performance result, due to the limitation of chi square test. 3) Activity identification and estimation without “other” activity should be done and compared with the current results, since the number of “other” activity is very limited in the current data set.

8.6. REFERENCE

Kitamura, R., Chen, C., Pendyala, R. M., & Narayanan, R. (2000). Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 27(1), 25-51.

Kitamura, R., Chen, C., & Pendyala, R. (1997). Generation of synthetic daily activity-travel patterns. Transportation Research Record: *Journal of the Transportation Research Board,* (1607), 154-162.

Chapter 9. Conclusions and Future Research

9.1. CONCLUSIONS AND CONTRIBUTIONS

This thesis is generated in the context of popularity of smart phone with GPS sensors and unavoidable demerits of traditional PT surveys. It tries to obtain the PT data (focusing on activity type) from continuous GPS trajectory data. Based on the obtained PT data, activity-travel pattern is analyzed trying to investigate the possible positive influence from multi-week and multi-month data. Besides, an activity sequence model is developed to regenerate the activities considering the activity identification error. The contributions and conclusions of this thesis are summarized as follows.

9.1.1. Contributions

An improved density-based algorithm with SVM is advanced for identifying activity stops from continuous GPS trajectories. In order to deal with the GPS data without the features of speed and acceleration, the density-based algorithm is improved by adding additional constraints from temporal and spatial respective to identify all the stop locations in the first step. Then SVMs is used to distinguish the activity stop location and non-activity stop location in the second step.

In order to fill in the gap that which machine learning method is the most suitable one for identifying activity type, four machine learning methods are tested and compared. It is found that classification tree performs superiority in time cost and accuracy among these four machine learning methods.

Besides, whether some techniques, such as data selection for training set and test set from distinctive season data, weather related features and GIS related features, can improve the accuracy of activity type identification is tested in this thesis, too.

And the activity-travel pattern is analyzed using the multi-month data focusing on the influence of weather’s temporal heterogeneity on trip chaining.

Furthermore, considering the accuracy of activity type identification in the estimation of multinomial logit model also fill in the gap of using PT data derived from GPS data in the activity-travel pattern generation.

9.1.2. Conclusions

A two-step methodology is proposed for identifying activity stops in continuous trajectories utilizing a variation of the DBSCAN algorithm and the SVMs method. In order to adjust DBSCAN to the context of GPS trajectories, two constraints are applied as improvements: a time sequence constraint and a direction change constraint. Application of this proposed methodology to GPS data collected using mobile phones in Nagoya area of Japan in 2008 demonstrates that the improved DBSCAN algorithm (C-DBSCAN) achieves an accuracy of

90% in identifying stop locations and the SVMs method is almost 96% accurate in distinguishing activity stops from non-activity stops.

Four machine learning methods are applied on identifying activity types with mobile phone GPS data collected in Nagoya Metropolitan area. Based on metrics of accuracy and time cost, it seems that classification tree shows superiority over one-against-rest SVM, neural networks, and discriminant analysis. However, none of the four kinds of methods can handle well identifying “recreation” activity and “others” activity at a satisfactory accuracy. This may be caused by the homogeneity of features in these two types of activities and deficiency of effective features that can identify these two activities.

The impact from GIS features and weather features in activity type identification and travel mode identification is summarized as follows. 1) Including weather related and GIS related features have different influence on each specific activity type or travel modes. 2) For the average accuracy of identification, including weather related features has negative influence identifying activity type but positive influence on identifying travel mode (when “bicycle” sample is included in the data set); including GIS related features has positive influence on both activity type identification and travel mode identification.

Methods of data selection for activity type identification and travel mode identification with data from several seasons: 1) For activity type identification, in the case of unbalanced composition of purpose type in different seasons and low accuracy of identifying some difficult activity types, test set composed of single season data same as those in training set cannot always achieve a satisfactory accuracy. Instead, it is better to use the mix-season data as the training set and GIS features to improve the accuracy of these difficult types of activities. And the test sets can be composed of summer data, or winter data or mixed season data. 2) For travel mode identification, test set composed of single season data same as those in training set indeed can achieve a satisfactory accuracy. But mixed season data as training set can achieve a satisfactory accuracy on the test set of summer data or winter data or mixed season data. And including GIS features can improve the accuracy result.

Ordered logit models are used to estimate the number of trips in a trip chain, the number of trip chains in a day and the total number of trips in a day. The result shows that the significant variables from weather dimension are related to snow and rainfall. To be specific, it reveals that snow accumulation will make people tend to make the trip chain more complex. Total precipitation in a day will make people tend to decrease the number of trip chains in a day. Compared to fine weather, weather condition of snow make people have less trips; but if people decides to make trips, s/he tends to make more trips when the total snow fall or snow accumulation increases. Since the conclusion are significant variables related to rainfall and snow, which corresponds to previous research with one-day trip data with spatial variability, it proves to some extent that it is applicable and acceptable to use data with spatial variability to replace the time-series data which abound with weather variability.

An activity sequence generation model is developed and estimated with the activity identified from GPS trajectory with error. The model is decomposed as a series of multinomial logit model with the activity that have already happened on the same day as a conditional probability to the next activity choice behavior. The data set has been split into two sub sets, one for training algorithm in machine learning and the other for predicting activity types as a test set. Then the predicted and actual activity types in the test set are used to estimate the activity sequence model respectively. The estimated results are used for prediction on the first sub set to compare the performance. The chi square test shows that the predictions with the two estimated results can generate the same results except the activity type of “eating out” and “shopping”.

9.2. FUTURE RESEARCH

There are also some limitations in this thesis and further research can be done to improve these limitations. Furthermore, some continuous analysis can be done based on the results in this thesis. To be specific, the future research includes the following points.

The improved density-based algorithm for identifying stop locations in GPS trajectories, C-DBSCAN is designed when the GPS data are collected in a very frequent level. It means this algorithm is not applicable to the GPS data collected with large intervals. A direction of future research can be identifying activity locations with GPS data collected in an extensive level.

Although the GIS shows an improvement on the activity type identification, there are still gap between the current achieved accuracy and the satisfactory accuracy. Future work in this field can focus on searching additional independent variables with positive effect to improve the activity type identification. These additional independent variable candidates can include visit frequency, accompaniment status, social media information, etc.

The sample size of analyzing trip chaining in Hakodate city is too small. Only two person’s data are used. The reason is the ground truth of each trip has to be checked with related GIS information, due to some mistakes by the participants. The future work is to include the other 18 participants’ data in the analysis as well as the demographic information.

When demonstrating the GPS trajectories on the GIS map, an interesting finding is that trivial travels which may be eliminated in the traditional person trip survey, since the participants may think they are not important. These trivial travels include the shopping at convenience stores, withdrawing/depositing money at ATMs, sending mails at post offices/ pillar boxes, etc. and these activities usually have a very short durations, usually from 1 minutes to 5 minutes at most. These activities seem to be insignificant. However, when intending to have these trivial activities, not only additional time is spent, but also the travel route may also change. And it will also influence the subsequent activities on the same day. As a result, to analyze the characteristics of these trivial activities, and to model when these activities happen as well as how these activities influence the route choice behavior and the subsequent activities are very necessary.

As far as activity sequence generation model is concerned, 1) more specific models should be estimated with bigger sample size data set. 2) Other tests, like Fisher's exact test should be used to compare the final performance result, due to the limitation of chi square test. 3) Activity identification and estimation without “other” activity should be done and compared with the current results, since the number of “other” activity is very limited in the current data set.

Since GPS data collection can ease the burden to the applicants in the survey, it is easier to collect the data lasting for multiple days. So the multi-day activity-travel patterns can be analyzed with these data. What is more, the multi-day travel demand model can be developed with these data.

Appendix A. Coordinate System Conversion

In this research, three types of coordinates systems are involved when dealing with the GPS trajectories, road networks, as well as POI data. In this Appendix, an introduction of how to convert among these three coordinate systems are given.

A.1. WORLD COORDINATE SYSTEM AND JAPAN COORDINATE SYSTEM

The coordinate system used in this research includes Japan coordinate system and world coordinate system. In order to calculate the distance, the conversion from one system to another is necessary. Since there is a shifting between these two coordinate systems, an approximate method advanced by *Geospatial Information Authority of Japan* was used to convert longitude and latitude in Japan coordinate system to World coordinate system. A demonstration of the shifting among these two coordinate systems can be found in Figure A.1.

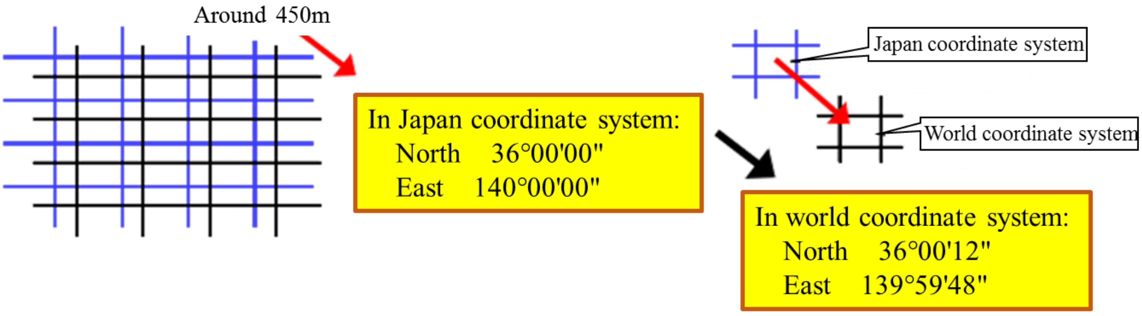


Figure A.1 The difference between world and Japan coordinate systems[2]

In the approximate method, the whole Japan was divided into several meshes and the magnitudes of shifting in longitude and latitude depend on the location of the mesh. And the corresponding shifting magnitude in each mesh can be found in this website: http://www.gsi.go.jp/MAP/NEWOLDBL/25000-50000/index25000-50000.html (in Japanese). Since the shifting magnitudes at each of the four points of the mesh are a bit different, the average of shifting magnitudes of the four points in the mesh are used. In case of the Hakodate city, the conversion formula is as follows.

$$\text{latitude\_w}=\text{latitude\_j}+9.225/3600 \qquad (A.1)$$

$$\text{longitude\_w}=\text{longitude\_j}-12.825/3600 \qquad (A.2)$$

where w is short for world coordinate system and j is short for Japanese coordinate system. The longitude and latitude are measured in degree.

The converted results have been tested by the TKY2JGD (online version) which can be

[2] The original figure is from: http://www.city.nagoya.jp/somu/cmsfiles/contents/0000004/4445/5-1mesh.pdf (in Japanese).

found at: http://vldb.gsi.go.jp/sokuchi/surveycalc/tky2jgd/main.html(in Japanese). The results by the formula and the software are almost the same.

A.2. WORLD COORDINATE SYSTEM AND CARTESIAN COORDINATE SYSTEM

For convenience of distance calculation, in this research, the longitude and latitude in the world coordinate system are converted to X and Y in the Cartesian coordinate system. The conversion formula developed by *Geospatial Information Authority of Japan* is used.

The x and y in the Cartesian coordinate system of an input GPS plot with latitude φ and longitude λ can be expressed as following formulas.

$$x = \bar{A}\left(\xi' + \sum_{j=1}^{5} \alpha_j \sin 2j\xi' \cosh 2j\eta'\right) - \bar{S}_{\varphi_0},$$
$$y = \bar{A}\left(\eta' + \sum_{j=1}^{5} \alpha_j \cos 2j\xi' \sinh 2j\,\eta'\right) \tag{A.3}$$

Where $\gamma = \tan^{-1}\left(\frac{\tau\bar{t}\lambda_c+\sigma t\lambda_s}{\sigma\bar{t}\lambda_c-\tau t\lambda_s}\right)$; $\mathrm{m} = \frac{\bar{A}}{a}\sqrt{\frac{\sigma^2+\tau^2}{t^2+{\lambda_c}^2}\left\{1 + \left[\frac{1-n}{1+n}\tan\varphi\right]^2\right\}}$; φ_0 and λ_0 are the latitude and longitude of origin point of Cartesian coordinate system; a is semi-major axis used in World Geodetic System GRS80 and its value is 6378137.0 meter; F is flattening used in World Geodetic System GRS80 and its value is 298.257222101; m_0 is the scale parameter of X axis in Cartesian coordinate system, and equals to 0.9999; $n = \frac{1}{2F-1}$;

Other parameters are calculated due to the following formulas.

$$t = \sinh\left(\tanh^{-1}\sin\varphi - \frac{2\sqrt{n}}{1+n}\tanh^{-1}\left[\frac{2\sqrt{n}}{1+n}\sin\varphi\right]\right) \tag{A.4}$$

$$\bar{t} = \sqrt{1+t^2} \tag{A.5}$$

$$\lambda_c = \cos(\lambda - \lambda_0) \tag{A.6}$$

$$\lambda_s = \sin(\lambda - \lambda_0) \tag{A.7}$$

$$\xi' = \tan^{-1}\left(\frac{t}{\lambda_c}\right) \tag{A.8}$$

$$\eta' = \tanh^{-1}\left(\frac{\lambda_s}{\bar{t}}\right) \tag{A.9}$$

$$\sigma = 1 + \sum_{j=1}^{5} 2j\alpha_j \cos 2j\xi' \cosh 2j\eta' \tag{A.10}$$

$$\tau = \sum_{j=1}^{5} 2j\alpha_j \sin 2j\xi' \sinh 2j\eta' \tag{A.11}$$

$$\alpha_1 = \frac{1}{2}n - \frac{2}{3}n^2 + \frac{5}{16}n^3 + \frac{41}{180}n^4 - \frac{127}{288}n^5 \tag{A.12}$$

$$\alpha_2 = \frac{13}{48}n^2 - \frac{3}{5}n^3 + \frac{557}{1440}n^4 + \frac{281}{630}n^5 \tag{A.13}$$

$$\alpha_3 = \frac{61}{240}n^3 - \frac{103}{140}n^4 + \frac{15061}{26880}n^5 \tag{A.14}$$

$$\alpha_4 = \frac{49561}{161280}n^4 - \frac{179}{168}n^5 \tag{A.15}$$

$$\alpha_5 = \frac{34729}{80640}n^5 \tag{A.16}$$

$$\bar{S}_{\varphi_0} = \frac{m_0 a}{1+n}\left(A_0 \frac{\varphi_0}{\rho} + \sum_{j=1}^{5} A_j \sin 2j\varphi_0\right) \tag{A.17}$$

$$\bar{A} = \frac{m_0 a}{1+n} A_0 \tag{A.18}$$

$$A_0 = 1 + \frac{n^2}{4} + \frac{n^4}{64} \tag{A.19}$$

$$A_1 = -\frac{3}{2}\left(n - \frac{n^3}{8} - \frac{n^5}{64}\right) \tag{A.20}$$

$$A_2 = \frac{15}{16}\left(n^2 - \frac{n^4}{4}\right) \tag{A.21}$$

$$A_3 = -\frac{35}{48}\left(n^3 - \frac{5}{16}n^5\right) \tag{A.22}$$

$$A_4 = \frac{315}{512}n^4 \tag{A.23}$$

$$A_5 = -\frac{693}{1280}n^5 \tag{A.24}$$

The longitude and latitude in degree need to be processed in radian first before being input in the calculation.

In case of Hakodate city, the point with latitude 41.5 degree and longitude 140.3 degree in the world coordinate system is used as the origin point in the Cartesian coordinate system. In the conversion of GPS data in Nagoya metropolitan area, the left-below point of mesh 523616 is used as origin point in the Cartesian coordinate system.